ENCYCLOPÆDIA
BRITANNICA



MACROPÆDIA

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



THE UNIVERSITY OF CHICAGO

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

The New

Encyclopædia Britannica

in 30 Volumes

MACROPÆDIA
Volume 1

Knowledge in Depth

FOUNDED 1768
15TH EDITION

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

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

Printed in U.S.A.

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

Aalto, Alvar

The Finnish architect Alvar Aalto, whose work exemplifies the best of 20th-century Scandinavian architecture, was one of the first to depart from the stiffly geometric designs common to the early period of the modern movement and to stress informality and personal expression. His style is regarded as both romantic and regional. He used complex forms and varied materials, acknowledged the character of the site, and gave attention to every detail of building. He also was active in city planning and furniture design. Aalto achieved an international reputation through his more than 200 buildings and projects, ranging from factories to churches, a number of them built outside Finland.

By courtesy of the Consulate General of Finland, New York



Aalto, 1970.

Early work. Hugo Alvar Henrik Aalto was born in Kuortane, in west central Finland, on February 3, 1898. His architectural studies at the Technical Institute of Helsinki in Otaniemi were interrupted by the Finnish War of Independence, in which he participated. Following his graduation in 1921, Aalto toured Europe and upon his return began practice in Jyväskylä, in central Finland. In 1927 he moved his office to Turku, where he worked in association with Erik Bryggman until 1933, the year in which he moved to Helsinki. In 1925 he married Aino Marsio, a fellow student, who served as his professional collaborator until her death in 1949. The couple had two children.

Important early commissions

The years 1927 and 1928 were significant in Aalto's career. He received commissions for three important buildings that established him as the most advanced architect in Finland and brought him worldwide recognition as well. These were the Turun Sanomat Building (newspaper office) in Turku, the tuberculosis sanatorium at Paimio, and the Municipal Library at Viipuri (now Vyborg). His plans for the last two were chosen in a competition, a common practice with public buildings in Finland. Both the office building and the sanatorium emphasize functional, straightforward design and are without historical stylistic references. They go beyond the simplified classicism common in Finnish architecture of the 1920s, resembling somewhat the building designed by Walter Gropius for the Bauhaus school of design in Dessau, Germany (1925–26). Like Gropius, Aalto used smooth white surfaces, ribbon windows (that is, windows in a continuous band), flat roofs, and terraces and balconies.

The third commission, the Viipuri Municipal Library, although exhibiting a similar dependence on European prototypes by Gropius and others, is a significant departure marking Aalto's personal style. Its spatially complex interior is arranged on various levels. For the auditorium portion of the library Aalto devised an undulating acoustic ceiling of wooden strips, a fascinating detail that, together with his use of curved laminated wood furniture of his own design, appealed both to the public and to those professionals who had held reservations about the clinical severity of modern architecture. The warm textures of wood provided a welcome contrast to the general whiteness of the building. It was Aalto's particular success here that identified him with the so-called organic approach, or regional interpretation, of modern design. He continued in this vein, with manipulation of floor levels and use of natural materials, skylights, and irregular forms.

By the mid-1930s Aalto was recognized as one of the world's outstanding modern architects; unlike many of his peers, he had an identifiable personal style.

Finnish pavilions for two world's fairs (Paris, 1937; New York, 1939–40) further enhanced Aalto's reputation as an inventive designer of free architectural forms. In these designs, both chosen in competition, he continued to use wood for structure and for surface effects. Further renown came in 1938, when the Museum of Modern Art in New York City held an exhibition of his work, showing photographs of his buildings and examples of his furniture.

Experiments in furniture

Aalto's experiments in furniture date from the early 1930s, when he furnished the sanatorium at Paimio. His furniture is noted for its use of laminated wood in ribbonlike forms that serve both structural and aesthetic ends. In 1935 the Artek Company was established by Aalto and Mairea Gullichsen, the wife of the industrialist and art collector, to manufacture and market his furniture. The informal warmth of Aalto's interiors is best seen in the much admired country home Villa Mairea, which he built for the Gullichsens near Noormarkku, Finland.

Mature style. The decade of the 1940s was not productive; it was disrupted by war and saddened by his wife's death. In 1952 he married Elissa Mäkiniemi, a trained architect, who became his new collaborator.

Aalto's commissions after 1950, in addition to being greater in number, were more varied and widely dispersed: a high-rise apartment building in Bremen (1958), a church in Bologna (1966), an art museum in Iran (1970). His continuing work in Finland, however, remained the measure of his genius. Many of his projects involved site planning of building groups. Two such projects were the master plans of colleges at Otaniemi (1949–55) and at Jyväskylä (1952–57). Aalto's experience in planning originated early with such industrial commissions as the Sunila cellulose factory (designed 1936–39, built 1951–54), which included workers' housing and was a triumph of comprehensive planning.

The single work that epitomizes Aalto's mature style is perhaps the Säynätsalo town hall group. Modest in scale in its forest setting, it nonetheless asserts a quiet force. Its simple forms are in red brick, wood, and copper, all traditional materials of Finland. Viewing it, a person feels the achievement of a perfect building, in that the essence of the time, the place, the people, and their purpose is brought into focus by the awareness of the architect.

There are two buildings by Aalto in the United States: Baker House dormitory, at the Massachusetts Institute of Technology in Cambridge, where Aalto was professor of architecture from 1945 to 1949; and the library of Mount Angel Abbey (a Benedictine college and seminary), in Oregon. Baker House, on the embankment of the Charles River, takes the form of a six-story serpentine wall of rough red brick. Each of the dormitory rooms, some of which are wedge-shaped, has a view either up or down the river. Like Baker House, the Mount Angel library is an example of Aalto's response to the site. Its interior construction follows the downward-sloping terrain, enclosing a dramatic, multilevel room that fans out from the entrance. The bookcases are arranged in radial pattern, and the room is lighted by curved skylights that repeat the rhythmic lines of the balconies.

Aalto received many honours. He was a member of the Academy of Finland (Suomen Aketemia) and was its president from 1963 to 1968; he was a member of the Congrès Internationaux d'Architecture Moderne from 1928 to 1956. His awards included the Royal Gold Medal for Architecture from the Royal Institute of British Architects (1957) and the Gold Medal from the American Institute of Architects (1963). He died in Helsinki on May 11, 1976.

Assessment. Aalto's preliminary plans were freely sketched, without the use of T-square and triangle, so that the unfettered creative urge for inventive shapes and irregular forms was allowed full play before functional relationships and details were resolved. The absence of theoretical rigidity revealed itself in his final designs, which happily retained the spontaneity and individuality of his early sketches. As a Swiss art historian expressed it, he dared "the leap from the rational–functional to the irrational–organic." Since Aalto's staff was small (some six to eight architects), all of the work bore the imprint of his personality.

Aalto wrote little to explain his work, but his architecture conveyed a variable, lively temperament, free from dogma and without monotony. His work was said to express the spirit of Finland and its people, primitive yet lyrical. His friendships with such artists as Fernand Léger, Jean Arp, and Constantin Brancusi may have nourished his fondness for curvilinear shapes. While his work was never compulsively innovative, neither was it static. His late designs showed an increased complexity and dynamism that some regarded as incautious. In particular, his work of the late 1960s was marked by splayed, diagonal shapes and clustered, overlapping volumes. Energy and imagination were ever present.

MAJOR WORKS

Turun Sanomat Building, Turku, Finland (1930); Municipal Library at Viipuri, now Vyborg in the Russian Soviet Federated Socialist Republic (designed 1927, built 1930–35, destroyed 1943); Sulfate Paper Mill at Toppila, Finland (1931); Sanatorium at Paimio, Finland (1933); cellulose factory at Sunila, Kotka, Finland (1936–39; extended 1951–54); Villa Mairea (Gullichsen House) at Noormarkku, Finland (1938–39); sawmill at Varkaus, Finland (1945); Baker House (senior dormitory), Massachusetts Institute of Technology, Cambridge, Massachusetts (1947–48); town hall group, Säynätsalo, Finland (1950–52); House of Culture, Helsinki, Finland (1955–58); house for Louis Carré, Bazoches, France (1956–58); church at Vuoksenniska, Imatra, Finland (1956–58); Nordjyllands Kunstmuseum, Ålborg, Denmark (1958–72); post and telegraph office, Baghdad, Iraq (1958); Community Centre, Wolfburg, Germany (1959–62); Community Centre, Seinäjoki, Finland (1962; theatre added, 1967); Edgar J. Kaufmann Conference Rooms, Institute of International Education, New York City (1965); Mount Angel Abbey Library, near Salem, Oregon (designed 1965–66, built 1967–70).

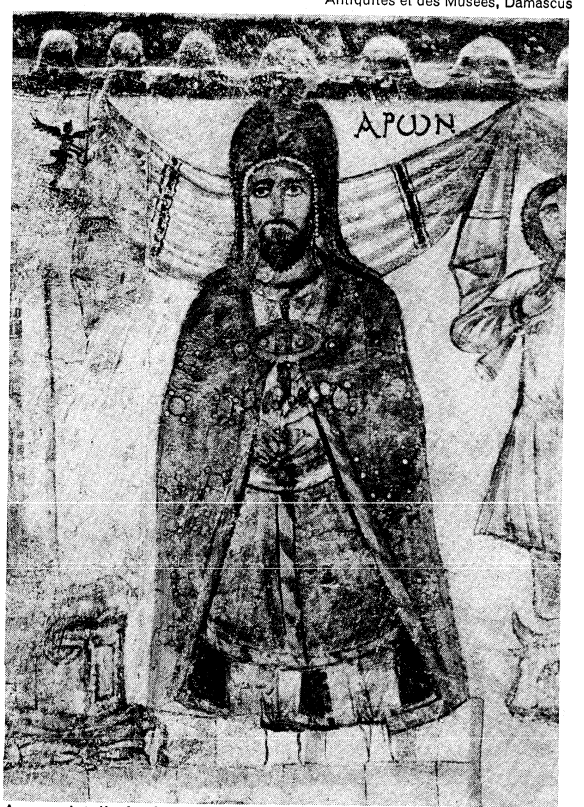
BIBLIOGRAPHY. A definitive review of Aalto's work, extremely well illustrated with photographs and plans, is the two-volume work edited by KARL FLEIG: *Alvar Aalto* (1963) and *Alvar Aalto 1963–1970* (1971). A brief essay by FREDERICK GUTHEIM, *Alvar Aalto* (1960), contains descriptive and critical comments as well as many photographs. A similar book is GEORGE BAIRD, *Alvar Aalto* (1970). A substantial chapter on Aalto appears in SIGFRIED GIEDION, *Space, Time and Architecture*, 5th ed. rev. (1967), and is an excellent critical treatment of Aalto's architecture. No purely biographical work exists.

(H.F.K.)

Aaron

Aaron is the traditional founder and head of the Jewish priesthood, who, in company with Moses, led the Israelites out of Egypt.

By courtesy of the Direction Generale des Antiquites et des Musees, Damascus



Aaron, detail of a fresco by an unknown artist, 3rd century, from the synagogue at Doura-Europus. In the National Museum, Damascus.

Life. Aaron is described in the Old Testament book of Exodus as a son of Amram and Jochebed of the tribe of Levi, three years older than his brother Moses. He acted together with his brother in the desperate situation of the Israelites in Egypt and took an active part in the Exodus. Although Moses was the actual leader, Aaron acted as his "mouth." The two brothers went to the Pharaoh together, and it was Aaron who told him to let the people of Israel go, using his magic rod in order to show the might of Yahweh. When the Pharaoh finally decided to release the people, Yahweh gave the important ordinance of the Passover, the annual ritual remembrance of the Exodus, to Aaron and Moses. But Moses alone went up on Mt. Sinai, and he alone was allowed to come near to Yahweh. Moses later was ordered to "bring near" Aaron and his sons, and they were anointed and consecrated to be priests "by a perpetual statute." Aaron's sons were to take over the priestly garments after him. Aaron is not represented as wholly blameless. It was he who, when Moses was delayed on Mt. Sinai, made the golden calf that was idolatrously worshipped by the people.

Aaron as "mouth" for Moses

Once a year, on Yom Kippur (the Day of Atonement), Aaron was allowed to come into the Holy of Holies, the most sacred part of the tabernacle, or sanctuary, in which

the Hebrew tribes worshipped, bringing his offering. Together with his sister, Miriam, Aaron spoke against Moses because he had married a foreigner (a Cushite woman); but as in the episode of the golden calf, the narrative tells how Aaron was merely reproved, though Miriam was punished, for the offense. In the rebellion of Korah the Levite, however, Aaron stood firmly at the side of Moses. According to Numbers 20, Aaron died on the top of Mt. Hor at the age of 123; in Deuteronomy 10, which represents another tradition, he is said to have died in Moserah and was buried there.

Aaron is a central figure in the traditions about the Exodus, though his role varies in importance. At the beginning he seems to be coequal with Moses, but after the march out of Egypt he is only a shadow at Moses' side. Moses is obviously the leading figure in the tradition, but it is also clear that he is pictured as delegating his authority in all priestly and cultic matters to Aaron and "his sons."

Aaron and the biblical critics. Scholars have long been aware that the figure of Aaron as it is now found in the Pentateuch, or first five books of the Old Testament, is built up from several sources or layers of traditions. According to Julius Wellhausen, a German biblical scholar, and his followers, the Jahwist source was the oldest one, followed in order by the Elohist, Deuteronomist, and Priestly code. Scholars have attributed the passages about Aaron to one or the other of these sources. Although their results differ, they do agree in ascribing about 90 percent of the material about Aaron to the Priestly source. According to Wellhausen, Aaron was not mentioned at all in the Jahwist narrative, but he may have been inserted by later redactors. It was Moses who was the hero of the priests before the exile, and it was Joshua, not Aaron, who officiated in the tabernacle.

Major source or tradition

Other scholars, such as Sigmund Mowinckel, believe that the narrative about the golden calf, which presents Aaron in an unfavourable light, was part of the ancient tradition in the Jahwist work, being the only passage in it that mentions him. This narrative, according to these scholars, originally came from the northern kingdom of Israel and described Aaron as the ancestor of the priests in northern Israel; later it was rewritten in a way defamatory to Aaron. But there are also features in the narrative that may indicate that a later source (or traditionist), the Elohist, tried to excuse Aaron and to put the main responsibility on the people. The Elohist narrator was credited with making Aaron the brother and helper of Moses, who stood at the side of Moses in the conflict with the Pharaoh and assisted him as a leader in battles and in the cult. It may also be the Elohist who provides the unfavourable story about Aaron's objection to Moses' wife. On the other hand, it seems to be the same narrator who mentions Aaron at the side of Moses in the revolt at Meribah, but here also Aaron, together with Moses, is actually reproached. There is reason to believe that Aaron was not mentioned in the Deuteronomist work by the original author but that his name has been added by a redactor. The main bulk of the traditions about Aaron and the frequent addition of "and Aaron" after the mention of Moses are found in the Priestly source, which was written at a time when the priests had a more dominant position in Judah than they had before the exile. By then Moses had ceased to be the hero of the priests, and Aaron had taken over that role.

Many modern scholars prefer to speak of traditions and layers of traditions where their predecessors spoke of sources, but, apart from this terminology, the view concerning Aaron has not greatly changed. There have been new attempts, however, to see the contrasting figures of Moses and Aaron in a new light. It has been suggested that the traditions about Moses represent a southern Judean tradition, while the old traditions about Aaron originated in the northern kingdom. It has also been indicated that the traditions about Moses are primarily concerned with a prophet, while those about Aaron are connected with priesthood. There may be a kernel of truth in all these suggestions, as also in the theory of Ivan Engnell that Moses represents the royal ideology while Aaron stands for priesthood, and priesthood alone. The standing struggle between the king and the leading priests is reflected both in the laws and in the narratives of the historical books.

Aaron in later Jewish and Christian thought. Aaron continued to live as a symbol in Jewish religion and traditions, and the position of the priests was strengthened after the exile. Also in the Qumrān sect, a Jewish community that flourished in the era immediately before and contemporary with the birth of Christianity, Aaron was a symbol for a strong priesthood, as can be seen from the Dead Sea Scrolls. At the end of time, men of the community should be set apart, as a select group in the service of Aaron. Only the sons of Aaron should "administer judgment and wealth," and according to the *Manual of Discipline* two messiahs were expected, one of Aaron, the priestly one, and one of Israel. According to a fragment found near Qumrān, the priest would have the first seat in the banquets in the last days and bless the bread before the Messiah of Israel. Here "the sons of Aaron" have the highest position.

Aaron as a symbol

In Talmud and Midrash (Jewish commentative and interpretative writings), Aaron is seen less as a symbol than as the leading personality at the side of Moses. The relationship between the two brothers is painted as prototypical in the *Haggada* ("Narrative"—the nonlegal parts of Talmud and Midrash). Rabbi Hillel, the great liberal sage, praised Aaron as peace loving, a man of goodwill, who wanted to teach his fellow men the Law.

In Jewish exegesis little is said about him, though he is mentioned as a man who created peace among men. Many attempts have been made to explain the episode of the golden calf. According to some exegetes, Aaron had to make the calf in order to avoid being killed. In the early 14th century, Gersonides explained that this would have been fatal not only for Aaron but even more for the people. Earlier, in the 11th century, Rashi contended that the calf was a symbol of the leader, Moses, who was at that time on the mountain. The relationship between Moses and Aaron is also discussed in the Talmud. Some traditionists have wondered why Aaron, and not Moses, was appointed high priest. The answer has been found in an indication that Moses was rejected because of his original unwillingness when he was called by Yahweh. It also seems to have been hard for some traditionists to accept that Aaron was described as older than Moses. The death of Aaron is related in the Midrash *Petirat Aharon*.

The first Christian communities admitted that Aaron, "the sons of Aaron," or "the order of Aaron" were symbols of the highest priesthood. But in the Letter to the Hebrews, Christ is described as a high priest according to the order of Melchizedek, which was set against "the order of Aaron." Of the Church Fathers, Cyril of Alexandria says that Aaron was divinely called to a priesthood in spirit and in truth and that he was a type of Christ. Gregory the Great translates the name Aaron as "mountain of strength" and sees in him a redeemer who mediated between God and man.

BIBLIOGRAPHY. F.S. NORTH, "Aaron's Rise in Prestige," in *Zeitschrift für die alttestamentliche Wissenschaft*, 66:191–199 (1954), is a short paper discussing Aaron's position; ROLAND DE VAUX, *Les Institutions de l'Ancien Testament*, 2 vol. (1958–60; Eng. trans., *Ancient Israel: Its Life and Institutions*, 1961), a comprehensive work, also gives a picture of Aaron and his place in ancient Israelite life and cult. WALTER BEYERLIN, *Herkunft und Geschichte der Ältesten Sinaitraditionen* (1961; Eng. trans., *Origins and History of the Oldest Sinaitic Traditions*, 1966), is mostly concerned with the traditions behind the story of Moses and Aaron. HAROLD H. ROWLEY, *Worship in Ancient Israel* (1967), gives a picture of religious life in Israel, including Aaron and his role. BERTIL GARTNER, *The Temple and the Community in Qumran and the New Testament* (1965), treats the religious cult in the Qumrān society, in which "Aaron" was synonymous with the chief priest.

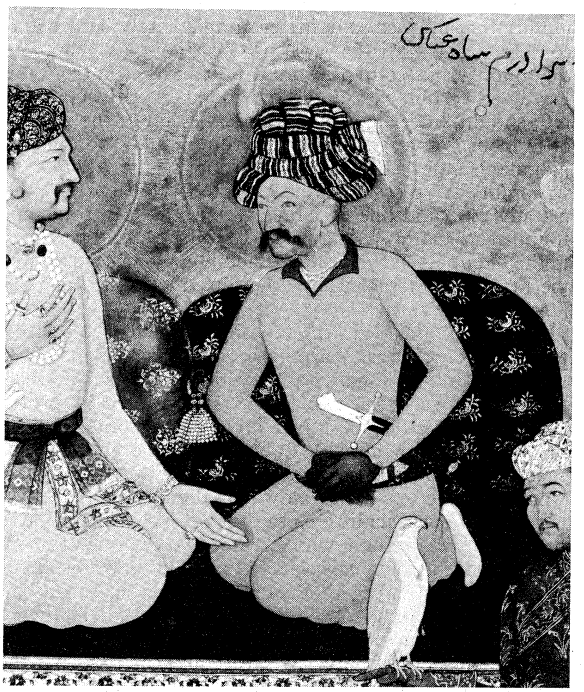
(A.S.K.)

ʿAbbās I the Great, of Persia

ʿAbbās I, styled "the Great," shāh (king) of Persia of the Ṣafavid dynasty, the first Ṣafavid monarch to create a standing army, won back substantial territory lost under

his predecessors to the Uzbeks and the Ottoman Turks. Although a just and courageous ruler whose reign was distinguished by its high artistic level, he tarnished his reputation by the murder and mutilation of many members of his family.

By courtesy of the Smithsonian Institution, Freer Gallery of Art, Washington, D.C.



ʿAbbās I (right) seated with the emperor Jahāngīr, detail of a painting by the Mughal school of Jahāngīr, *c.* 1620. In the Freer Gallery of Art, Washington, D.C.

Early years

The third son of Solṭān Moḥammad Shāh, ʿAbbās was born on January 27, 1571, and came to the throne in October 1588, at a critical moment in the fortunes of the Ṣafavid dynasty. The weak rule of his semi-blind father had allowed the tribal emirs, or chiefs, of the Turkmen tribes, who had brought the Ṣafavids to power and still constituted the backbone of Ṣafavid military strength, to usurp the authority of the shāh. Moreover, the inter-tribal factionalism of these Turkmens (known as Kizilbash, "Red Heads," because of the distinctive red headgear they had adopted to mark their adherence to the Ṣafavids) had so weakened the state that its traditional enemies, the Ottoman Turks to the west and the Uzbeks to the east, had been able to make large inroads into Persian territory.

Shāh ʿAbbās thus had two immediate tasks: to restore internal security and reassert the authority of the monarchy and to expel Ottoman and Uzbek troops from Persian soil. Since he was unable to fight a war on two fronts simultaneously, in 1589–90 he signed a peace treaty with the Ottomans, thus freeing himself for an offensive against the Uzbeks. By the treaty, large areas in west and northwest Persia were ceded to the Ottomans. Despite the breathing space thus gained, ʿAbbās for ten years was unable to launch a major offensive against the Uzbeks, and Iran suffered further loss of territory both to the Uzbeks and to the Mughals of India.

The delay was caused by ʿAbbās' decision to create a standing army—a concept novel to Ṣafavid kings, who traditionally levied armies in time of need from the tribal cavalry. The creation of a standing army immediately caused a budgetary problem, because the old tribal cavalry had been paid from the revenues of the provinces governed by Kizilbash chiefs. ʿAbbās solved the problem in the short term by bringing a number of these provinces directly under the control of the Shāh; the taxes in these new "crown" provinces were remitted to the royal treasury. In the long run the inevitable result of this policy, the reduction in the numbers of Kizilbash troops, seriously weakened the country's military strength.

The new standing army was composed mainly of Georgians, Armenians, and Circassians, who had been brought to Persia as prisoners during the reign of ʿAbbās' grandfather, and their descendants. After their conversion to Islām, they were trained for service either in the army or in the administration of the state or the royal household. Shāh ʿAbbās felt he could rely on the loyalty of these *ghulāms* ("slaves") of the Shāh, as they were known, and he used them to counterbalance the influence of the Kizilbash, whom he distrusted. *Ghulāms* soon rose to high office and were appointed governors of crown provinces.

Eventually, ʿAbbās was able to take the offensive against his external foes. In 1598 he inflicted a major defeat on the Uzbeks and regained control of Khorāsān. From 1602 onward he conducted a series of successful campaigns against the Ottomans and recovered the territory lost to them.

Rebuilding of Isfahan

After his great victory over the Uzbeks, ʿAbbās transferred the capital from Kazvin to Isfahan. Under his guidance, Isfahan rapidly became one of the most beautiful cities in the world. He adorned the city with many mosques and theological colleges and constructed numerous caravansaries and public baths. He laid out the city with spacious boulevards and a splendid square. The Shāh's building energies were not confined to Isfahan; the extension and restoration of the famous shrine at Meshed and the construction, along the swampy littoral of the Caspian Sea, of the celebrated stone causeway, designed to give access to his favourite winter retreats, were among his most notable achievements.

To the new capital at Isfahan came ambassadors from European countries, merchants seeking to establish trade relations, representatives of foreign monastic orders seeking permission to found convents at Isfahan and elsewhere, and gentlemen of fortune, such as the brothers Sir Anthony and Sir Robert Sherley—the former an adventurer, the latter a loyal servant of the Shāh who distinguished himself in the wars against the Ottomans. The reign of Shāh ʿAbbās was a period of intense commercial and diplomatic activity, and, in the Persian Gulf, the Portuguese, the Dutch, and the English strove to make themselves masters of trade there and in the Indian Ocean.

ʿAbbās' reign also marks a peak of Persian artistic achievement. Under his patronage, carpet weaving became a major industry, and fine Persian rugs began to appear in the homes of wealthy European burghers. Another profitable export was textiles, which included brocades and damasks of unparalleled richness of colour and intricacy and freedom of design. The production and sale of silk was made a monopoly of the crown. In the illumination of manuscripts, bookbinding, and ceramics, the work of the period of ʿAbbās is without equal; in painting it is among the most notable in Persian history.

Assessment

Shāh ʿAbbās was a courageous and energetic man and a ruler with a passionate zeal for justice and the welfare of his subjects. He frequented teahouses and other meeting places of the ordinary people in order to learn of extortion and oppression on the part of his officials; his punishment of corrupt officials was swift. He showed unusual religious tolerance, granting privileges to many Christian groups.

The dark side of his character was reserved for his own sons and members of his own family. The experiences of his youth, when he was marked for execution by his uncle, Shāh Esmāʿīl II, had left him with a morbid fear of conspiracy. Originally, he followed the practice of his predecessors in appointing the princes of the blood royal as provincial governors, but after a series of revolts and intrigues in favour of his sons, the royal princes were confined to the harem, where their only companions were women and eunuchs. As his obsessive fear of assassination increased, ʿAbbās began to put to death or to blind any member of the royal family who caused him anxiety in this regard. In this way, one son was executed (an act that caused ʿAbbās bitter remorse) and two were blinded, and his father and brothers were blinded and imprisoned.

ʿAbbās died on January 19, 1629, without an heir capable of succeeding him.

Though ʿAbbās possessed great stature as a monarch—even in an age notable for its outstanding rulers—his great achievement in first saving the Ṣafavid Empire from collapse and then raising it to new heights of splendour is marred by his treatment of his own family and the fact that his reforms contained within them the seeds of the future decay of both dynasty and state.

BIBLIOGRAPHY. L.L. BELLAN, *Chah ʿAbbās I: sa vie, son histoire* (1932), the only biography of Shāh ʿAbbās I in any European language—generally accurate; R.M. SAVORY, "ʿAbbās I" in the *Encyclopaedia of Islam*, new ed., vol. 1, pp. 7–8 (1960); for general background to the period of Shāh ʿAbbās, the reader is referred to: J. CHARDIN, *Voyages du Chevalier Chardin*, 4th ed., 4 vol. (1811); SIR JOHN MALCOM, *The History of Persia from the Early Period to the Present Time* (2 vol., 1815); and V. MINORSKY, *Tadhkirat al-Mulūk* (1943), especially the introduction, commentary, and appendixes.

(R.M.Sa.)

ʿAbd al-Malik ibn Marwān

The reign of ʿAbd al-Malik ibn Marwān (685–705), the fifth caliph of the Umayyad Arab dynasty (661–750), witnessed the re-establishment of Umayyad power throughout the Muslim Empire as well as a major reorganization of government administration. A wise and tenacious ruler, ʿAbd al-Malik was able to bring peace and prosperity to an empire torn by civil war and to ensure the continuation of the rule of his dynasty for the next half century. (For a discussion of religious and dynastic conflicts that divided the Muslim world in this period see CALIPHATE, EMPIRE OF THE.)

Early years and life in Medina

ʿAbd al-Malik was born in 646 or 647, the son of Marwān ibn al-Ḥakam, fourth Umayyad caliph. He spent the first half of his life with his father in Medina, where he received religious instruction and developed friendly relations with the pious circles of that city that were to stand him in good stead in his later life. At the age of 16, he was entrusted by his kinsman, the caliph Muʿāwiyah, with administrative responsibilities. He remained at Medina until 683, when he and his father were driven out of the city by Medinese rebels in revolt against the central government in Damascus. He then met the Syrian Umayyad army that was marching on Medina and gave its commander advice about the best means of attacking the city, advice that was followed and proved successful. When the caliph Yazīd died in November 683, Marwān was proclaimed caliph in 684 and was able to effect a partial rally of Umayyad rule, but at the cost of a bitter feud that arose between northern and southern Arab tribes. When Marwān died in 685 and ʿAbd al-Malik succeeded to the caliphate, the forces opposing the Umayyads were still formidable.

There were, first, the northern Arab tribes who, under their leader Zufar, were holding out in northern Syria and Iraq. They were finally pacified only in 691. The second focus of resistance was in Iraq, where three main groups, opposed to each other but united in their resistance to the Umayyads, held sway: the Khārijites, the Shīʿites, and the forces of the anticaliph ʿAbd Allāh ibn az-Zubayr, who was proclaimed caliph in Mecca in 685 and had received at least nominal allegiance from many provinces. The initial attempts by the former Umayyad governor of Iraq, ʿUbayd Allāh ibn Ziyād, to regain the province failed, and he was killed by the Shīʿites in 686. For three years ʿAbd al-Malik made no further attempt to interfere in Iraq but bided his time as the various groups in Iraq exhausted themselves in internecine warfare. Muṣʿab, the brother of the anticaliph Ibn az-Zubayr, defeated the Shīʿites in 687 but then had to deal with the Khārijites, committing a large part of his forces.

The decisive battle with Muṣʿab in 691

ʿAbd al-Malik first took the field against Muṣʿab in 689 but had to turn back to quell a rebellion in Damascus. In the following year, the campaign again proved fruitless. Only after the defeat of the northern Arab tribes in 691 was ʿAbd al-Malik finally able to face Muṣʿab. The decisive battle took place at Dayr al-Jā Thalīq. The forces of Muṣʿab were weakened by their wars against the Khārijites, and ʿAbd al-Malik bribed many of them to desert Muṣʿab, who was then killed in battle. The whole of Iraq now fell into his hands, and the only remaining centre of opposition was the now aging anticaliph, Ibn az-Zubayr. ʿAbd al-Malik publicly chided him for his temerity and then sent his famous governor al-Ḥajjāj to Arabia. Al-Ḥajjāj besieged Ibn az-Zubayr in Mecca and killed him in September 692. The Muslim community was finally unified.

At first, the re-establishment of Umayyad rule was more apparent than real. The Khārijites were still either restless or in open revolt. The Khārijites in Persia were especially dangerous. It was only after ʿAbd al-Malik had appointed al-Ḥajjāj to govern Basra that campaigns against them began to prove successful (the Persian Khārijites were finally wiped out in 697). But north of Kūfah, another Khārijite trouble centre developed. In 695, these Khārijites captured Mosul and occupied large areas of central Iraq. Al-Ḥajjāj, leading his Syrian troops, defeated them too in 697. The Khārijite movement, however, remained strong, especially among the Bakr tribes between Mosul and Kufah.

Al-Ḥajjāj had now become governor of all the eastern provinces. He was a ruthless and efficient administrator, intent upon pacifying all the provinces entrusted to him by ʿAbd al-Malik. A great Muslim army, led by an Arab aristocrat, Ibn al-Ashʿath, and operating in the Afghanistan region, mutinied, swore allegiance to its commander, and turned back to Iraq. Al-Ḥajjāj, with the aid of Syrian reinforcements, was able to defeat the rebels, and their leader was murdered in 704 in Afghanistan. Al-Ḥajjāj, realizing that he could no longer trust the Iraqis, built a new city, Wāsiṭ, which he planned as a garrison city for Syrian troops and as his private residence. Thereafter, he ruled Iraq as enemy territory.

Under ʿAbd al-Malik, the conquest of North Africa was resumed in 688 or 689. There, the Arabs were opposed by both the native Berbers and the Byzantines. The governor appointed by ʿAbd al-Malik succeeded in winning the Berbers over to his side and then captured Carthage, seat of the Byzantine province, in 697. Other coastal cities fell, and the work of pacification and Islāmization continued apace. ʿAbd al-Malik also resumed campaigns against the Byzantines in Anatolia in 692, but no permanent conquest ensued. These campaigns were partly designed to keep the Syrian troops fit.

Assessment of ʿAbd al-Malik as an Umayyad ruler

In general, Umayyad rule was greatly strengthened by ʿAbd al-Malik, who enjoyed good relations with the Medinese religious circles, an element with considerable moral influence in the Islāmic world. ʿAbd al-Malik was more pious than any of his Umayyad predecessors. His long sojourn in Medina had enabled him to know the sentiments of Medinese religious scholars. As caliph, he treated them respectfully, and his private life was close to their ideals. As a result, many were to abandon their earlier opposition to Umayyad rule.

ʿAbd al-Malik adopted Arabic instead of the local languages as the language of administration. Government officials had been mostly non-Muslim, but the measures of ʿAbd al-Malik enabled Arab Muslims more easily to control affairs of government. A new Muslim currency was also struck, modelled on Greek and Persian coinage, but with Muslim inscriptions. A wave of Islāmization set in, but the privileged position of the Arabs was maintained. In fact, the problem of non-Arab Muslims grew more acute and was to become one of the main threats to Umayyad rule in later years.

The Umayyad family lived in Damascus and surrounded the Caliph. Many of them were appointed as governors, but many were also recalled for inefficiency. ʿAbd al-Malik enjoyed the support of his clan, but he was more autocratic than Muʿāwiyah, the first Umayyad caliph, with whom he is often compared. He abandoned the policy of consulting with a council of advisers and reserved all major decisions for himself. Despite his religious interests and ideals (*e.g.*, he built the Dome of the Rock in Jerusalem), he was a master politician. In Syria, he succeeded in placating the northern Arab tribes, to the chagrin of the southern Arabs.

ʿAbd al-Malik was a shrewd judge of character. His choice of al-Ḥajjāj as viceroy of the East was a wise one, and he supported his lieutenant loyally. In appearance, he was dark, thickset, and had a long beard. He was nicknamed "Dew of the Stone" for his miserliness. The sources describe him as eloquent in his speech and a lover of poetry. He maintained his calm during periods of crisis and was decisive in his opinions but was capable of great cruelty if necessary. He pursued his enemies relentlessly and closely supervised all affairs of state.

Shortly before his death, the question of succession became acute. His brother, ʿAbd al-ʿAzīz, governor of Egypt, had been designated by their father to succeed ʿAbd al-Malik. Against the advice of his courtiers, ʿAbd al-Malik had begun to take steps to exclude his brother from succession in favour of his own children. He had tried to pressure ʿAbd al-ʿAzīz to renounce his claims, but without success. Luckily for ʿAbd al-Malik, ʿAbd al-ʿAzīz died in May 705. ʿAbd al-Malik now felt free to name three of his own children to succeed him, al-Walīd, Sulaymān, and Yazīd. ʿAbd al-Malik died in Damascus shortly thereafter, in October 705, and was succeeded without difficulty by his eldest son, al-Walīd.

BIBLIOGRAPHY. The best account of ʿAbd al-Malik's reign is in J. WELLHAUSEN, *Das arabische Reich und sein Sturz* (1902; Eng. trans., *The Arab Kingdom and Its Fall*, 1927). Also useful are SIR WILLIAM MUIR, *The Caliphate: Its Rise, Decline, and Fall*, new ed. (1924); and P.K. HITTI, *History of the Arabs*, 8th ed. (1964).

(T.Kh.)

ʿAbd al-Muʾmin

No Berber sovereign of the Middle Ages achieved as much or exerted a greater influence on the course of events than the Almohad caliph ʿAbd al-Muʾmin, who in the second half of the 12th century founded an Islāmic empire that included most of Spain and North Africa.

Early life

ʿAbd al-Muʾmin was born (date unknown) at Tāgrā, a small village near Oran, in what is now Algeria. He came from a humble family: his father had been a potter. He seems to have been well instructed in the Muslim faith, and must have had a good knowledge of Arabic, for he wished to continue his studies at one of the centres of Muslim learning in the East. A chance meeting with Ibn Tūmart, a Berber religious reformer, made him abandon this idea and begin his brilliant career.

Around 1117, Ibn Tūmart, the founder of the Almohad movement, was returning from a long stay in the East. He landed at Mahdīyah in Tunisia, and began a journey to southern Morocco, his native country. Wherever he stopped along the way, he proclaimed a twofold message: strict adherence to the doctrine of the oneness of God (hence the name Almohads or al-Muwaḥḥidūn, Unitarians) and scrupulous observance of Islāmic law. ʿAbd al-Muʾmin heard Ibn Tūmart preach at Mellala, near Bejaïa, Algeria. He was an attentive listener, and from that time attached himself to the man who had revealed to him the true doctrine.

ʿAbd al-Muʾmin does not seem to have played any special role among Ibn Tūmart's disciples during the slow journey that took them to Marrakesh. But when his master declared his opposition to the ruling Almoravid regime, proclaimed himself the *mahdī* ("divinely-guided one"), and took refuge in the remote High Atlas region, ʿAbd al-Muʾmin went with him. Ibn Tūmart won a following in the mountains and founded a small Almohad state there, centred around the village of Tinmel. When al-Bashīr, the reformer's second in command, was killed in an attack on Marrakesh, ʿAbd al-Muʾmin took his place and became Ibn Tūmart's designated successor. The *mahdī* died in 1130. His death was kept secret at first to allow ʿAbd al-Muʾmin—a stranger to the High Atlas—time to win support from the Almohad leaders. When he was proclaimed leader of the Almohads, he assumed the prestigious title of caliph.

His first task was to carry on the struggle against the Almoravids. Learning from the failure at Marrakesh, he realized that he must conquer Morocco from the mountains. On the plains, the Christian knights who served the Almoravids could easily repulse the Almohads' Berber infantry. He spent the next 15 years winning control of the High Atlas, Middle Atlas, and Rif regions, finally moving into his native country, north of Tlemcen.

Near that town, the Almoravids, having suffered the loss of Reverter, the leader of their Catalan mercenaries, were defeated by ʿAbd al-Muʾmin in open battle in 1145. The Almohad forces then moved west, subjugating Morocco's Atlantic coastal plain. They then laid siege to Marrakesh and took it by storm in 1147, massacring the Almoravid inhabitants.

Conquest of the empire

Arab historians have left a description of the man who had now become master of Northwest Africa. He was a sturdy Berber of medium height, with dark hair and regular features. A good soldier, with great courage and endurance, he was at the same time learned in Islām and a gifted orator. Although he had personal charm and could, when necessary, show patience and moderation, he was at times as harsh as his master, Ibn Tūmart. When a revolt broke out in the Atlantic plain area following the capture of Marrakesh, he conducted a methodical purge there in which more than 30,000 people were executed.

ʿAbd al-Muʾmin left neither memoirs nor a political testament; his ideas must be deduced from his actions. His newfound power and his very success raised problems that demanded immediate solutions.

The capture of Marrakesh posed the moral question of whether to abandon this city founded by the Almoravid heretics, whom he had exterminated without pity. He contented himself with destruction of their palace and mosques and retained Marrakesh as the capital of his new empire.

Soon he had to choose between two imperial policies: to complete the conquest of North Africa or to concentrate his energies on Spain, where the Christians were threatening the former Almoravid domains. Showing good judgment as well as feeling for his native country, he gave priority to North Africa.

In 1151 he subjugated the area around Constantine and on his way home fought a battle near Sétif against a coalition of Arab tribes that had been wandering over the Berber country for a century, gradually destroying its peasant and sedentary way of life. ʿAbd al-Muʾmin was victorious, but instead of punishing these people who had showed themselves to be the worst enemies of the Berbers and the Almohad government, he thenceforth came to rely on them to strengthen his dynasty against opposition from the family of Ibn Tūmart. He also wished to use the Arab cavalry in his holy war against the Christians in Spain.

In 1158–59, ʿAbd al-Muʾmin conquered Tunisia and Tripolitania. This marked the zenith of Berber power in Islām: a Berber caliph reigned over all of North Africa west of Egypt, and his authority was acknowledged by most of Muslim Spain as well.

ʿAbd al-Muʾmin's government

Even while he was pursuing his conquest, ʿAbd al-Muʾmin had established a central government for his empire. To the traditional clan organization of the Maṣmudah and other Berber peoples supporting the Almohads, he added an organization to promote the spread of Almohad doctrine and a central administration (the *makhzan* modelled on those of Muslim Spain, which was staffed largely by Spanish Muslims. A government land registry was established to assure the dynasty regular revenue. ʿAbd al-Muʾmin fully accepted the responsibilities of an art patron, but remembering the puritanical austerity of Ibn Tūmart, he sometimes imposed on the mosques built for him by Andalusian artisans a plainness that became more artificial than elaborate ornamentation.

ʿAbd al-Muʾmin died at Rabat, Morocco, in the spring of 1163. His work, faithfully carried on by his successors Abū Yaʿqūb Yūsuf (reigned 1163–84) and Abū Yūsuf Yaʿqūb al-Manṣūr (1184–99), was maintained for more than half a century. Disturbances caused by the rebellious Arab tribes impoverished the country without endangering the dynasty. After their defeat by the Spanish Christians at Las Navas de Tolosa in 1212, however, the Almohads began to decline, and their empire soon disintegrated.

Though in the long run ʿAbd al-Mu'min's successors proved unable to perpetuate his achievements, he himself had written one of the most glorious chapters in the history of the Muslim West.

BIBLIOGRAPHY. Information about ʿAbd al-Mu'min and his work may be found in the general histories by HENRI TERRASSE, *Histoire du Maroc*, vol 2, pp. 261–316 (1950); and AMBROSIO HUICI MIRANDA, *Historia política del Imperio Almohade*, 2 vol. (1956–57), both with numerous bibliographic references.

(H.-L.-É.T.)

ʿAbd ar-Raḥmān III an-Nāṣir

ʿAbd ar-Raḥmān III an-Nāṣir (reigned 912–961), the eighth prince and first caliph of the Umayyad Arab Muslim dynasty of Spain, is generally regarded as the greatest ruler of his dynasty and the one who restored and enhanced the power and prestige of Umayyad rule in Spain and beyond.

Early years of ʿAbd ar-Raḥmān's reign

ʿAbd ar-Raḥmān was born in January 891. He succeeded his grandfather ʿAbd Allāh as amir (prince) of Córdoba in October 912 at the age of 21. Because of his intelligence and character he had been the obvious favourite of his grandfather, who had designated him heir presumptive in preference to the other royal princes. In appearance he is described as having been light-skinned, handsome, thickset, and short-legged. He appeared to be very short when he walked but was imposing on horseback.

Public homage was paid to him in Córdoba immediately after his accession. He set about at once and with great energy to restore the authority of Córdoba in Spain—an authority that had been curtailed during the latter years of the reign of his grandfather by a host of rebels entrenched in mountain forts throughout the land. Ten days after his accession he had the head of the first rebel exhibited in Córdoba. Thereafter, for a score or so of years, he led almost annual expeditions against the rebels, first in southern and later in central and eastern Spain.

His greatest enemy was a crypto-Christian rebel, ʿUmar ibn Ḥafṣūn, lord of Bobastro. ʿAbd ar-Raḥmān's strategy was one of continuous harassment of Ibn Ḥafṣūn's forts. Beginning with the campaign of Monteleón, ʿAbd ar-Raḥmān captured 70 forts in the provinces of Elvira, Granada, and Jaén—all of which had been directly or indirectly controlled by Ibn Ḥafṣūn. In 913 Seville was captured, followed by Algeciras, Rayyu, Sidonia, and Carmona. When Ibn Ḥafṣūn died in 917, the rebellion collapsed. His children were captured or killed, and the centre of the rebellion, Bobastro, was finally stormed in 928. In 933 Toledo fell after a bitter siege, and, with its fall, the last Muslim centre of resistance to Cordoban hegemony disappeared.

Campaigns against the Christian north

Meanwhile, ʿAbd ar-Raḥmān also had to check threats from the Christian north. The main danger came from the Kingdom of Leon. An expedition commanded by Ordoño II, then vassal king of Galicia and later king of Leon, into Muslim territory in the summer of 913, especially his sack of Evora (Talavera) and the massacre of its Muslim population, produced widespread resentment in Muslim Spain. ʿAbd ar-Raḥmān decided to counterattack, which he began in earnest in 920, leading the campaign of Muez in person. He captured the forts of Osma and San Esteban de Gormaz and then inflicted a crushing defeat on the combined armies of Leon and Navarre at Valdejunquera on July 26, 920. Four years later, in the spring of 924, he led another campaign into Navarre and sacked the capital, Pamplona. With these two campaigns, ʿAbd ar-Raḥmān was able to secure his frontiers with Christian Spain for the next seven years. But the next King of Leon, Ramiro II, who ascended the throne in 932, proved a formidable adversary and began immediately to mount attacks against Muslim territory. The encounter between the two rulers finally took place in 939, when, at the "ditch" of Simancas, Ramiro severely defeated the Muslims and ʿAbd ar-Raḥmān barely escaped with his life. After that defeat ʿAbd ar-Raḥmān resolved never to personally lead another expedition. The Christian victory, however, was not followed up. When Ramiro died in 950 and civil war broke out in the Christian territories, ʿAbd ar-Raḥmān made good his earlier losses so thoroughly that in 958 Sancho, exiled king of Leon, Garcia Sánchez, king of Navarre, and his mother, Queen Toda, all paid personal homage to ʿAbd ar-Raḥmān in Córdoba.

In North Africa the policy of ʿAbd ar-Raḥmān was directed against the Fāṭimids in al-Qayrawān (now in modern Tunisia). In order to check their control over North Africa he financed rebellions against them and sent naval expeditions to sack the coastal cities. The city of Ceuta was fortified in 931 as a base of operations in North Africa. Toward the end of his reign, however, Fāṭimid power increased, and the Fāṭimid general Jawhar was able to repulse the allies of ʿAbd ar-Raḥmān. The struggle with the Fāṭimids, however, was inconclusive and was to continue throughout the 10th century.

Significance of the man and his reign

As a result of his early successes, and probably at his own suggestion, some of his court poets urged ʿAbd ar-Raḥmān to adopt the title of caliph. He assumed that dignity in 929, shortly after the fall of Bobastro, and chose the honorific title an-Nāṣīr li-Dīn Allāh (He Who Fights Victoriously for the Religion of God). His reasons were, internally, to enhance his prestige and, externally, to counter the Fāṭimid claim to this honour. The consolidation of power brought great prosperity to Muslim Spain—one indication of which was his building of a mint where pure gold and silver coins were struck. ʿAbd ar-Raḥmān was also a great builder; he renovated and added considerably to the Great Mosque at Córdoba and to the royal palace. At vast expense he built a new royal city, Madīnat az-Zahrā', to house his household and government. He kept a very strict control over the affairs of state and his civil service, changing his governors frequently to avoid the growth of local dynasties. In 949 he executed his own son for conspiring against him.

Christian and Jewish communities flourished during his tolerant reign. His fame spread so far beyond his domains that Córdoba by the end of his reign enjoyed almost as much fame as Constantinople in the Mediterranean world. In Córdoba he received emissaries from such distant rulers as Otto I of Germany and the Byzantine emperor. Córdoba was said to have contained 3,000 mosques and more than 100,000 shops and houses. His reign, the second longest of any Muslim caliph, afforded his wise and courageous policies the fullest chance of development. ʿAbd ar-Raḥmān died in October 961, at the height of his power and fame.

BIBLIOGRAPHY. The best account of ʿAbd ar-Raḥmān's reign is in E. LEVI-PROVENCAL, *Histoire de l'Espagne musulmane*, vol. 2, pp. 1–164 (1950). Also useful are W.M. WATT, *A History of Islamic Spain* (1965); and S. LANE-POOLE, *The Story of the Moors in Spain* (1886; reprinted as *The Moors in Spain*, 1967).

(T.Kh.)

Abdelkader

Abdelkader (ʿAbd al-Qādir), born September 6, 1808, the son of the head of the Qādirīyah Muslim brotherhood Mahieddin (Muḥyī ad-Dīn), became a well-known military figure after the French occupation of the Algerian coast, leading the Algerians in their battle against the invading French until 1847. A political leader who founded a modern state, and a religious leader as well, Abdelkader embodied the resistance and growing nationalism of the Algerian people.

His physical handsomeness and the qualities of his mind had made him popular even before his military exploits. Of medium height, lithe and elegant, with regular features and a black beard, his demeanour was exceptionally refined, and his life-style simple. He was known as a religious and educated man, who could excite his coreligionists with his poetry and oratorical eloquence.

Algeria was an Ottoman regency when the French army landed there in 1830. The government was controlled by a dey (governor) and by the Turkish Janissaries who had chosen him. These rulers, supported by the Koulouglis (people of mixed Turkish and Algerian ancestry) and by certain privileged tribes, and aided by the fact that they

were of the same religion as the people, long held Algeria firmly in their grip.

Nevertheless, the Algerians detested them, and there were continual rebellions in the early 19th century. As a result, the country was left too divided to oppose the Christian invaders.

The western tribes laid siege to Oran and tried to organize themselves, unified by their common Muslim religious sentiment, which was cultivated by the schoolmasters and particularly by members of the religious brotherhoods. The leader of one of the brotherhoods, Mahieddin, director of the *zāwiyah* (religious school) near Mascara, was asked to lead the harassment of the French troops in Oran and Mostaganem.

In November 1832, the aging Mahieddin had his young son Abdelkader elected in his place. The youth, already renowned for his piety and military prowess, took over the war of harassment. The ensuing Desmichels Treaty of 1834 gave him the whole interior of the Oran, with the title commander of the believers. In a move to unify his new territories, Emir Abdelkader, taking advantage of this treaty, imposed his rule on all the tribes of the Chelif, occupied Miliana and then Médéa, and succeeded in defeating Gen. Camille Trézel at Macta. Although pressed by generals Bertrand Clauzel and T.R. Bugeaud, he managed to rally support from Algerians who had become indignant over the French use of violence. By able negotiation, he convinced General Bugeaud to sign the Treaty of Tafna (1837), which further increased his territory and made him master of the whole interior of the provinces of Oran and the Titteri, with the French having to be content with a few ports.

Creation of a new state

In two years the Emir had organized a true state, the capital of which was sometimes Mascara and sometimes the fortress of Tiaret, (now Tagdempt). He established juridical equality among population groups by suppressing the privileges of the warlike tribes (*makhzen*) and by imposing equal taxes on all his subjects. First he extended his influence to the Sahara by fighting at-Tijīnī, who dominated the southern oases, and rallying the desert peoples to him. Then he strengthened his authority in the valley of the Chelif and in the Titteri as far as the borders of the province of the east, where he was resisted by the *bey* of Constantine, Hajj Ahmed. He also exacted harsh punishment of the Koulouglis of Zouatna, who had joined the French. By the winter of 1838, his authority extended across the borders of Kabylie and, in the south, from the oasis of Biskra to the Moroccan border. To destroy the power of at-Tijīnī, he besieged his capital, Aïn Mahdi, for six months and demolished it, while all the Saharan tribes paid him homage.

Abdelkader was an absolute leader, who only rarely called in the grandees to advise him. Algerian religious sentiment was his support, the one force that could bring his subjects together and unify them in the face of the invader. But that did not prevent him from employing competent persons of all nationalities, whether Jews or Christians, to help him build a modern state. The best known of these Europeans was the future diplomat Léon Roches, who later recounted his adventures in a fanciful book, *Trente-deux ans à travers l'Islam* ("Thirty-two Years Through Islām"). Abdelkader organized a regular army of approximately 2,000 men, to be supported by either volunteers or contingents furnished by the tribes. As towns near French territory would have been too vulnerable, he fortified interior sites, such as Sebdou, Saïda, Tiaret, Taza, and Boghar, where he opened arsenals, warehouses, and workshops, and where he stored surplus crops whose sales were to finance his arms purchases, mainly in England. He set up a new administration, with officials on fixed salaries. He taught his people austerity and set a personal example, living without ceremony in a tent. By expanding education, he slowly spread the concepts of independence and nationality to his people.

When the columns of the Duc d'Orléans crossed the Iron Gates, the Emir took it as a violation of the territories granted him by the Treaty of Tafna. Even though he was still far from having completed his own work of organization, he made a surprise attack and destroyed the French colonization of the Mitidja Plain. From then on the war languished until General Bugeaud was named governor general in 1840. Bugeaud convinced the French government to arm him for the conquest of all Algeria. The resulting war was bitter and lasted seven years. The Emir avoided big battles, preferring to use his rifle-armed cavalry in incessant skirmishes, from which it would retreat almost as soon as it had fired. But he was fighting a French army composed of infantry that Bugeaud had organized into extremely mobile columns, and he had to contend with the ruinous devastation of the countryside practiced by Bugeaud and his lieutenants so as to force the starving inhabitants to desert their leader.

Responses to the French policy of conquest

In 1841 the French destroyed the Emir's fortified sites, and he was forced to become a nomad in the interior of Oran. The following year he lost Tlemcen, and communication with his Moroccan allies became difficult. Yet, despite further reverses and French penetration in the south, he succeeded in reaching Morocco. But after Bugeaud's defeat of the Moroccans at Isly, the Sultan was forced to hold Abdelkader in the midst of his empire. The Emir, however, proved to have unflagging energy. Taking advantage of a revolt in the Dahra, he re-entered Algeria, took the Sidi Brahim outpost, and penetrated deep into the interior, all the while escaping the pursuing French columns. In July 1846, with only a handful of men left, Abdelkader again took refuge in Morocco, the sultan of which by then considered him to be a burden. Deprived of this last area of support, Abdelkader returned to Algeria and with great dignity turned himself over to Gen. Christophe de Lamoricière and to Bugeaud's successor, King Louis-Philippe's son, the Duc d'Aumale, who promised him transport to the east.

Surrender and later life

Louis-Philippe, however, failed to respect his son's promise. Abdelkader was held prisoner in France, first at the Château de Pau, where he learned the principles of Freemasonry, and later at Amboise. It was the prince-president Louis-Napoléon who, in 1852, authorized his return to Bursa, and then to Damascus, where he led an exemplary life and wrote *Rappel à l'intelligent, avis à l'indifférent* ("Call to the Intelligent, Warning to the Indifferent"). The French government provided him with a large pension and with a Kabyle guard, and even attempted to obtain a throne for him somewhere between Turkey and Egypt, which they wished to remove from Ottoman control. At the time of the 1871 Algerian insurrection, he disowned one of his sons who had tried to arouse the tribes of southern Constantine.

When he died on May 26, 1883, he was respected by all. French efforts to make him the symbol of Algerian support for colonial rule were erroneous. Abdelkader believed he was carrying out God's will in admitting that his political role had ended. Present-day Algerians consider him to be the greatest hero of their people.

BIBLIOGRAPHY. PAUL AZAN, *L'Émir Abd el Kader, 1808–1883, du fanatisme musulman au patriotisme français* (1925), the most complete work, but currently out of date; MARCEL EMERIT, *L'Algérie à l'époque d'Abd-el-Kader* (1951), documents and commentaries; RENE GALLISSOT, "Abd el-Kader et la nationalité algérienne," *Revue Historique*, 233:339–368 (1965); CHARLES HENRY CHURCHILL, *The Life of Abdel Kader, Ex-Sultan of the Arabs of Algeria* (1867), the only biography currently available in English.

(M.E.)

ʿAbduh, Muḥammad

An Egyptian religious scholar and liberal reformer, Muḥammad ʿAbduh laid the foundations for a generation of Muslim intellectuals of the late 19th and early 20th centuries who sought to revitalize Islāmic teachings and institutions and to reconcile them with the demands of modern civilization. ʿAbduh's achievement was to stimulate the conviction among the new secularly educated classes of Egypt and other Muslim countries that their style of life was compatible with religious faith and that Islāmic principles, properly understood, required vigorous efforts to improve the conditions of their society.

ʿAbduh was born in 1849 in a village of the Nile Delta. He attended the mosque school in Ṭanṭā and subse-

quently al-Azhar University in Cairo, receiving the degree of *ʿālim* (scholar) from the latter in 1877. After an early infatuation with mysticism, in 1872 he fell under the influence of Jamāl ad-Dīn al-Afghānī, the revolutionary pan-Islāmic preacher of Persian origin who had settled in Cairo and who stimulated ʿAbduh's interest in theology, philosophy, and politics. In punishment for political activity, Afghānī was expelled from Egypt in 1879 and ʿAbduh was exiled to his village, but the next year ʿAbduh's fortunes changed. He became editor of the government's official gazette, which he made a platform for preaching resistance to Anglo-French political encroachment and the need for social and religious reform. He was implicated in ʿUrābī Pasha's rebellion against foreign control in 1882 and, following the British military occupation of Egypt, was exiled. Rejoining Afghānī in Paris for several months in 1884, ʿAbduh helped his mentor publish the revolutionary journal *al-ʿUrwat al-wuthqā* ("The Firmest Bond"), which was smuggled to Egypt, India, and elsewhere. After brief visits to England and Tunisia, ʿAbduh settled for three years in Beirut and taught in an Islāmic college there.

Early career and exile

In 1888 ʿAbduh was permitted to return to Egypt, where he began a judicial career that spanned the rest of his life. He was appointed a judge in the National Courts of First Instance, then in 1891 at the Court of Appeal; in 1899, with British help, he became *muftī* (Islāmic legal counsellor) of Egypt. In the latter post he effected several reforms in the administration of Islāmic law and of religious endowments. He also issued advisory opinions to private petitioners, including such controversially liberal judgments as the permissibility of eating meat slaughtered by Christian and Jewish butchers and of accepting interest paid on loans. ʿAbduh also lectured at al-Azhar and, against much conservative opposition, induced reforms in the administration and curriculum of that ancient institution. He established a benevolent society that operated schools for poor children. He served on the Legislative Council, preaching political cooperation with Britain and a long-term effort to bring about legal and educational reforms in Egypt; these views, differing markedly from those he had espoused earlier in life under Afghānī's influence, earned him the approval of Lord Cromer, the British Resident, but also the hostility of the khedive (ruling prince) ʿAbbās Ḥilmī and of the nationalist leader Muṣtafā Kāmil. Late in life ʿAbduh learned French and pursued an interest in European thought.

Judicial career

In addition to his numerous articles in the official gazette and *al-Urwat al-wuthqā*, ʿAbduh's most important writings included *Risālat at-tawḥīd* ("Treatise on the Oneness of God"); a polemic on the superiority of Islām to Christianity in its inherent receptivity to science and civilization; and a fragmentary commentary on the Qurʾān, completed after his death by a disciple. In theology ʿAbduh sought to establish the harmony of reason and revelation, the freedom of the will, and the primacy of the ethical implications of religious faith over ritual and dogma. He deplored the blind acceptance of traditional doctrines and customs and asserted that a return to the pristine faith of the earliest age of Islām would not only restore the Muslims' spiritual vitality but would provide an enlightened criterion for the assimilation of modern scientific culture.

Religious and philosophical beliefs

In matters of Islāmic law, which governed Muslim family relationships, ritual duties, and personal conduct, ʿAbduh tried to break through the rigidities of scholastic interpretation and to promote considerations of equity, welfare, and common sense, even if this occasionally meant disregarding the literal texts of the Qurʾān. From his death to the present day ʿAbduh has been widely revered as the chief architect of the modern reformation of Islām. ʿAbduh died on July 11, 1905, in a suburb of Alexandria.

BIBLIOGRAPHY. The only full-length biography of ʿAbduh is in Arabic: MUHAMMAD RASHID RIDA, *Tarīkh al-Ustādh al-Imām ash-Shaikh Muḥammad Abduh*, vol. 1 (1931). The most extensive western-language treatment is CHARLES C. ADAMS' classic, *Islam and Modernism in Egypt* (1933), which also deals at length with Ridā and other followers of ʿAbduh. A substantial analysis of ʿAbduh's thought and legacy is found in MALCOLM H. KERR, *Islamic Reform: The Political and Legal Theories of Muhammad ʿAbduh and Rashīd Ridā* (1966); ELIE KEDOURIE, *Afghani and ʿAbduh* (1966), portrays ʿAbduh as a clandestine religious and political subversive. An authoritative summary of ʿAbduh's life and work is by JOSEPH SCHACHT in *The Shorter Encyclopaedia of Islam* (1953).

(M.H.K.)

Abel, Niels Henrik

Niels Henrik Abel was a Norwegian mathematician, recognized only after his death as a pioneer in the development of modern mathematics.



Abel, lithograph after a drawing by Johan Gørbitz, 1826.
By courtesy of the Royal Norwegian Embassy, Washington, D.C.

Abel was born on August 5, 1802, on the island of Finnøy, near Stavanger, Norway, where his father was a poor Protestant minister. The family soon moved to the parish of Gjerstad, near the town of Risør (southeast Norway), where the boy grew up. In 1815, when he entered the cathedral school in Oslo, his mathematical talent was recognized by a teacher who introduced him to the classics in mathematical literature and proposed original problems for solution. Thoroughly challenged, Abel studied the works of the 17th-century English mathematician and physicist Isaac Newton and the contemporary mathematicians Leonhard Euler (German), Joseph-Louis Lagrange (French), and Carl Friedrich Gauss (German) and learned to detect gaps in their mathematical reasoning.

Early recognition of his talent

Although when Abel's father died in 1820 the family was left in straitened circumstances, the boy was able to enter the University of Christiania (Oslo) in 1821 because his teacher contributed and raised funds.

On graduation from the University in 1822, Abel continued his studies with further subsidies obtained by his teacher. His first papers, published in 1823 in the new periodical *Magazin for Naturvidenskaberne*, were on functional equations and integrals, his solution of an integral equation being the first. Abel's friends urged the Norwegian government to grant him a fellowship for study in Germany and France. While waiting for the royal decree to be issued, in 1824 he published at his own expense his proof of the impossibility of solving algebraically the general equation of the fifth degree, which he hoped would bring him recognition. He sent the pamphlet to Gauss, who dismissed it, failing to recognize that the famous problem had indeed been settled.

Abel spent the winter of 1825–26 with Norwegian friends in Berlin, where he met August Leopold Crelle, civil engineer and self-taught enthusiast of mathematics, who became his close friend and mentor. With Abel's warm encouragement, Crelle founded the *Journal für die reine und angewandte Mathematik* ("Journal for Pure and Applied Mathematics"), the first volume of which (1826)

contains papers by Abel, including a more elaborate version of his work on the quintic equation. Other papers dealt with equation theory, functional equations, integration in finite forms, and problems from theoretical mechanics.

Abel's early mathematical training had been in the formal school typified by Euler. In Berlin new directions in mathematics stimulated him to do further independent work. Soon distracted socially, however, Abel travelled throughout Europe.

Arriving in Paris in the summer of 1826, he called on the foremost mathematicians and completed a memoir on transcendental functions. In this major work he presented a theory of integrals of algebraic functions, in particular the result known as Abel's theorem: there is a finite number, or genus, of independent integrals of this nature. This theorem forms the basis for the later theory of Abelian integrals and Abelian functions. Abel was accepted with restrained civility in Paris, for his work was still unknown. He submitted his memoir for presentation to the Academy of Sciences, hoping that it would establish his reputation; but he waited in vain. Before leaving Paris thinking he had a persistent cold, Abel consulted a physician, who informed him he had tuberculosis.

Abel's theorem

Abel returned to Norway heavily in debt. He subsisted by tutoring, by receiving a small grant from the university, and, in 1828, by accepting a substitute teaching position. His poverty and ill health did not decrease his production; he wrote a great number of papers, principally on equation theory and elliptic functions. Among them are the theory of the Abelian equations with Abelian groups. He rapidly developed the theory of elliptic functions in competition with Karl Gustav Jacobi.

By this time Abel's fame had spread to all mathematical centres, and strong efforts were made to secure a suitable position for him by a group from the French academy, who addressed Bernadotte, the king of Norway–Sweden; Crelle worked to secure a professorship for him in Berlin.

In the fall of 1828, Abel became seriously ill, and his condition deteriorated on a sled trip at Christmas time to visit his fiancée at Froland, where he died on April 6, 1829. The French Academy of Sciences published his memoir in 1841.

BIBLIOGRAPHY. The *Oeuvres complètes de N.H. Abel* (1839; new ed., 2 vol., 1881) contains all his works. For biographical discussions, see OYSTEIN ORE, *Niels Henrik Abel, Mathematician Extraordinary* (1957), with bibliography; ERIC T. BELL, *Men of Mathematics* (1937, reprinted 1961); and GANESH PRASAD, *Some Great Mathematicians of the Nineteenth Century*, vol. 1 (1933). A translation of a portion of Abel's memoir on the quintic is contained in DAVID E. SMITH, *A Source Book in Mathematics* (1929). See also C.A. BJERKNES, *Niels Henrik Abel: Mémorial publié à l'occasion du centenaire de sa naissance* (1902).

(Ed.)

Abelard, Peter

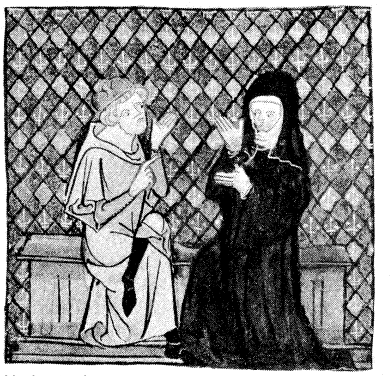
Peter Abelard, the most famous and the most controversial teacher of his age, was a logician, moral philosopher, and theologian. He was also a gifted and original poet, the founder of a famous convent for women, and the tragic lover of Héloïse. He was twice condemned for heresy by ecclesiastical councils in France.

Early life

The outline of Abelard's career is well-known, largely because he described so much of it in his famous *Historia calamitatum* ("History of My Troubles"). He was born the son of a knight in 1079 at Le Pallet in Brittany south of the Loire River. He sacrificed his rights of inheritance and the prospect of a military career in order to study philosophy, particularly logic, in France. He provoked bitter quarrels with two of his masters, Roscelin of Compiègne and William of Champeaux. The two men represented opposite poles of philosophy. Roscelin was a Nominalist who asserted that universals are nothing more than mere words; William in Paris upheld a form of Platonic Realism according to which universals exist. Abelard in his own logical writings brilliantly elaborated an independent philosophy of language. While showing how words could be used significantly, he stressed that language itself is not able to demonstrate the truth of things (*res*) that lie in the domain of physics.

Abelard was a peripatetic both in the manner in which he wandered from school to school at Paris, Melun,

By courtesy of the Musee Conde, Chantilly, France; photograph, Giraudon



Abelard, with Héloïse, miniature portrait by Jean de Meun, 14th century. In the Musée Condé, Chantilly, France.

Corbeil and elsewhere and as one of the exponents of Aristotelian logic who were called the Peripatetics. In 1113 or 1114 he went north to Laon to study theology under Anselm of Laon, the leading biblical scholar of the day. He quickly developed a strong contempt for Anselm's teaching, which he found vacuous, and returned to Paris. There he taught openly but was also given as a private pupil the young Héloïse, niece of one of the clergy of the cathedral of Paris, Canon Fulbert. Abelard and Héloïse fell in love and had a son whom they called Astralabe. They then married secretly. To escape her uncle's wrath Héloïse withdrew into the convent of Argenteuil outside Paris. Abelard suffered castration at Fulbert's instigation. In shame he embraced the monastic life at the royal abbey of Saint-Denis near Paris and made the unwilling Héloïse become a nun at Argenteuil.

Career as a monk

At Saint-Denis Abelard extended his reading in theology and tirelessly criticized the way of life followed by his fellow monks. His reading of the Bible and of the Fathers of the Church led him to make a collection of quotations that seemed to represent inconsistencies of teaching by the Christian Church. He arranged his findings in a compilation entitled *Sic et Non* ("Yes and No"); and for it he wrote a preface in which, as a logician and as a keen student of language, he formulated basic rules with which students might reconcile apparent contradictions of meaning and distinguish the various senses in which words had been used over the course of many centuries. He also wrote the first version of his book called *Theologia*, which was formally condemned as heretical and burned by a council held at Soissons in 1121. Abelard's dialectical analysis of the mystery of God and the Trinity was held to be erroneous, and he himself was placed for a while in the abbey of Saint-Médard under house arrest. When he returned to Saint-Denis he applied his *Sic et Non* methods to the subject of the abbey's patron saint; he argued that St. Denis of Paris, the martyred apostle of Gaul, was not identical with Denis of Athens (also known as Dionysius the Areopagite), the convert of St. Paul. The monastic community of Saint-Denis regarded this criticism of their traditional claims as derogatory to the kingdom; and, in order to avoid being brought for trial before the King of France, Abelard fled from the abbey and sought asylum in the territory of Count Theobald of Champagne. There he sought the solitude of a hermit's life but was pursued by students who pressed him to resume his teaching in philosophy. His combination of the teaching of secular arts with his profession as a monk was heavily criticized by other men of religion, and Abelard contemplated flight outside Christendom altogether. In 1125, however, he accepted election as abbot of the remote Breton monastery of Saint-Gildas-de-

Rhuys. There, too, his relations with the community deteriorated, and, after attempts had been made upon his life, he returned to France.

Héloïse had meanwhile become the head of a new foundation of nuns called the Paraclete. Abelard became the abbot of the new community and provided it with a rule and with a justification of the nun's way of life; in this he emphasized the virtue of literary study. He also provided books of hymns he had composed, and in the early 1130s he and Héloïse composed a collection of their own love letters and religious correspondence.

Final years

In about 1135 Abelard went to the Mont-Sainte- Geneviève outside Paris to teach and to write in a blaze of energy and of celebrity. He produced further drafts of his *Theologia* in which he analyzed the sources of belief in the Trinity and praised the pagan philosophers of classical antiquity for their virtues and for their discovery by the use of reason of many fundamental aspects of Christian revelation. He also wrote a book called *Ethica* or *Scito te ipsum* ("Know Thyself"), a short masterpiece in which he analyzed the notion of sin and reached the drastic conclusion that human actions do not make a man better or worse in the sight of God, for deeds are in themselves neither good nor bad. What counts with God is a man's intention; sin is not something done (it is not *res*); it is uniquely the consent of a human mind to what it knows to be wrong. Abelard also wrote *Dialogus inter philosophum, Judaeum et Christianum* ("Dialogue between a Philosopher, a Jew, and a Christian"), and a commentary on St. Paul's letter to the Romans, the *Expositio in Epistolam ad Romanos*, in which he outlined an explanation of the purpose of Christ's life, which was to inspire men to love him by example alone.

On the Mont-Sainte-Geneviève Abelard drew crowds of pupils, many of them men of future fame, such as the English humanist John of Salisbury. He also, however, aroused deep hostility in many by his criticism of other masters and by his apparent revisions of the traditional teachings of Christian theology. Within Paris the influential Abbey of Saint-Victor was studiously critical of his doctrines, while elsewhere William of Saint-Thierry, a former admirer of Abelard, recruited the support of Bernard of Clairvaux, perhaps the most influential figure in Western Christendom at that time. At a council held at Sens in 1140, Abelard underwent a resounding condemnation, which was soon confirmed by Pope Innocent II. He withdrew to the great monastery of Cluny in Burgundy. There, under the skillful mediation of the abbot, Peter the Venerable, he made peace with Bernard of Clairvaux and retired from teaching. Now both sick and old, he lived the life of a Cluniac monk. His death at the nearby priory of Saint-Marcel at Chalon-sur-Saône probably occurred in 1144. His body, first sent to the Paraclete, now lies alongside that of Héloïse in the cemetery of Père-Lachaise in Paris. Epitaphs composed after his death suggest that Abelard impressed some of his contemporaries as one of the greatest thinkers and teachers of all time.

BIBLIOGRAPHY. *The Letters of Abelard and Héloïse have* been translated into English by C.K. SCOTT MONCRIEFF (1925, reprinted 1964) and by J.T. MUCKLE (1947), who has also published separately a translation of *The Story of Abelard's Adversities* (1954). The historical novel *Peter Abelard*, by HELEN WADDELL (1933), is justly popular and has been reprinted many times. More general studies are ETIENNE GILSON, *Héloïse et Abélard* (1938; Eng. trans., 1953); and LIEF GRANE, *Peter Abélard* (1964; Eng. trans., 1970). Information concerning editions of Abelard's writings and studies of his thought may be found in D.E. LUSCOMBE's edition and English translation of *Peter Abelard's "Ethics"* (1971). On the reception of Abelard's thought and teaching in his own time, see D.E. LUSCOMBE, *The School of Peter Abelard* (1969). The best study of Abelard's philosophy is JEAN JOLIVET, *Arts du langage et théologie chez Abélard* (1969).

(D.E.L.)

Abraham

Abraham, the first of the Hebrew patriarchs, is the father of faith for the three great monotheistic religions: Judaism, Christianity, and Islām. He is a central figure in the biblical accounts as the recipient of a divine call, revelation, and promise; the prototype of the faithful man tested by God; and the founding ancestor of the chosen people, the first of the patriarchs of Israel. He is presented as a seer or prophet and the paradigmatic "friend of God."

The critical problem of a "biography" of Abraham. There can be no biography of Abraham in the ordinary sense. The most that can be done is to apply the interpretation of modern historical finds to biblical materials so as to arrive at a probable judgment as to the background and patterns of events in his life. This involves a reconstruction of the patriarchal age (of Abraham, Isaac, Jacob, and Joseph; early 2nd millennium BC), which until the end of the last century was unknown and considered virtually unknowable. It was assumed, based on a presumed dating of hypothetical biblical sources, that the patriarchal narratives in the Bible were only a projection of the situation and concerns of a much later period (9th–5th centuries BC) and of dubious historical value. Several theses were advanced to explain the narratives; *e.g.*, that the patriarchs were mythical beings or the personifications of tribes or folkloric or etiological (explanatory) figures created to account for various social, juridical, or cultic patterns. However, after World War I, archaeological research made enormous strides with the discovery of monuments and documents, many of which date back to the period assigned to the patriarchs in the traditional account. The excavation of a royal palace at Mari, an ancient city on the Euphrates, for example, brought to light thousands of cuneiform tablets (official archives and correspondence and religious and juridical texts) and thereby offered exegesis a new basis, which specialists utilized to show that in the biblical book of Genesis, narratives fit perfectly with what, from other sources, is known today of the early 2nd millennium BC, but imperfectly with a later period. A biblical scholar in the 1940s aptly termed this result "the rediscovery of the Old Testament."

Thus, there are two main sources for reconstructing the figure of father Abraham: the book of Genesis, from the genealogy of Terah, Abraham's father, and his departure from Ur to Harran in chapter 11 to the death of Abraham in chapter 25; and recent archaeological discoveries and interpretations, concerning the area and era in which the biblical narrative takes place.

The biblical account. According to the biblical account, Abram (The Father [or God] Is Exalted), who is later named Abraham (The Father of Many Nations), a native of Ur in Mesopotamia, is called by God (Yahweh) to leave his own country and people and journey to an undesignated land, where he will become the founder of a new nation. He obeys the call unquestioningly and (at 75 years of age) proceeds with his barren wife, Sarai, later named Sarah (Princess), his nephew Lot, and other companions to the land of Canaan (between Syria and Egypt). There the childless septuagenarian receives repeated promises and a covenant from God that his "seed" will inherit the land and become a numerous nation. He not only has a son, Ishmael, by his wife's maidservant Hagar but, at 100 years of age, by Sarah, a legitimate son, Isaac, who is to be the heir of the promise. Yet, he is ready to obey God's command to sacrifice Isaac, a test of his faith, which he is not required to consummate in the end because God substitutes a ram. At Sarah's death, he purchases the cave of Machpelah near Hebron, together with the adjoining ground, as a family burying place. It is the first clear ownership of a piece of the promised land by Abraham and his posterity. Toward the end of his life, he sees to it that his son Isaac marries a girl from his own people back in Mesopotamia rather than a Canaanite woman. He dies at the age of 175 and is buried next to Sarah in the cave of Machpelah.

Personal characteristics

Abraham is pictured with various characteristics: a righteous man, with wholehearted commitment to God, a man of peace (in settling a boundary dispute with his nephew Lot), a quick-acting warrior (he rescues Lot and his family from a raiding party), compassionate (he argues and bargains with God to spare the people of Sodom

and Gomorrah), hospitable (he welcomes three visiting angels), and an unscrupulous liar to save his own skin (he passes off Sarah as his sister and lets her be picked by the Egyptian pharaoh for his harem). He appears as both a man of great spiritual depth and strength and a person with common human weaknesses and needs.

The Genesis narrative in the light of recent scholarship. The saga of Abraham unfolds between two landmarks, the exodus from "Ur of the Chaldeans" (Ur Kasdim) of the family, or clan, of Terah and "the purchase of" (or "the burials in") the cave of Machpelah. Tradition seems particularly firm on this point. The Hebrew text, in fact, locates the departure specifically at Ur Kasdim, the Kasdim being none other than the Kaldu of the cuneiform texts at Mari. It is manifestly a migration of which one tribe is the centre. The leader of the movement is designated by name: Terah, who "takes them out" from Ur, Abram his son, Lot the son of Haran, another son of Terah, and their wives, the best known being Sarai, the wife of Abram. The existence of another son of Terah, Nahor, who appears later, is noted.

Ur Kasdim seems to have been the Sumerian city of Ur, today Tall al-Muqayyar (or Mughair), about 200 miles (300 kilometres) southeast of Baghdad in Lower Mesopotamia, which was excavated with great success from 1922 to 1934. The identification is disputed by some scholars, but most scholars continue to locate Ur of the Chaldeans at Tall al-Muqayyar, and therefore in the Sumerian city of Ur, capital of the 3rd dynasty of Ur (*c.* 2112–*c.* 2004 BC), which the patriarchal tribe left either at that time or soon after, for reasons that are impossible to determine. At any rate, it is certain that the cradle of the ancestors was the seat of a vigorous polytheism whose memory had not been lost and whose uncontested master in Ur was Nanna (or Sin), the Sumero-Akkadian moon god. "They served other gods," Joshua, Moses' successor, recalled, speaking to their descendants at Shechem.

After the migration from Ur, the first important stopping place was Harran, where the caravan remained for some time. The city has been definitely located in Upper Mesopotamia, between the Tigris and the Euphrates, in the Balikh Valley, and can be found on the site of the modern Harran in Turkey. It has been shown that Harran was a pilgrimage city, for it was a centre of the Sin cult and consequently closely related to the moon god cult of Ur. The Mari tablets have shed new light on the patriarchal period, specifically in terms of the city of Harran.

The Hebrews and Benjaminites in the Mari tablets

There have been many surprising items in the thousands of tablets found in the palace at Mari. Not only are the Ḫapiru ("Hebrews") mentioned but so also remarkably are the Banu Yamina ("Benjaminites"). It is not that the latter are identical with the family of Benjamin, a son of Jacob, but rather that a name with such a biblical ring appears in these extrabiblical sources in the 18th century BC. What seems beyond doubt is that these Benjaminites (or Yaminites, meaning "Sons of the Right," or "Sons of the South," according to their habits of orientation) are always indicated as being north of Mari and in Harran, in the temple of Sin. The patriarchs at Harran noted in chapter 11 of Genesis and the Benjaminites at Harran mentioned in the Mari text provide a perfect concordance. There are also other parallels: these Benjaminites were in the habit of using fire beacons, which their descendants were to employ much later in Israel. They are pillagers and raiders like the people of Jacob's blessing listed in chapter 49 of Genesis, in which the tribe of Benjamin is described exactly in this way. And in the tablets of Mari, there is frequent mention of "the city of Nakhur" (probably the equivalent of Nahor), of a Laba-an (Laban) and of several Aramu (Aramaeans). It is impossible not to be reminded that Laban, Jacob's father-in-law, is called "the Aramaean" several times in Genesis and that the nomadism of the ancestors applies just as well to the incessant movings of the tribes (the Haneans, Sutaeans, Ḫapiru, Benjaminites, and Benê-sim'âl) that abound in the Mari texts. All of this documentation leads so close to the image reflected in the book of Genesis that it is possible to understand the enthusiasm of William F. Albright, an eminent biblical archaeologist, when he writes: one can no longer doubt the substantial correctness of the Hebraic tradition which has Abraham's family come from the Harran region, between the Euphrates and Khabur. The portrait of the customs and relationships rings true in the light of the newly discovered sources.

From Harran toward Canaan

The Bible has given no information on the itinerary followed between Ur and Harran. Scholars think that the caravan went up the Euphrates, then up the Balikh. After a stay of indeterminate length in Harran, the Bible says only that Terah died there, at the age of 205, and that Abraham was 75 when he took up the journey again with his family and his goods. This time the migration went from east to west, first as far as the Euphrates, which they may have crossed at Carchemish, since it can be forded during low-water periods. Here again, the Mari texts supply a reference, for they indicate that there were Benjaminites on the right bank of the river, in the lands of Yamhad (Aleppo), Qatanum (Qatna), and Amurru. Since the ancient trails seem to have been marked with sanctuaries, it is noteworthy that Nayrab, near Aleppo, was, like Harran and Ur, a centre of the Sin cult and to note that a few kilometres south of Aleppo, on the road to Ḥamāh, there is still today a village that bears the name of Benjamin. The route is in the direction of the "land of Canaan," the goal of the journey. If the tribe made a slight detour then, since it was to settle in Sichem (Shechem) and pass on to Damascus, this would explain the presence of a foreigner, Eliezer, a native of Damascus, who became the "servant" of Abraham, a sort of majordomo or steward, having authority over Abraham's entire household. However, this is uncertain, for the text is unreliable, specifically on the subject of the origins of the steward, whose name, Eliezer (My God is Help), presents no difficulty.

If a stop in Damascus is assumed, the caravan must next have crossed the land of Bashan (the Ḥawrān of today), first crossing the Jabboq, then the Jordan at the ford of Dāmiyā, arriving in the heart of the Samaritan country, to reach at last the plain of Shechem, today Balāṭah, at the foot of the Gerizim and Ebal mountains. Shechem was at the time a political and religious centre, the importance of which has been perceived more clearly as a result of recent archaeological excavations. From the mid-13th to the mid-11th centuries BC, Shechem was the site of the cult of the Canaanite god Baʿal-Berit (Lord of the Covenant); at this site a magnificent *matzeva* (sacred stone pillar) was discovered. There was also the "oak of Moreh" mentioned in the account of Abraham's journey to Shechem, an essential element in the local Semitic cult. The architecture uncovered by archaeologists, at least that of the palace, would date back to the 18th century, the period in which the presence of the patriarchs in Shechem is placed.

Bethel, Hebron, and the oaks at Mamre

The next stopping place was in Bethel, identified with present-day Baytīn, north of Jerusalem. Bethel was also a holy city, whose cult was centred on El, the Canaanite god *par excellence*. Its name does not lend itself to confusion, for it proclaims that the city is the *bet*, "house" or temple, of El (God). The Canaanite sanctuary was taken over without hesitation by the patriarchs, who built their altar there and consecrated it to Yahweh, at least if the Yahwistic tradition in Genesis is to be believed. Certainly this memory remained particularly alive in the hearts of the Israelites. Bethel remained for centuries the sanctuary *par excellence* for them. Even more than Abraham, they associated Jacob with it when, on the road to Harran, he had the dream of the ladder. But if Jacob stopped to spend the night in this place, it was because the spot had been marked in the past by the presence of an ancestor, namely, Abraham.

The latter had not yet come to the end of his journey. Between Shechem and Bethel he had gone about 50 kilometres. It was about as far again from Bethel to Hebron, or more precisely to the oaks of Mamre, "which are at Hebron" (according to the Genesis account). The location of Mamre has been the subject of some indecision. At the present time, there is general agreement in setting it 1.5

miles (two kilometres) northwest of Hebron at Rāmat al-Khalīl, an Arabic name which means the Heights of the Friend, the friend (of God) being Abraham.

Mamre, where there was an oak (if not a grove of oaks), marked the site of Abraham's encampment, but this did not at all exclude episodic travels in the direction of the Negeb, to Gerar and Beersheba. Life was a function of the economic conditions of the moment, of pastures to follow and to find, and thus the patriarchs moved back and forth between the land of Canaan and the Nile delta. They remained shepherds and never became cultivators. It was in Mamre that Abraham received the revelation that his race would be perpetuated, and it was there that he learned that his nephew Lot had been taken captive. The latter is an enigmatic episode, an "erratic block" in a story in which nothing prepared the way for it. Suddenly, the life of the patriarch was inserted into a slice of history in which several important persons ("kings") intervene: Amraphel of Shinar, Arioch of Ellasar, Ched-or-laomer of Elam, and Tidal of Goiim. Scholars of previous generations tried to identify these names with important historical figures—*e.g.* Amraphel with Hammurabi of Babylon—but little remains today of these suppositions. The whole of chapter 14 of Genesis, in which this event is narrated, differs completely from what has preceded and what follows. It may be an extract from some historical annals, belonging to an unknown secular source, for the meeting of Melchizedek, king of Salem and priest of God Most High (El ʿElyon), and Abraham is impressive. The king-priest greets him with bread and wine on his victorious return and blesses him in the name of God Most High.

The meeting with Melchizedek, king-priest of Salem

In this scene, the figure of the patriarch takes on a singular aspect. How is his religious behaviour to be characterized? He swears by "the Lord God Most High"—*i.e.*, by both Yahweh and El ʿElyon. It is known that, on the matter of the revelation of Yahweh to man, the biblical traditions differ. According to what scholars call the Yahwistic source (J) in the Pentateuch (the first five books of the Bible), Yahweh had been known and worshipped since Adam's time. According to the so-called Priestly source (P), however, the name of Yahweh was revealed only to Moses. It may be concluded that it was probably El whom the patriarchs, including Abraham, knew. As noted before, in Mesopotamia the patriarchs worshipped "other gods." On Canaanite soil, they met the Canaanite supreme god, El, and adopted him, but only partially and nominally, bestowing upon him qualities destined to distinguish him and to assure his pre-eminence over all other gods. He was thus to become El ʿOlam (God the Everlasting One), El ʿElyon (God Most High), El Shaddai (God, the One of the Mountains), and El Roʾi (God of Vision). In short, the god of Abraham possessed duration, transcendence, power, and knowledge. This was not monotheism but monolatry (the worship of one among many gods), with the bases laid for a true universalism. He was a personal god too, with direct relations with the individual, but also a family god and certainly still a tribal god. Here truly was the "God of our fathers," who in the course of time was to become the "God of Abraham, Isaac, and Jacob."

It is not surprising that this bond of the flesh should still manifest itself when it came to gathering together the great ancestors into the family burial chamber, the cave of Machpelah. This place is venerated today in Hebron, at the Ḥaram al-Khalīl (Holy Place of the Friend), under the mosque, and seems to have retained all its mystery. Abraham, "the friend of God," was forevermore the depositary of the promise, the beneficiary of the Covenant, sealed not by the death of Isaac but by the sacrifice of the ram that was offered up in place of the child on Mount Moriah.

BIBLIOGRAPHY. A study of the patriarch Abraham is based on scriptural documentation and particularly on Genesis. One should therefore first consult a critical translation of the first book of the Old Testament as one finds in a basic work such as: EPHRAIM A. SPEISER (ed. and trans.), *Genesis: Introduction, Translation, and Notes*, "Anchor Bible" (1964). Also worthwhile are the atlases and albums that give illustrated introductions: L.H. GROLLENBERG, *Atlas van de Bijbel*, 3rd ed. (1955; Eng. trans., *Atlas of the Bible* 1956); and G.E. WRIGHT and F.V. FILSON (eds.), *The Westminster Historical Atlas to the Bible*, rev. ed. (1956). Articles under "Abraham" in several dictionaries on the Bible, such as *The Interpreter's Dictionary of the Bible* (1962) and *The Westminster Dictionary of the Bible* (1944) are also useful.

The world of the patriarchs has called forth numerous works, among which these take into account the most recent archaeological discoveries: H. CAZELLES, "Patriarches" in *Supplément au Dictionnaire de la Bible*, 7 (1961); WILLIAM F. ALBRIGHT, *The Biblical Period from Abraham to Ezra* (1963); H.H. ROWLEY, "Recent Discovery and the Patriarchal Age," *Bulletin of the John Rylands Library*, 32:44–79 (1949); ROLAND DE VAUX, "Les Patriarches hébreux et les découvertes modernes," *Revue Biblique* (1946–49); and *Les Institutions de l'Ancien Testament*, 2 vol. (1958–60; Eng. trans., *Ancient Israel: Its Life and Institutions*, 2nd ed., 1965). Centred on the city of Ur and the discoveries that were made there is SIR LEONARD WOOLLEY, *Abraham: Recent Discoveries and Hebrew Origins* (1936); some of his analogies must be rejected. On the texts gathered at Mari and dating from the 18th century BC, which are of very great importance for biblical exegesis, see G. DOSSIN, "Benjaminites dans les textes de Mari," *Mélanges syriens*, 2:981–996 (1939); and J.R. KUPPER, *Les Nomades en Mésopotamie au temps des rois de Mari* (1957). A. PARROT, *Abraham et son temps* (1962; Eng. trans., *Abraham and His Times*, 1968), presents a synthesis of the double archaeological and epigraphical documentation.

(A.Pa.)

Abrasives

Abrasives are sharp, hard materials used to wear away the surface of softer, less resistant materials. Included within the term are both natural and manufactured substances ranging from the relatively soft particles used in household cleansers and jeweller's polish to the hardest known material, the diamond. Abrasives are indispensable to the manufacture of nearly every product made today.

Abrasives are used in the form of grinding wheels, coated abrasives (sandpaper), honing stones, polishes, cutoff wheels, tumbling and vibratory mass-finishing media, sandblasting, pulpstones, ball mills, and still other tools and products. Only through the use of abrasives is industry able to produce the highly precise components and ultrasmooth surfaces required in the manufacture of automobiles, airplanes, and space vehicles, mechanical and electrical appliances, and machine tools.

HISTORY

The use of abrasives goes back to earliest man's rubbing of one hard stone against another to shape a weapon or a tool. The Bible mentions a stone called *shamir* that was very probably emery, a natural abrasive still in use today. Ancient Egyptian drawings show abrasives being used to polish jewelry and vases. A statue of a Scythian slave, called "The Grinder," in the Uffizi art gallery in Florence, shows an irregularly shaped natural sharpening stone used to whet a knife.

Sand and pieces of flexible hide were early man's sandpaper. Later, craftsmen tried to fix abrasive grains to flexible backings with crude adhesives. A 13th-century Chinese document describes the use of natural gums to fix bits of seashell to parchment. About two centuries later, the Swiss began coating crushed glass on a paper backing.

Early sand and glass abrasives lacked sharpness, and by the 19th century early abrasive products like the natural sandstone that man had formed into the "grinding wheel" no longer met the needs of developing industry. In 1873 Swen Pulson, working in the Norton and Hancock Pottery Company, Worcester, Massachusetts, won a jug of beer by betting that he could make a grinding wheel by combining emery with potter's clay and firing them in a kiln. Pulson succeeded on his third try; this incident signalled the end of unsatisfactory glue-and-silicate bonded products and the birth of the vitrified grinding wheel.

Just before the turn of the century, when nature's abrasives emery, corundum, and garnet were falling short of industry's demands, man discovered how to make aluminum oxide and silicon carbide with the electric fur-

First synthetic diamonds

nace. In 1955 the General Electric Company succeeded in manufacturing diamond. Like other man-made abrasives synthesized diamond proved superior, in many applications, to the natural product, which had been used in grinding wheels since 1930.

Once used only when precise dimensional accuracy and smooth surfaces were necessary, abrasives have become a widely applied industrial tool. Higher grinding-wheel speeds, higher horsepowered grinding machines, and improved abrasives and abrasives products have steadily augmented their role.

NATURAL ABRASIVES

Most natural abrasives have been replaced by manufactured or synthetic abrasives. This is because nearly all industrial applications demand consistent properties from lot to lot for uniform manufacturing results. With the exception of natural diamond, most of nature's abrasives are too variable in their properties. Natural abrasives, however, are still important.

Diamond. Diamond is the hardest abrasive, and in fact the hardest material known. India was the main source of diamonds until the 1860s when they were discovered in South Africa, which still supplies 95 percent of all diamonds mined. Lesser quantities are found in Brazil, Australia, Guyana, and Venezuela.

Stones not suitable for use as gems (called bort or boart) are designated for industrial purposes. The portion used as abrasive is crushed into various sizes for grinding wheels, lapping (fine polishing) powders, coated abrasive belts, and polishing disks. Industrial diamonds vary in colour from off-white to black.

The shape of a diamond particle is important to its performance. Natural diamonds are sorted for shape. Blocky shapes are suited for more severe grinding operations such as cutting concrete building blocks. More splintery shapes are reserved for grinding the cemented tungsten-carbide cutting tools used in industry.

Corundum. Corundum is a naturally occurring form of aluminum oxide, found mainly in South Africa. Other deposits have been found in India, Canada, and the southern part of the United States. Corundum varies in its properties from location to location and even from within the same location, making it unsuitable for use in grinding wheels. Its principal utility is in polishing and glass grinding.

Emery. Also a natural form of aluminum oxide, emery is generally more impure than corundum. It consists of smaller aluminum oxide crystals, imbedded in an iron oxide matrix. Principal sources are Turkey and Greece. Its rather round grain shape makes it more suitable for polishing than for grinding. Along with corundum, emery was an important early abrasive and is still used in emery cloth sandpaper.

Garnet. Far superior to flint, and still valuable for coated abrasive products in the woodworking, leather, and shoe industries, garnet possesses considerable toughness. When its cutting crystal edges break down, garnet fractures sharply to present new cutting edges.

Of the seven known forms of garnet, the best for abrasive use comes from the Adirondack Mountains in New York and is mined by a process similar to panning for gold.

Flint. Flint, or flint quartz, is the most commonly used abrasive in the manufacture of coated abrasive paper (sandpaper). Though flint is found all over the world, not all varieties are suitable as abrasives. The best type varies in colour from a dull grayish white to a faint pink and comes from the eastern United States, where it is mined from open pits. Lumps, varying in size from four to ten inches (ten to 25 centimetres) in diameter, are crushed into sharp crystals. Flint lacks durability and hardness and loses its abrasive cutting action rather quickly.

Quartz. Quartz is the main ingredient of sandstone, which was for years hewn from natural deposits and formed into the first grinding wheels. Until recently these were used in the glass, cutlery, and woodworking edge-tool industries. The best stones came from Ohio and West Virginia. Although still in use where production rate is not a factor, as on a farm, they have been replaced by higher speed, higher efficiency manufactured grinding wheels and equipment.

In other forms, quartz is found in dense stones in the United States and in Belgium. These are used as honing and sharpening stones for putting the final, keen edge on cutlery and various cutting tools and are used for honing machined parts and products to accurate size and smooth finish.

Quartz in the form of sand is used as an inexpensive, disposable abrasive for sandblasting buildings and metal structures such as bridges, when it is impractical to recover the abrasive. In a crushed or grain form, it is used in buffing and polishing wheels.

Pumice. Volcanic in origin, pumice is the frothy part of lava. The best quality, in lump form, is mined in Italy. Its mild action is utilized in metal polishing, glass cutting, furniture finishing, scouring powders, and soaps.

Tripoli. Tripoli (or Diatomite) is also used in scouring powders and soaps, in grease sticks for buffing, and as a polish ingredient. The principal United States sources are Missouri and Oklahoma. Other sources are Iceland and Kenya.

Other natural abrasives. Talc is the softest and mildest abrasive. It was used in the past to polish grains of rice. In addition to sandstone, other varieties of silica, such as rottenstone, which comes from the weathering of impure limestones, and diatomite, an opaline silica consisting of the fossilized remains of microscopic plants, are used in polishes. Calcined (roasted) diatomite is an effective polishing agent for use on nickel plate. Others include feldspar, chert, and chalk.

MANUFACTURED ABRASIVES

By the 1970s, manufactured abrasives represented the larger share of world production and use. The two most important manufactured abrasives, silicon carbide and aluminum oxide, were developed in the late 19th century by use of the electric furnace. These proved to have the consistent quality required by industry. Their principal value is for use in grinding wheels, coated abrasive products, and polishing and buffing wheels. Because large amounts of electric power are required in their manufacture, electric furnace abrasives are made where power is cheap and plentiful.

Silicon carbide. Silicon carbide was first produced by Edward G. Acheson, an electrical engineer in Monongahela City, Pennsylvania. In 1891 he combined a mixture of clay and powdered coke in a small electric furnace. The resulting hard, shiny crystals were able to scratch glass and were first used for polishing precious stones. Initially, the price of silicon carbide was nearly equal to that of the gemstones it polished.

In the manufacture of silicon carbide, a mixture of pure silica sand and carbon in the form of finely ground coke is built up around a carbon electrical conductor within a brick furnace. Electrical current is passed through the carbon conductor, generating heat and bringing about a chemical reaction in which the carbon in the coke and silicon in the sand combine to form silicon carbide. A little sawdust is added to the mixture. The sawdust burns, making the mixture porous and permitting the escape of great quantities of inflammable gas.

At the completion of the furnace run the product consists of a core of silicon carbide crystals loosely knitted together, surrounded by partially or entirely unconverted raw material. The quality of the abrasive produced depends largely upon the choice of raw materials and the accuracy of the control and operation of the furnace.

A number of companies produce silicon carbide under such trade names as Carborundum, Crystolon, Carbolon, etc.

Silicon carbide manufacture

The method used to manufacture silicon carbide today is basically the same as that developed by Acheson. The major changes are in the size of the furnace and control of the ingredients, techniques, and the understanding of the process. Time for a furnace run is about 36 hours, during which temperatures vary from 4,000° to 4,500° F (2,200° to 2,480° C) in the core and drop to about

2,500° F (1,370° C) at the outer edge. The furnaces are the shape of a large trough with electrodes at either end. The power amounts to over 100,000 kilowatt-hours per run. Furnaces are about 40 feet long, 10 feet high, and 10 feet wide and can produce about 25,000 pounds of silicon carbide per run.

Aluminum oxide. Manufacture of aluminum oxide abrasive closely followed the development of silicon carbide. Although there are records of previous experiments to produce emery and corundum abrasives, the world's first successful attempts were made in the laboratories of the Ampere Electro-Chemical Company at Ampere, New Jersey, in 1897. Rock bauxite, the main ingredient of which is aluminum oxide, was used as the raw material.

The results were sufficient to arouse interest. Several hundred tons of the new abrasive were distributed throughout industry for testing, and the outcome of this testing was the construction of a plant at Niagara Falls in which the manufacture of crystalline aluminum oxide on a commercial scale was begun in 1901.

Bauxite is a claylike substance. Found near the surface, it is one of earth's most plentiful minerals. It is also the source of aluminum. Principal mines are in Surinam, Guyana, and the United States. Bauxite consists of aluminum oxide, Al_2O_3, combined with water and associated with varying quantities of other impurities. By heating the ore to drive off the combined water before the material goes into the electric furnace (an operation called calcining), the percentage of aluminum oxide is greatly increased over its content in the natural state.

Arc-type electric furnaces are used for manufacturing aluminum oxide. Two carbon electrodes are suspended in the furnace. The furnace is charged with the crude ore. Current passes from one electrode to the mass, passes through it, then arcs back to the other electrode. This produces an intense heat of over 2,000° F, which melts the bauxite and reduces the major portion of the impurities to their base metals. Small percentages of coke and iron borings combine with the impurities, thus helping to settle them to the bottom of the mass. During the furnace run of approximately 16 hours, fresh bauxite charges are continually added.

At the end of the furnace run, the large steel jacket or outer shell of the furnace is lifted and the hot chunk of aluminum oxide is allowed to cool. Cooling time varies from one day to one week depending on the type of crystal structure desired.

Both the chemical composition and the crystalline structure of the abrasive are important in the control and variation of the abrasive properties. A wide variation is possible; this characteristic has been utilized in tailoring aluminous abrasives to specific industrial applications. The types vary in colour from one that is pure, snow-white, of large crystals, and friable, used for the grinding of heat-sensitive metals, to the nearly black, finely crystalline, less pure kinds used for heavy-duty operations in foundries and steel mills.

Recent developments include alloy abrasives, made by adding, for example, zirconium oxide to aluminum oxide and the spinel (a hard crystalline mineral) and aluminum oxide combinations.

Produced by a number of companies under a variety of trade names, aluminum oxide is the most important abrasive used.

Synthetic diamond. Synthetic diamonds, initially more expensive than natural diamonds, have proved to be more efficient in the grinding wheels used for manufacturing and sharpening cutting tools made of cemented tungsten carbide. Because of the more irregular surface of the synthesized diamond, the bond between the grinding wheel and the diamond particles lasts longer.

Diamond is now manufactured in a variety of shapes and sizes for different types of grinding operations. By changes in the process, the diamond can be manufactured in shapes suitable for cutting ceramics and glass.

Metallic coating on the surface of the manufactured diamond also improves bonding. Nickel and copper have proved to be the most successful coatings. Both are heat conductors and serve to dissipate the heat of grinding and, thus, act as a bonding interlayer, which improves bonding strength.

Boron carbide. A product of the electric furnace, boron carbide is harder than silicon carbide but is not suitable for use in grinding wheels or coated abrasive products. It does have some utility as a lapping (fine polishing) abrasive, substituting for diamond in the polishing of hard materials. It is made from the combination of coke and boric acid at very high temperature. In solid form, boron carbide becomes extremely wear resistant.

Boron nitride. Cubic boron nitride was first synthesized by the General Electric Company. Its hardness is reported to be twice that of silicon carbide and two and one half times that of aluminum oxide. High-pressure techniques are used in making the material, which has strong crystals and sharp points with clearly defined edges. Despite its hardness, the initial forms of boron nitride did not have the mechanical strength necessary for grinding.

Cubic boron nitride with a bonded nickel coating is a useful abrasive. Although more expensive than industrial diamond abrasive, its grinding efficiency on certain tool-steel applications has made it economically feasible to use. Continued experiments in its manufacture may result in lower costs for this newest of manufactured abrasives.

Other manufactured abrasives. Other manufactured materials employed as abrasives are neither as hard nor as abrasive on work surfaces as those listed above. Glass beads manufactured specifically for use in blasting machines, utilizing air pressure to propel them against a workpiece, remove rust and scale. Basically a cleaning abrasive, they do not materially alter the dimensions of the workpiece.

Metal shot made of bronze, iron, steel, and steel alloy is used for shot blasting and cleaning in foundries, and by forging companies to clean castings and parts and to alter surface finishes. Shot is also used for relieving internal stresses.

Steel wool, made by combing the wire that forms steel fibres (with equipment similar to that used in carding wool), is used to clean and finish wood and many other materials. The size of the steel fibres is changed by varying the size of the teeth in the card-clothing, the more abrasive product being the one with coarser wires.

ABRASIVE PROPERTIES AND MANUFACTURE

One of the most important properties necessary in an abrasive material is hardness. The abrasive must be harder than the material it is to grind, polish, or remove. Hardness of the various abrasive materials is measured on either the Mohs scale or the Knoop hardness scale. The Mohs scale, first described in 1812, measures resistance to indentation as judged by which material will scratch another. This scale, which assigns numbers to natural minerals, has been widely accepted and is used by mineralogists. The Knoop hardness test employs an elongated diamond-shaped indenting device and measures the indentation it makes in a given test material. With the Knoop test, the hardness of extremely brittle materials including glass and even diamonds can be measured without harming either the indenter or test piece (see MATERIALS TESTING).

Measuring hardness

Toughness or body strength characteristics are significant to abrasive function. Ideally, a single abrasive particle resharpens itself by the breakdown of its dull cutting or working edge, which exposes another cutting edge within the same particle. In manufactured abrasives it is possible to achieve some degree of control over this property by varying grain shape during the crushing or sizing operation, by making changes in the purity of the abrasive, by alloying abrasives, and by controlling the crystal structure within abrasive grains. Thus abrasives can be developed to meet the operating conditions found in a variety of applications.

Interaction between the abrasive and the material being ground prevents the use of one abrasive as a universal medium. When silicon carbide is used on steel, or alu-

minum oxide on glass, there is some reaction, which has yet to be clearly defined. Regardless, the effect is rapid dulling and inefficient abrasive action. Attrition resistance is the name given to this third, very significant property.

Preparation and sizing. All abrasives, with the exception of the naturally appearing fine powders such as talc and iron oxide, must be crushed to the particle size required for use. Sizes in use vary from 4 grit, which measures about ¼ inch (six millimetres) in diameter, to as fine as 900 grit, which measures about six microns (0.00024 inch) or about one-tenth the thickness of a human hair. In some cases, even finer powders are required when used for the polishing of scratch-free surfaces on high-quality optical lenses and mirrors for high-power telescopes.

For the coarser sizes the crushed grain is measured by a series of test screens as established in most countries by government standards. For sizes 240 grit and smaller, the size of the grains is usually measured by a settling rate or sedimentation test.

Crushing methods have a significant effect on the cutting action and the strength of abrasive grains. Heavy crushing pressures, for example, tend to create splintery, sharp, and weak grains. These penetrate easily, remove material at a fast rate, and their cutting edges break down easily for resharpening. This shape is required in many coated abrasive products.

When a grinding wheel or coated abrasive belt is to be used to grind away large amounts of material, under heavy pressures, a strong pulverized abrasive grain stands up longer. The sharp edge is not as necessary because heavy pressures are used that effect the penetration of the abrasive into the material being ground.

Abrasive product manufacture. The most important abrasive product manufactured is the grinding wheel. Made of abrasive grain and a binder, or "bond," it is a self-sharpening cutting tool. As the grains on the periphery become dull, they are shed from the surface of the grinding wheel and fresh, sharp cutting edges are exposed. The ability to resharpen is controlled by the nature of the bond used and the ratio of abrasive to bond, measured by volume.

In the manufacture of grinding wheels, abrasives of the proper size and bonding materials are weighed out and mixed together in a power mixing machine. When thoroughly mixed, measured amounts of the abrasive and bond mixture are evenly distributed in steel molds. The mold is placed in a powerful hydraulic press, and the mixture is compressed to the desired wheel size, allowing some oversize in dimensions for finishing operations. Pressures vary with wheel size and may be upward of several hundred tons. Some shaving and shaping of the wheels is done before they are baked or fired.

The majority of grinding wheels made have a vitreous, ceramic bond, made of clays and feldspars. The so-called vitrified wheel is fired in high-temperature kilns at temperatures of 2,300° F (1,260° C). Electric-, oil-, and gas-fired kilns are used. The length of the "burn" varies with wheel size and can be as long as two weeks.

The remaining 35–40 percent of the grinding wheels manufactured have organic type bonds using either resins, rubber, or shellac as the bonding material. These wheels are baked at temperatures between 300° and 400° F (150° and 200° C).

The lower temperature permits the inclusion of steel rings, molded-in threaded bushings, or fibreglass reinforcements, which become baked into the wheel and serve to make it more resistant to breakage from side pressure. The bushings help to hold the wheel in place on certain grinding machines.

Organic wheels can be made much thinner than vitrified wheels, and they are used in place of metal saw blades for cutting a great variety of materials.

Truing a grinding wheel

Nearly all grinding wheels must be finished after they have been baked or fired. In a process called truing, the wheels are cut to final size, and the outside glazed layers resulting from the kiln are removed, making the sides of the wheel parallel, and the size of the arbor hole accurate; at the same time the working surface of the wheel is sharpened. Wheels are trued with conical steel cutters, by rubbing in beds of steel shot, and by actually grinding them with grinding wheels.

Grading of wheels assures they have the correct resistance to wear. Grade or hardness, determined by the amount of bond, permits the grinding wheel to keep itself sharp and free cutting in a variety of conditions. Grinding wheels used for heavy grinding operations are "harder" and are made with greater amounts of bond, retaining the abrasive particles longer under severe conditions such as those found in steel mills and foundries. In the tool room where industrial cutting tools are sharpened, softer wheels, with less bond per unit of abrasive, are required so that as soon as the abrasive grains start to dull, and the possibility of building up heat develops in the sensitive cutting tool being ground, the wheel will resharpen and shed the dulled grains.

Finally, grinding wheels are checked for balance to assure that they will run without vibration.

Grinding wheels six inches in diameter and larger are usually speed tested. The wheel is rotated at a speed at least 50 percent greater than the maximum allowable operating speed. This is a nondestructive measure of the wheel's bursting strength.

The same basic processes are used in the manufacture of abrasive bricks, sticks, and formed shapes. These are used as rubbing blocks, sharpening stones, honing stones, and shaped abrasive media for tumbling or mass finishing.

Sandpapers (coated abrasive) are the next most significant abrasive product. They consist, basically, of a single layer of abrasive particles held to a flexible backing material.

Manufacture begins with huge rolls of backing material, either paper, cloth, or a combination of the two. The backing is fed to the making machine where the first layer of adhesive is applied. Next, the layer of abrasive is applied, either by gravity or electrostatically. The electrostatic method orients the slivery type of abrasive used, with the sharp ends facing out. The process can also control the spacing of the grains. This is an advantage in that wide spacing helps to alleviate loading problems when grinding soft, stringy, or gummy materials.

After being coated with the abrasive layer, the product is draped in long festoons in which it dries partially. Then it is run through another sizing operation and a second layer of adhesive is applied. The product is draped, allowed to dry thoroughly, and wound into large-diameter rolls.

The adhesives used to bond the abrasive to the backing are water soluble, waterproof, or a combination of the two. Water-soluble types are used for dry grinding operations and in household sanding, and occasionally on wood-sanding commercial operations. Resin or resin-over-glue types currently in use have the flexibility associated with the soluble types. The all-resin is best for severe operations; its properties are such that the heat of grinding actually increases the adhesive's holding strength.

Disks for right-angle grinders are die cut. Sheets are cut in the standard lengths and widths used in production and household applications. Strips are slit, cut to length and their ends joined to make up the coated abrasive belts that have become an essential part of industry, replacing several kinds of abrasive wheels used previously.

The cutting action of coated abrasive products is determined by the abrasive used, the grit size, the density or spacing of the grit, the strength of the adhesive, and the flexibility of the backing material. Flexibility can be achieved by flexing or breaking the bond.

Other products use the abrasive in the form of grains or powders. In addition to the sizing operation, many types are specifically treated, by calcining, acid, or heating, to make them more suitable for use as lapping abrasive or perhaps as sandblasting grain. For use in lapping and polishing, the abrasive is usually mixed with a vehicle such as mineral or seal oil. Polishing sticks consist of waxes or greases impregnated with various-sized abrasive

grains, depending on the particular requirements of the work.

Two materials used really for cleaning rather than grinding are still abrasive in nature. Glass beads, pressure blasted onto a surface, remove rust, scale, and carbon. These have replaced much hand cleaning with steel wool. Steel wool still has some applications.

INDUSTRIAL APPLICATIONS

Grinding, the most important abrasive application, is in some way involved in the manufacture of almost every product. This use may be direct, as when the product requires pieces that must be made within close dimensional tolerance limits, or a very smooth surface, or when used on materials too hard to be machined by conventional cutting tools, or indirect, as when, for example, grinding wheels are used to sharpen cutting tools. The materials that are used to make cutting tools must of course be hard in order for the cutting tool to cut and retain its sharp edge. Abrasive grinding wheels are the only means for sharpening the dull edges of such tools.

Grinding wheels in use in industry today rotate at peripheral speeds up to 180 miles (290 kilometres) per hour. The abrasive wheel may throw a 40-foot stream of bright yellow sparks and remove upward of 1,000 pounds (450 kilograms) of metal per hour while grinding out the imperfections from the surface of a bar of stainless steel. Or the grinding wheel may be as small as 0.022 inch (0.55 millimetre) in diameter, may rotate at 150,000 revolutions per minute, and may grind miniature precision ball bearings to accuracies measured in millionths of an inch.

In the automotive industry, only abrasives can produce the tight fit required between piston rings and cylinders to prevent the escape of compressed gasoline vapours. Valves and valve seats are ground. Bearing surfaces in the engine, transmission, and wheels need specific finish, size, and roundness to assure frictionless rotation. These can be achieved only with abrasive tools.

Abrasive machining, the use of abrasives rather than high-speed steel or tungsten carbide cutting tools, makes use of the self-sharpening grinding wheel and eliminates tool sharpening costs. The ability to grind hardened materials without the previously necessary prehardening machining saves intermediate part-handling operations.

Speed and improved grinding systems, machines, and grinding tools are the main reason for the increased importance of grinding. One-hundred-horsepower motors, automatic loading equipment, high-speed grinding wheels removing unbelievable amounts of hard-to-grind materials, ultrafine tolerances, and machines costing $100,000 and more are part of the new abrasives systems that are capable of fantastic rates of production. Yet, the abrasive products portion of the total cost to grind the part may be as low as 5 percent, even though the grinding wheel, with diamond as the abrasive, might cost $14,000. Or the wheel might cost 28 cents and be used to shape a die used in the manufacture of tableware.

Relatively new products, such as plastic bonded tumbling media for mass finishing a multitude of parts, have eliminated time-consuming, hand-deburring operations, a plague to the aircraft industry, in which the high cost of labour makes handwork prohibitively expensive. Parts with rough edges are tumbled in a rotating barrel of loose abrasive or preshaped abrasive pieces. As the mass slides, burrs are ground away, surfaces are finished, and edges are smoothed. When the same parts and media are mixed in a vibrating tub, the process becomes even more productive.

Abrasive wheels have replaced steel saws in many places. Thin, abrasive cutoff wheels are capable of sawing through nearly every material known, at rates faster than those of metal saws, while generating less heat and producing a better cut surface. Some space-age metals, because of their hardnesses, can be cut only with abrasive wheels. Granite, marble, slate, and various building blocks are cut to size with diamond abrasive wheels. Grooves for expansion joints and for the reduction of wet-weather skidding accidents are cut in concrete runways and highways by blades with a metal centre, onto which are brazed metal segments with the diamond abrasive mixed throughout.

Tool sharpening

The sharpening of all types of tools continues to be a major grinding operation. Drills, saws, reamers, milling cutters, broaches, and the great spectrum of knives are kept sharp by abrasives. Coarser grit products are used for their initial shaping. Finer grit abrasives produce keener cutting edges. Ultrasharp tools must be hand honed on natural sharpening or honing stones. Even grinding wheels themselves may require some sharpening. Specially designed steel disks or diamond tools are used to remove dull abrasive cutting edges and create a sharp cutting surface.

In foundries and steel mills, grinding wheels and coated abrasive belts remove the unwanted portions of castings, forgings, and billets. Abrasive grit is pressure blasted against the metal to clean it in preparation for painting. Metal shot is used on softer metallic castings.

The roster of unusual applications for abrasives includes frozen fish cut into fish sticks; animal-gut sutures and guitar strings ground to constant diameters; human skin blemishes and birthmarks removed or bones shaped by plastic surgery; spots and discolorations removed from suede clothing; toothpicks ground round; and stairs made skidproof by abrasive grain.

STATISTICS

Grinding wheels range in size from ten-ton pulpstones used to chew pulpwood into 80 tons of newsprint per day to 0.022-inch-diameter mounted wheels grinding miniature bearings. Some abrasives used for polishing are so tiny that if scented the minute particles would pass for talcum powder. Yet, when made into a polishing compound, they grind as surely, yet minutely, as the 72-inch-diameter wheels used to grind files and saws.

Competition in the abrasives market (about $1,000,000,000 annually in the non-Communist world) has prompted manufacturers throughout the world to develop more efficient abrasive tools and machines. Technological advances have increased the efficiency of diamond abrasive wheels five times in the last 14 years. The zirconia–aluminum oxide abrasive wheel used in steel mills grinds eight times faster, at one-tenth the abrasive cost, and, despite a fivefold increase in labour and overhead rates, at 25 percent of the total cost of 1948.

Yet, despite many advances, the basics of the grinding operation are not clearly understood and are still the subject of research. One industry problem is the multiplicity of sizes and varieties. One leading abrasive manufacturer stocks 1,000 different types and sizes, 38,000 sandpapers, and 250,000 grinding wheels. Experts are still unable to specify the exact optimum combination of inputs for a well-defined grinding operation on a specific part. There are still too many variables.

Even in the 1970s, when an exceptionally smooth, polished finish or an especially sharp edge is required, the job usually calls for the highly skilled craftsman employing hand tools and methods essentially similar to those used centuries ago.

BIBLIOGRAPHY. KENNETH B. LEWIS, *The Grinding Wheel* (1959), published by the Grinding Wheel Institute, contains general information on the various types of grinding operations, presented from a nontechnical point of view. Another nontechnical source is P.S. HOUGHTON, *Grinding Wheels and Machines* (1963). The Norton Company, Worcester, Massachusetts, publishes, and updates annually, about 16 booklets of general and semitechnical information on abrasives; titles include *Lectures on Grinding, Lectures on Coated Abrasives,* and *Abrasives and Grinding Wheels.* Another source of nontechnical information is the booklet series published by the Abrasive Grain Association, Cleveland, Ohio; titles include *History of Abrasive Grain, Physical and Chemical Properties of Abrasive Grain, Use of Abrasive Grain in the Manufacture of Coated Abrasives, Use of Abrasive Grain in Polishing,* and *Use of Abrasive Grain in Lapping and Metallographic Polishing. Proceedings* of the INDUSTRIAL DIAMOND ASSOCIATION and of the AMERICAN SOCIETY FOR ABRASIVE METHODS are published annually with contents consisting of reprinted versions of papers presented at their conferences. Topics vary from very general to highly technical. An important periodical, published eight times each year, is *Grits and Grinds.*

Various papers on abrasives are also available from technical societies such as the Society of Manufacturing Engineers; the American Foundrymen's Society; and the American Society for Abrasive Methods.

(D.J.B.)

Abū Ḥanīfah

Abū Ḥanīfah an-Nuʿmān ibn Thābit, 8th-century theologian and jurist, founded the first of four orthodox Islāmic schools of law. As a jurist, he reviewed the then existing body of legal doctrines, formulated views on new questions, and integrated these into a coherent system by anchoring them to a well-defined legal theory. The school of Abū Ḥanīfah acquired such prestige that its doctrines were applied by a majority of Muslim dynasties. Even today it is widely followed in India, Pakistan, Turkey, China, Central Asia, and Arab countries.

Abū Ḥanīfah was born in 699 in Kūfah, an intellectual centre of Iraq. He belonged to the *mawālī*, the non-Arab Muslims, who pioneered intellectual activity in Islāmic lands. The son of a merchant, young Abū Ḥanīfah took up the silk trade for a living and earned renown for honest dealing. It provided him with a steady and handsome income, out of which he donated generously to charity, helping scholars particularly.

Abū Ḥanīfah had a well-rounded education. In early youth he was attracted to theological debates, in which his sharp wit and logical mind won him laurels. Disenchanted with theology, he turned to law and for about 18 years was a disciple of Ḥammād (died 738), then the most noted Iraqi jurist. After Ḥammād's death, he became his successor. He also learned from several other scholars, notably the Meccan traditionist ʿAṭāʾ (died *c.* 732), and the founder of the Shīʿite law, Jaʿfar aṣ-Ṣādiq (died 765). Abū Ḥanīfah's mind was also matured by extensive travels, contacts with a variety of people, exposure to the stimulating influences of the heterogeneous, dynamic, and advanced society of Iraq, and conversance with business practices.

Tall, of medium weight, Abū Ḥanīfah was impressive in his bearing and dignity. He had a serious and sober disposition and was known for piety and religious earnestness and devotion. Having an independent income, he neither needed nor cared for governmental patronage, and this helped Islāmic law develop independently of political authority. His independence and fearlessness, his disregard for worldly vanities, and, above all, his piety made him a paragon of Muslim scholarship.

By Abū Ḥanīfah's time a vast body of legal doctrines had accumulated as a result of the endeavour to apply Islāmic norms to legal problems. The disagreements in these doctrines had rendered the development of a uniform code, based on sound principles, a need of the hour. Abū Ḥanīfah responded by thoroughly scrutinizing the current doctrines in collaboration with his students, several of them outstanding scholars. He had each legal problem discussed thoroughly before formulating doctrines definitively.

Contribution to Islāmic legal thought

Being a speculative jurist, Abū Ḥanīfah brought about systematic consistency in legal doctrines. In his doctrines, emphasis shifts from material to systematic considerations. Again and again he disregarded established practices and considerations of judicial and administrative convenience in favour of systematic and technical legal considerations. His legal acumen and juristic strictness were such that Abū Ḥanīfah reached the highest level of legal thought achieved up to his time. Compared with his contemporaries, the Kufan Ibn Abī Laylā (died 765), the Syrian Awzāʿī (died 774), and the Medinese Mālik (died 795), his doctrines are more carefully formulated and systematically consistent and his technical legal thought more highly developed and refined.

Before Abū Ḥanīfah's time, doctrines had been formulated mainly in response to actual problems, whereas he attempted to solve problems that might arise in the future. By the introduction of this method—a characteristic of Abū Ḥanīfah—the area of law was considerably enlarged. The method presupposed advanced legal thought —thought that was further refined by the practice of this method. It was presumably because of this, coupled with Abū Ḥanīfah's somewhat rationalist orientation and his reserve about traditions that were not highly authenticated, that his school was denounced as the school of *raʾy* (independent opinion) as opposed to that of Ḥadīth (authoritative tradition). This, however, was a polemical allegation rather than an objective statement of the school's standpoint.

Although it was not Abū Ḥanīfah's primary concern, his contribution to theology was also significant. He took distinct and moderate positions on several theological questions, stimulating the development of the Māturīdī school.

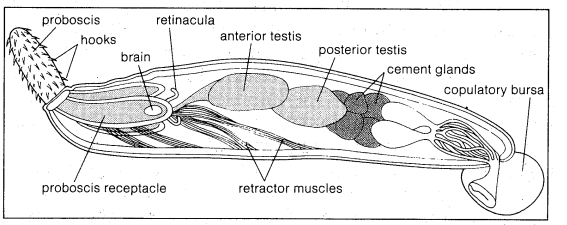
Thanks to his temperament and academic preoccupation, Abū Ḥanīfah took no direct part in power struggles, despite his antipathy toward the Umayyads and ʿAbbāsids, the ruling dynasties of that time. His sympathies lay with the ʿAlids, whose revolts he supported with both words and money. This partly explains why Abū Ḥanīfah adamantly refused to accept a judgeship and, consequently, suffered severe persecution under both the dynasties. In 767 he died in Baghdad, either while imprisoned or soon after release.

BIBLIOGRAPHY. For bibliographical information on Abū Ḥanīfah, see the article by J. SCHACHT in the *Encyclopaedia of Islam*, vol. 1, pp. 123–124, new ed. (1960); and *An Introduction to Islamic Law* (1964). For most authentic sources of Abū Ḥanīfah's legal doctrines, see the works of his disciples, Abū Yūsuf (d. 798) and al-Shaybānī (d. 805). For theological doctrines, see A.J. WENSWINCK, *The Muslim Creed* (1932); and J.A. WILLIAMS (ed.), *Islam* (1961).

(Z.I.A.)

Acanthocephala

The spiny-headed worms constitute the phylum Acanthocephala. A proboscis, which bears hooks, gives the group its name. There are about 600 recorded species, all of which are parasites in vertebrates as adults and in arthropods as juveniles. The adults are usually less than one centimetre in length, but some reach half a metre or more. Acanthocephalans are found throughout the world, especially in fish.



Male *Acanthocephalus ranae*, internal structure.

Natural history. Acanthocephalans are bisexual, and mating occurs in the intestines of vertebrate hosts. The eggs, which actually are shelled larvae, are excreted with the feces of the host. No further development occurs until the eggs are eaten by an arthropod, which is the intermediate host. After its release in the gut of the arthropod, the larva, called an acanthor, bores through the gut wall into the host's blood cavity (hemocoel), is encapsulated there, and develops into a new stage called an acanthella. The acanthella, a miniature version of the adult, withdraws its armed proboscis before entering a resting stage during which it is known as a cystacanth. Once again, no further development occurs until the cystacanth has been ingested by its definitive host, a vertebrate. The young acanthocephalan emerges inside the vertebrate's intestine, uses its proboscis to bore into the gut wall, and matures there.

If a cystacanth is swallowed by the wrong host, it may bore through the gut wall into the body cavity where it encysts and remains infective until the accidental host is eaten; then the cystacanth emerges again in its new host. This behaviour, incorporated into some acanthocephalan life cycles, may provide the only means for completing the life cycle if the definitive host does not feed directly on the intermediate host. Although usually fish parasites,

acanthocephalans also occur in amphibians, reptiles, birds, and mammals. Man is only rarely infected and then accidentally. Because they do little damage to their hosts, acanthocephalans are of no economic importance.

Form and function. The body of an acanthocephalan is divided into a proboscis and an elongated cylindrical trunk. The proboscis, which usually bears backwardly pointing hooks, can be withdrawn into the trunk, which may also bear hooks or spines. Usually white, acanthocephalans may occasionally be yellow, orange, or red in colour. The internal anatomy is simple; there is no gut, and the bulk of the trunk is occupied by a proboscis receptacle and the muscles used to retract it. Passing backward from the base of this receptacle are structures known as ligament sacs. One of these sacs in the male encloses two testes and a number of cement glands. The female has two sacs that do not necessarily persist after she matures. Externally the male possesses an eversible copulatory bursa, the female a simple vagina.

Evolution and classification. The affinities of the acanthocephalans are uncertain. They possess a pseudocoelom; thus, although they are closely related to members of the phylum Aschelminthes and possibly to the rotifers or priapulids, acanthocephalans are sufficiently different from these groups to justify their classification as a distinct phylum. Attempts to derive the acanthocephalans from the tapeworms have not been accepted.

Features used in the classification of acanthocephalans include the nature and distribution of hooks and spines, which are not of any proved phylogenetic significance. For this reason a number of classifications exist. The classification given below is based on the one accepted by Libbie Henrietta Hyman (1951) and Pierre-Paul Grassé (1961) in their widely used textbooks.

Annotated classification.

PHYLUM ACANTHOCEPHALA

Adults parasitic, usually in fish; juveniles parasitic in arthropods; usually one centimetre long; about 600 species; proboscis with hooks; bisexual; worldwide distribution.

Order Archiacanthocephala

Small proboscis hooks, arranged concentrically; males usually with eight cement glands; excretory system often present; ligament sacs persist in mature female; parasites of terrestrial vertebrates and arthropods.

Order Palaeacanthocephala

Proboscis hooks alternate in radial rows; males usually with six cement glands; no excretory system; ligament sacs do not persist in females; parasites of aquatic vertebrates and crustaceans.

Order Eoacanthocephala

Proboscis hooks arranged radially; cement gland cells difficult to distinguish; no excretory system; ligament sacs persist in female; parasites of fishes, reptiles, and aquatic arthropods.

Alternative classifications are given by references listed by W.L. Nicholas (1967).

BIBLIOGRAPHY. D.W.T. CROMPTON, *An Ecological Approach to Acanthocephalan Physiology* (1970), a comprehensive book on the biology of the acanthocephalans; P.P. GRASSE (ed.), *Traité de zoologie*, vol. 4, pt. 1, *Platyhelminthes, Mésozoaires, Acanthocéphales, Némertiens* (1961), an important modern text on this subject (in French); L.H. HYMAN, *The Invertebrates*, vol. 3, *Acanthocephala, Aschelminthes and Entoprocta* (1951), a good general account now partially outdated, should be read in conjunction with W.L. NICHOLAS, "The Biology of the Acanthocephala," in *Advances in Parasitology*, vol. 5 (1967), an excellent modern review; S. YAMAGUTI, *Systema Helminthum*, vol. 5, *Acanthocephala* (1963), comprehensive taxonomic work containing nearly all described species.

(F.E.G.C.)

Acarina

Mites and ticks, close relatives of spiders, scorpions, and harvestmen, constitute the Acarina (Acari, Acarida), which is part of the arthropod class Arachnida. The acarines may be separated into three orders, Opilioacariformes, Parasitiformes, and Acariformes, consisting of seven suborders and 333 families.

General features. *Size range and diversity of structure.* Some mites are less than 0.5 millimetre in length; the largest ticks are slightly over 30 millimetres. Nymphs and adults generally have four pairs of legs, the larvae three. Members of some families have a dense covering of stiff hairs, or setae, while others are almost nude. Soft-bodied mites and ticks have only a few hardened plates or shields; others that are strongly sclerotized have numerous hardened plates. Acarines range in shape from minute, soft-bodied, elongated or circular individuals to large, hard-bodied spider-like ones. Some are flattened; others assume grotesque body shapes as they become engorged with food or filled with eggs.

By courtesy of (B) G.W. Wharton, *Science*, 167: 1382 (1970); (C, bottom) from G.O. Evans, J.G. Sheals, and D. Macfarlane, *The Terrestrial Acari of the British Isles*, vol. 1, p. 139, figure 160, Trustees of the British Museum, (Natural History); (others) E.W. Baker and G.W. Wharton, *An Introduction to Acarology*

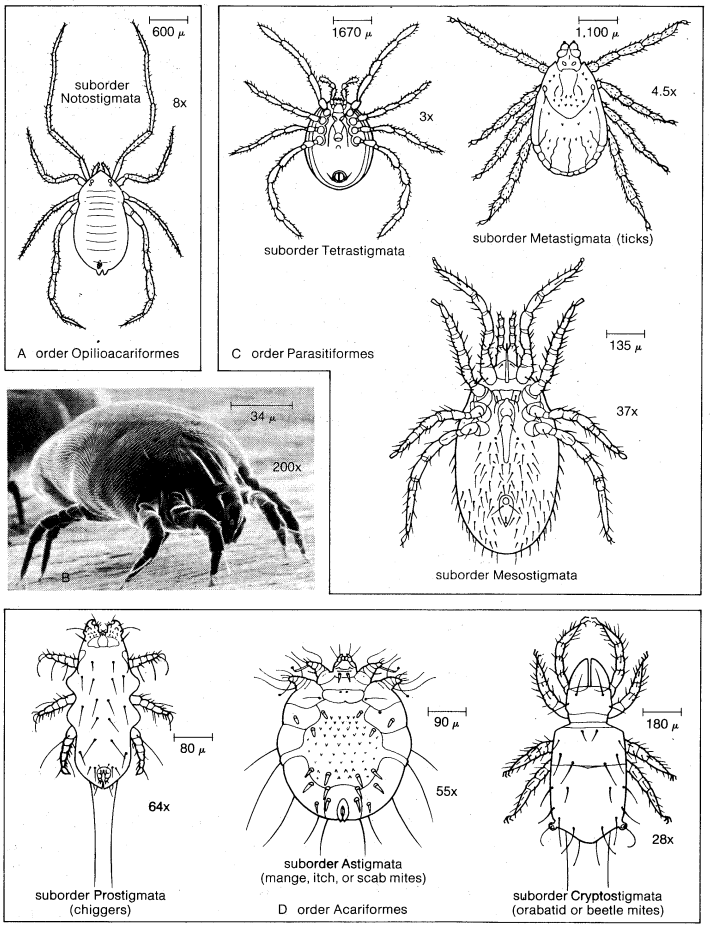


Figure 1: *Diversity of Acarina.*
(A) *Opilioacarus segmentatus*, female, dorsal view.
(B) Protonymph of the North American house-dust mite (*Dermatophagoides farinae*). (C, left) *Holothyrus nitidissimus*, male, ventral view; (C, right) *Dermacentor variabilis*, female, dorsal view; (C, bottom) *Ornithonyssus bacoti*, ventral view.
(D, left) *Myobia musculinus*, female, dorsal view; (D, centre) *Sarcoptes scabiei*, female, dorsal view; (D, right) *Parhypochthonius aphidinus*, female, dorsal view.

Distribution and abundance. Mites and ticks are distributed throughout the world in almost every conceivable habitat and frequently occur in large numbers. They are recorded as high as 5,000 metres on the slopes of Mt. Everest and as deep as 5,200 metres in the northern Pacific Ocean. Over 50 terrestrial species are known from the Antarctic. A few mites have been found drifting high in the atmosphere.

Acarines have been found in diverse habitats that include hot springs, caves, harsh deserts, and tundras. Although many of the approximately 30,000 species described thus far are free-living, some are parasites of animals. The total number of species of mites and ticks has been estimated at more than 500,000.

Importance. The Acarina are an economically important arachnid group; the ticks (Metastigmata), for example, surpass all other arthropods (except mosquitoes)

Carriers of disease

in the number of diseases they transmit to man. Many mites also are intermediate hosts of diseases transmissible to man, his domesticated animals, and his crops. Others are pests as a result of their biting or feeding habits or the damage they cause to food and related products.

In the mite suborder Tetrastigmata (order Parasitiformes), one species of *Holothyrus* is known to secrete an irritant substance that is toxic to fowl and man. The house-mouse mite (*Liponyssoides sanguineus*), which transmits the disease rickettsialpox to man, probably is the most important species in the suborder Mesostigmata (order Parasitiformes) from a medical standpoint. Widespread species such as the tropical fowl mite (*Ornithonyssus bursa*), northern fowl mite (*Ornithonyssus sylviarum*), and chicken mite (*Dermanyssus gallinae*) also are important pests of poultry and man.

Eriophyid and tetranychid mites (Prostigmata), widespread and abundant plant-feeding species, frequently cause serious injury to, or death of, the host plant; the former are the only phytophagous Acarina known to transmit plant viruses. The mange, itch, or scab mites (Astigmata) occur on many different animals including man. House-dust allergy is caused by species of *Dermatophagoides*, a common domestic arthropod.

The beetle mites (Cryptostigmata), among the most numerous soil arthropods, are important in the development of soil fertility and also act as intermediate hosts for important tapeworm parasites of domestic animals.

Rocky Mountain spotted fever, a rickettsial disease that occurs in the United States, is transmitted to man by the bite of several species of hard ticks (Ixodidae), especially the Rocky Mountain wood tick (*Dermacentor andersoni*) and the American dog tick (*Dermacentor variabilis*). Relapsing fever, an important bacterial disease throughout the world, is transmitted to man by certain species of soft ticks (Argasidae) of the genus *Ornithodoros*. Texas cattle fever is a widespread protozoan disease transmitted by cattle ticks (*Boophilus*); the disease, no longer prevalent in the U.S. because the tick has been eliminated, remains important in many tropical and subtropical countries. Other diseases transmitted to animals by ticks include anaplasmosis, tularemia, Q fever, Colorado tick fever, hemorrhagic fever, and tick-borne encephalitis (see DISEASE, HUMAN; and DISEASES OF ANIMALS).

The chiggers (Prostigmata), important pests of man, also transmit scrub typhus (tsutsugamushi disease), a rickettsial disease occurring in the Asiatic-Pacific region.

Natural history. *Reproduction and life cycle.* The sexes occur separately in Acarina; *i.e.*, there are both males and females. Most species lay eggs (oviparity), but, in some parasitic ones, the eggs hatch within the female, and the young are born alive (ovoviviparity). Many species also can reproduce by parthenogenesis; *i.e.*, by the development of unfertilized eggs. At least one tick, *Amblyomma rotundatum*, is exclusively parthenogenetic, and males are unknown; another, *Haemaphysalis longicornis*, has a distinct parthenogenetic population in which males are extremely rare and infertile.

Fertilization

Although there is considerable diversity in mating behaviour throughout the Acarina, the actual process of sperm transfer occurs directly or in packets called spermatophores. The male spermatozoa may be introduced by the male copulatory structure (aedeagus) directly into the female genital opening or, as in some Astigmata, into a special female copulatory structure called a bursa copulatrix. The males of species that use the latter method may have special copulatory structures (*e.g.*, suckers, spurs, or enlarged legs) for grasping the female. Some males produce a sealed packet containing spermatozoa (spermatophore) that is transferred to the female genital opening, either directly by the mouthparts of the male or indirectly by deposition on a surface, after which the female places it in her genital opening. Eggs begin to develop after mating and fertilization; although only a few eggs develop simultaneously in many Acarina, large numbers develop at the same time in ticks and some mites. Eggs are deposited haphazardly on food material by many plant- and grain-feeding species, and are hidden in the soil by predatory soil-inhabiting species. One

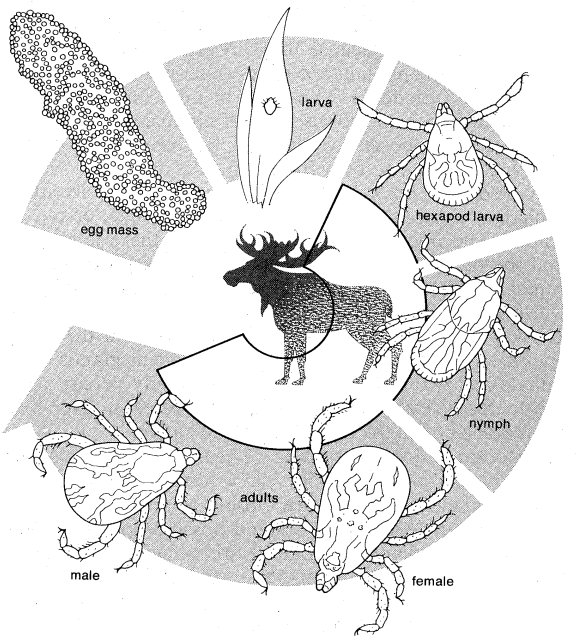


Figure 2: Life cycle of the hard tick (*Dermacentor*).
All stages except those of egg and larva require an animal host.
Based on H.D. Pratt and K.S. Littig, *Ticks of Public Health Importance and their Control* (1967)

predatory mite, *Cheyletus eruditus*, broods her small cluster of eggs and drives other arthropods from them.

The primitive life cycle among species that lay eggs has four active immature stages—hexapod larva, protonymph, deutonymph, and tritonymph. There are many deviations from this primitive type of life cycle, which is found only in the Cryptostigmata and some Prostigmata. The hexapod larva, characterized by three pairs of legs, is common to all families of Acarina except Eriophyidae (Prostigmata), whose members have only two pairs of legs in all active stages. Among the Parasitiformes, the Mesostigmata lack the tritonymphal stage; the Metastigmata may have only one nymphal stage (Ixodidae) or as many as eight (Argasidae). Some Prostigmata (Podapolipidae) develop directly from egg to larviform adults; others have from one to three nymphal stages. Many Astigmata (order Acariformes) have a nonfeeding, or hypopal, stage between the protonymphal and tritonymphal stages; it frequently occurs during adverse environmental conditions.

Ecology. The Acarina are a diverse group of the class Arachnida, in both habits and habitats. Based on habits, they may be considered either free-living or parasitic.

Free-living Acarina

The free-living Acarina include species from all of the suborders except Metastigmata (ticks). The beetle mites (Cryptostigmata), largely fungal feeders, are extremely numerous in the surface layers of soil; it is estimated that as many as 6,000,000 members of one species occur in an acre of pasture soil. One interesting species (*Symbioribates papuensis*) lives in fungi that grow on the backs of large weevils found in high altitude moss forests of New Guinea.

Some families of free-living mites have specialized stylet-like feeding organs (chelicera), which can pierce plant cells and are used to suck out their contents; feeding by members of the family Eriophyidae (Prostigmata), for example, causes the formation of galls, dwarfing of shoots, and malformation of fruits and leaves. The family Tetranychidae (Prostigmata) contains the spider mites, which are foliage feeders; several species, capable of producing silk, spin a light web over plant leaves.

The families Acaridae and Glycyphagidae (Astigmata), pests of stored grain and cereal products, have blunt, toothed chelicera that enable them to scrape and gouge their food material; certain species of these two families frequently increase to tremendous numbers in foodstuffs, causing "grocers' itch" or "copra itch" in man.

A large group of free-living predatory mites lives in

soil, humus, other organic matter, and water; these mites prey on small arthropods, their eggs, and other small invertebrates. Some species of Macrochelidae (Mesostigmata), for example, feed on housefly eggs or other arthropods; the Phytoseiidae (Mesostigmata) are important predators of plant-feeding mites. Predatory mites typically are long-legged, fast-moving acarines, with specialized chelicera. Ground-inhabiting species usually have well-developed hard plates, or shields; those found on vegetation are not so heavily sclerotized. Many aquatic species have long swimming hairs, or setae, on the legs.

Many free-living mites utilize insects or other arthropods to disperse themselves; this nonparasitic association is known as phoresy. Adults of the genus *Dinogamasus* (Laelapidae), for example, live in a special mite pouch (acarinarium) on an abdominal segment of certain carpenter bees.

Parasitic Acarina

Parasitic species are known in all acarine groups except the beetle mites (Cryptostigmata). The majority are external parasites, including those most important to man and his domesticated animals. A few species are internal parasites of animals. One ecological group contains families found only on the skin surface or feathers of the host. Psoroptidae (Astigmata), or scab mites, for example, attack the skin surface of mammals and feed on skin scales; the continuous abrasion of the skin causes a lesion over which a protective scab eventually forms. The families Analgidae, Proctophyllodidae, and many Dermoglyphidae (Astigmata), which are found on the feathers of birds and feed on body and feather scales, apparently do not harm the host. Different genera of feather mites prefer certain types of feathers and specific locales on the feathers.

A second ecological group contains several families of mites that burrow into skin, hair follicles, or quills of the host and feed on fatty secretions, lymph material, or, occasionally, blood. The families Sarcoptidae (Astigmata) and Psorergatidae (Prostigmata) contain species that burrow just beneath the skin of a mammalian host, causing mange or itch; Demodicidae (Prostigmata), Audycoptidae, and Rhyncoptidae (Astigmata) also occur in hair follicles or skin of mammals. Members of the family Ophioptidae (Prostigmata) form small pits in the scales of snakes; and Syringophilidae (Prostigmata) infest the quills of birds. The Knemidocoptidae (Astigmata), skin inhabitants of birds, may attack the legs of the host, causing a condition called scaly leg.

A third ecological group contains species that pierce the skin and suck up tissue fluid or blood without actually invading the tissues. These species may spend either short or long intervals on the host.

The larvae of Trombiculidae (Prostigmata), the chiggers, parasitize many different types of vertebrates and a few invertebrates, feeding on host tissues; chigger nymphs and adults, however, are free-living and predatory. Species of Dermanyssidae, Macronyssidae, and many Laelapidae (Mesostigmata) feed on blood or tissue secretions of mammals, birds, and reptiles. Some species, which spend much of their time off the host and in the nest, frequently are referred to as nest parasites, or nidicolous species. The protonymph of one species of Macronyssidae, found in the mouth of a long-nosed bat, causes destruction of tissues and loss of teeth; and a species of Spinturnicidae (Mesostigmata) is found in the anal opening of certain species of cave-dwelling bats. Most members of the tick families Argasidae and Ixodidae (Metastigmata) are obligate blood-sucking parasites of vertebrates; most Argasidae nymphs and adults feed on the host only a few minutes, but most stages in the life cycle of Ixodidae remain attached to the host for several days.

A few families of mites are found in additional parasitic associations with vertebrates. Rhinonyssidae, Entonyssidae, and Halarchnidae (Mesostigmata), for example, are respiratory parasites of birds, snakes, and mammals, respectively. Gastronyssidae (Astigmata) occur in the stomach, nasal openings, or eyes of bats, and the nonfeeding (hypopal) stage of certain species of Hypoderidae (Astigmata) occurs in the tissues of birds.

Several families of mites have established a parasitic relationship with invertebrates. Larvae of Trombidiidae and Erythraeidae (Prostigmata), as well as those of other families, are parasitic on insects; later stages are free-living. In contrast, nymphs and adults of the aquatic family Unionicolidae (Prostigmata) parasitize mollusks and sponges, and the larval stage is free-living. Certain species of Pyemotidae and Tarsonemidae (Prostigmata) and other families are parasitic on invertebrates throughout the life cycle. The breathing tubes (tracheae) of some Hymenoptera (bees, wasps) and Orthoptera (grasshoppers) may be invaded by certain species of Scutacaridae and Podapolipidae (Prostigmata); one species, for example, causes suffocation in honeybees (Isle of Wight disease). Otopheidomenidae (Mesostigmata), which occur in the auditory structures of moths and beneath the wings of bugs, may cause deafness in the moth host. Ewingidae (Astigmata) are parasites on the gills of crabs.

Form and function. *External features.* The subclass Acarina is distinguished by the lack of body segmentation (secondarily developed in a few families), a characteristic shared only with the spiders among the arachnids. An anterior region called the gnathosoma contains the mouth, specialized feeding appendages (chelicerae), and segmented structures called palps or pedipalps. The mouth or buccal cavity joins the pharynx internally; paired salivary glands may discharge into the mouth or in front of its opening. The chelicerae are basically three-segmented pincerlike appendages; as a result of the diverse feeding habits of some mites, however, chelicerae sometimes are modified as piercing organs or stylets. The palps, which may be simple sensory structures or predatory organs modified for grasping or piercing, usually have six segments—coxa, trochanter, femur, genu, tibia, tarsus; frequently there also is a clawlike apotele (modified distal segment of appendage).

Major body regions

Behind the gnathosoma is a large region (idiosoma), which bears the legs, the genital and anal openings, and an assortment of tactile and sensory structures; respiratory pores (spiracles) and sclerotized shields of various shapes and sizes usually are present. The functions of the idiosoma parallel those of the abdomen, thorax, and portions of the head of insects. Although nymphs and adults commonly have four pairs of legs, some Prostigmata (Eriophyidae, Podapolipidae, Tenuipalpidae) and Astigmata (Evansacaridae, Teinocoptidae) have one to three pairs. Legs have the same basic segmentation as palps; fusion or division of segments frequently occurs. The tarsus may be terminated either by several sensory hairs (setae) or by a clawlike or suckerlike apotele. The legs, which frequently bear ridges and spurs, always have tactile and sensory setae that follow a fixed pattern in position and number; the arrangement of leg setae is helpful in establishing systematic relationships. The first pair of legs usually functions in locomotion but sometimes is modified as a sensory or predatory structure.

The genital opening, usually located on the underside between the legs, frequently is protected by one or more shields or flaps and has two or three pairs of discs; in some Prostigmata (Cloacaridae, Demodicidae, Myobiidae, Ophioptidae, Podapolipidae), however, the male genitalia (aedeagus) is located on the dorsal side. The anal opening, also generally on the underside, is surrounded by a shield in the Mesostigmata and is always closed by a pair of valves; in some Prostigmata (Penthaleidae) and Astigmata (Chirorhynchobiidae, Knemidocoptidae), the anal opening is located on the dorsal side. As in all arthropods, the cuticle of acarines is secreted by an outer cell layer called the epidermis and contains numerous pore canals. The cuticles of many Acarina absorb water from the air, enabling them to avoid desiccation.

There are many different types of sensory receptors, most of which are setal. The setae, of many shapes and sizes, may be hollow chemoreceptors or solid, tactile structures. Other specialized setae, known as trichobothria, pseudostigmatic organs, eupathidia, or famuli, occur only in the Acariformes. A sensory pit called Haller's organ contains sensory setae and is found on the tarsal

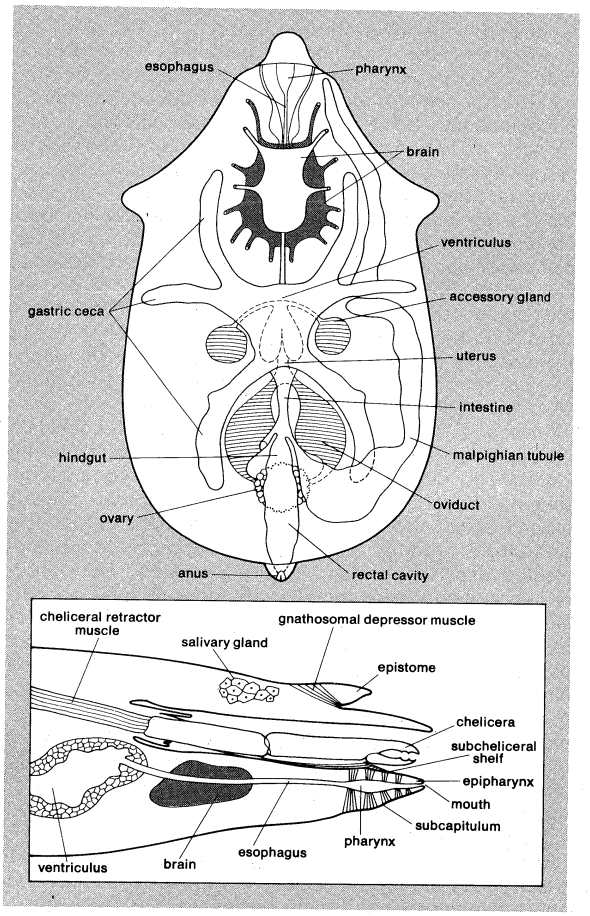


Figure 3: *Internal features of a mite.*
(Top) Dorsal internal aspect. (Bottom) Midsagittal section through the anterior region.
From G.W. Krantz, *A Manual of Acarology*

segment of the first pair of legs of all ticks (Metastigmata). One to three pairs of eyes are present on the anterior of the idiosoma in Notostigmata, a few Astigmata, and many Metastigmata and Prostigmata. A single median eye also may be present in some Prostigmata.

Internal features. One, three, or four pairs of spiracles (respiratory pores) are present in the idiosomal region of all acarine suborders except Astigmata and some Prostigmata and Cryptostigmata, which respire through the cuticle or have other types of respiratory organs. The spiracles open internally into a system of tubes (tracheae) that radiates throughout the body. Tracheae function in the exchange of oxygen and carbon dioxide and help to control water loss in certain species.

The idiosomal body cavity contains various organ systems bathed in blood, or hemolymph. Circulation of blood carrying the products of digestion and excretion may be by the movement of muscles or by a heart, which is present in some suborders. The blood of some soft ticks (Argasidae) differs from that of other acarines biochemically, although similar morphologically.

Digestive system

The digestive system consists of an anterior muscular pharynx, long narrow esophagus, ventriculus (stomach), short intestine, hindgut, and posterior rectal cavity. The ventriculus may have several paired gastric cecae (blind sacs), which function partly as food storage organs and enable some acarines to go long periods without feeding. One or two pairs of excretory organs (malpighian tubules), which may open into the hindgut, absorb waste material from the body cavity and transform it into an organic compound called guanine, which passes into the hindgut, is mixed with other residues, and eventually excreted from the anal orifice. Some Metastigmata and Prostigmata have no connection between the ventriculus and hindgut, and the latter serves as an excretory organ.

The male and female reproductive systems vary in structure and position. The gonads may be paired (primitive condition), united, or multiple (testes in some Prostigmata). There usually are accessory glands associated with the reproductive systems, especially in males.

A well-developed central nervous system surrounds the esophagus in the idiosoma. A series of nerves radiating from one part of the brain innervates the legs, digestive system, musculature, and genitalia. The mouthparts are innervated by another part of the brain.

Classification. *Distinguishing taxonomic features.* Features used by acarologists to distinguish the orders and suborders of Acarina include the presence or absence of spiracles; if present, the type, relative position, and structures associated with the spiracles are important. Other distinguishing features include the position and form of the palpal apotele (modified distal segment of palpus) when present, the type and location of sensory organs, the type of mouthparts, and the presence or absence of eyes.

Annotated classification. The ordinal classification presented here is after Gerald W. Krantz (1970). The process of evolution in mites and ticks has resulted in a degree of specialization and morphological diversity not found in any other group of arachnids. Acarina, believed to be derived from two lines of descent, have a long, though meagre, fossil history dating from the Devonian Period (about 395,000,000 years ago).

SUBCLASS ACARINA (Acari, Acarida; mites and ticks)
Arachnids usually without visible segmentation; mostly minute in size; larvae usually with 3 pairs of legs, adults usually with 4 pairs; diverse habitats include plants, soil, animals, stored foods, fresh and marine water; parasitic forms may transmit diseases; worldwide distribution; about 30,000 described species.

Order Opilioacariformes (mites)
Weakly sclerotized mites superficially resembling members of arachnid subclass Opiliones; 1,000–2,500 microns (μ) in size; body divided into hairy anterior portion bearing 2 or 3 pairs of simple eyes and nude posterior portion with over 200 pores, 4 pairs of spiracles, and 12 secondarily developed segments; terminal palpal apotele a pair of claws; rutellae (hypertrophied setae) present; all coxae of legs movable; first 2 pairs of legs with 6 segments, last 2 with divided trochanter, 7 segments; tritosternal base divided.

Suborder Notostigmata
Terrestrial, under stones and other debris in semi-arid habitats; recorded from southwestern U.S., South America, Central Asia, Africa, and the Mediterranean region; carnivorous and possibly omnivorous; of no economic importance; 1 family, Opilioacaridae, and about 11 species.

Order Parasitiformes (mites and ticks)
Small to large in size; usually heavily sclerotized mites; 1 or 2 pairs of spiracles on posterior portion of body; peritremes, or grooves, present or absent; palpal apotele present or absent; tarsi of 1st pair of legs with sensory organs.

Suborder Tetrastigmata
Heavily sclerotized; 2,000–7,000 μ in size; eyes absent; pair of coxal glands opening at base of coxae of 1st pair of legs; pair of spiracles behind coxae of 3rd and 4th pair of legs, peritremes present; palpal apotele present; terrestrial, under stones and decaying vegetation; recorded from Indo-Pacific region; carnivorous; of minor economic importance; 1 family, Holothyridae, and about 10 species.

Suborder Mesostigmata
Generally with a number of sclerotized plates; 200–2,000 μ in size; eyes absent; pair of spiracles between coxae of 2nd to 4th pair of legs; usually associated with elongated peritremes; palpal apotele present; tritosternum usually well developed but reduced to absent in some parasitic families; majority free-living in soil or decaying organic matter; many parasites of vertebrates (except amphibians and fish) and invertebrates; some economically important; cosmopolitan; about 60 families, 8,000 species.

Suborder Metastigmata (ticks)
Largest Acarina; adults 2,000–30,000 μ in size, eyes present or absent; pair of lateral spiracles enclosed in spiracular plate (modified peritreme) anterior or behind coxae of 4th pair of legs; palpal apotele and tritosternum absent; sensory organ on tarsus of 1st pair of legs a pit (Haller's organ); mouthparts modified to form hypostome (holdfast organ) with teeth turned backward; active stages mostly external parasites

(some nonfeeders), feeding primarily on blood of vertebrates (except fish); many members economically important as disease carriers; cosmopolitan; 3 families, Argasidae (soft ticks), Nuttalliellidae, and Ixodidae (hard ticks); about 850 species.

Order Acariformes (mites)

Small to large in size; generally weakly sclerotized mites (except most Cryptostigmata); palpal apotele and posterior body spiracles absent; anterior portion of body with or without sensory organs that are specialized setae or structures.

Suborder Prostigmata

Most heterogeneous suborder includes chiggers; 100–16,000 μ in size; weakly sclerotized with few plates; eyes present or absent; spiracles, when present, at or between bases of chelicerae or on anterior portion of body; chelicera stylet-like, chelate (pincerlike), or reduced; diverse in habitat and habits —terrestrial, aquatic, marine, parasitic, predatory, phytophagous; many species economically important; cosmopolitan; about 113 families, 10,000 species.

Suborder Astigmata

Homogeneous group includes mange, itch, or scab mites; weakly sclerotized and slow moving; 200–1,500 μ in size; eyes rarely present, spiracles absent; palps single segmented (sometimes with 2 false segments); chelicera chelate; true claws absent; rodlike sensory setae on tarsus of 1st pair of legs; wide range of terrestrial habitats; parasitic, predatory, or feed on decaying material; some species economically important; cosmopolitan; about 41 families, 6,100 species.

Suborder Cryptostigmata (orabatid or beetle mites)

Usually strongly sclerotized and slow moving, 200–1,500 μ in size; eyes and spiracles absent; pseudospiracles generally present, palpi without claws, 3–5 segmented; chelicerae usually chelate; rutella present; tarsi with 1–3 claws; ventrally with various shields; majority terrestrial in forest humus and soil, a few aquatic; feed on algae, fungi, or decaying material; of some economic importance; cosmopolitan; about 114 families, 5,000 species.

Critical appraisal. Acarologists are not in total agreement as to the classification of mites and ticks. Much of the disagreement concerns the levels at which taxonomic divisions should be made and the use of ordinal names; consequently, seemingly different classifications are actually more alike than they may initially appear to be. Knowledge of Acarina systematics is so incomplete that any classification proposed below the ordinal level is difficult and may quickly become obsolete.

In recent years, three major classifications, including the one used above, have been proposed for the mites and ticks. One classification, developed in 1952, considered the mites and ticks as members of the order Acarina, which consisted of five suborders (Onychopalpida, Mesostigmata, Ixodides, Trombidiformes, and Sarcoptiformes). Another system in 1961 gave the mites and ticks subclass rank under the name Acari, with seven orders, listed as suborders above.

BIBLIOGRAPHY. G.W. KRANTZ, *A Manual of Acarology* (1970), a recent systematic arrangement, with keys to families, brief notes on important species, and many illustrations; T.E. HUGHES, *Mites or the Acari* (1959), a good general discussion, with many illustrations and examples; G.O. EVANS, J.G. SHEALS, and D. MACFARLANE, *The Terrestrial Acari of the British Isles*, vol. 1, *Introduction and Biology* (1961), a good general discussion of biology and higher classification, with some emphasis on species occurring in the British Isles; E.W. BAKER and G.W. WHARTON, *An Introduction to Acarology* (1952), an important systematic review, with keys to families, lists of genera, and brief notes on biology; E.W. BAKER *et al.*, *Guide to the Families of Mites* (1958), an important systematic review, with keys to families and many figures, and *A Manual of Parasitic Mites of Medical or Economic Importance* (1956), a good general discussion of some economically important mites.

(N.A.W.)

Accelerators, Particle

Particle accelerators are devices in which charged particles such as electrons, protons, and other ions are accelerated to very high velocities. The unit of energy used in accelerator technology is the electron volt (eV), which is the energy acquired by a particle with an electrical charge equal to the charge of an electron, after passing through a potential difference of one volt. The larger scale measurements are: keV = 1,000 eV; MeV = 1,000,000 eV; and GeV = 1,000,000,000 eV.

Particle accelerators are used as sources of various types of radiation. They are capable of creating high-energy beams of electrons, protons, heavy ions and beams of secondary particles that result from the interaction of accelerated particles with matter; for example, hard gamma radiation, positrons, antiprotons, neutrons, various mesons, and neutrino emissions (see also PARTICLES, SUBATOMIC).

Use of accelerators

Accelerators are the main tools of contemporary experimental nuclear physicists, who use them for fundamental research on the nature of nuclear forces, the exploration of transuranium (heavier than uranium) elements, and the detection of new unknown particles. Accelerators are also used in metallurgy for flaw detection in thick metal products such as heavy castings, in technology for producing radioactive isotopes, in medicine for treating malignant tumours, and in the food industry for sterilizing food products.

Every accelerator must have access to a source of charged particles. Thermionic emission (emission of electrons from a heated surface) is usually used to produce electrons. The production of ions is a more complex process. Positively charged ions can be obtained by means of an arc or glow discharge in a gas. Under the influence of the discharge, neutral atoms give off one or more electrons, becoming singly or multiply charged ions. These ionized particles are drawn away from the source and directed into the accelerator. Acceleration occurs by means of an electric field that is created along the trajectory of the particles.

The path travelled by the charged particles in an accelerator can be either linear or cyclic. In linear accelerators the particles move in a straight line and pass through a series of accelerating devices. The final energy of the particles depends on the number of these devices and the magnitude of the accelerating voltage in each. In cyclic accelerators the particles move in a magnetic field, either in a spiral trajectory or in a trajectory that is very close to circular. In one revolution the particles pass by one or more of the accelerating devices. The final energy depends on the number of revolutions and the magnitude of the energy acquired during each revolution. Because the particles may make millions of revolutions and traverse a distance measured in hundreds of thousands of kilometres, provision must be made for the stable motion of the particles. This is achieved by proper design of the magnetic field (magnetic focussing).

To avoid collisions with air atoms, the acceleration is conducted in a vacuum chamber, in which the pressure of the remanent (residual) air does not exceed 10^{-6} or, in some machines, 10^{-9} millimetre of mercury.

Because of the large energies involved, accelerators are usually physically huge. The lengths of linear accelerators and the radii of cyclic accelerators approach kilometre distance. The large dimensions of such installations are combined with meticulous detail and fine adjustment. The design and construction of accelerators utilize the latest developments in advanced fields of science and technology, such as electronics, radio and electrical engineering, magnetic optics, and plasma physics.

HISTORY

The first man-made nuclear reaction was accomplished by the English physicist Ernest Rutherford in 1919 by irradiating a sheet of matter with a stream of alpha particles (nuclei of helium atoms). After that, numerous studies were carried out involving the bombardment of the nuclei of light elements with the help of particles emitted from naturally radioactive elements. Very soon, natural sources of fast particles ceased to meet the demands of experimenters. Research began on the methods of accelerating charged particles. The first experiments involving the fission (breaking apart) of lithium atoms with protons, using man-made accelerators of up to 500 keV, were carried out by the English physicists J.D. Cockcroft and E.T.S. Walton in 1932. Protons were accelerated in special constant-voltage tubes. The original

high-voltage electrostatic generator was completed in 1931 by R.J. van de Graaff of the United States. Voltages created with this device reached 1.5×10^6 volts. The Cockcroft–Walton and van de Graaff accelerators (discussed below in greater detail) were still being used in the early 1970s in a number of laboratories.

The idea for the linear resonance accelerator was first expressed in 1924, and an actual model of an accelerator of this type for the acceleration of alkali metal ions (sodium, potassium) was built in 1927 under the supervision of the Swiss physicist R. Wideröe.

An important step in the further development of accelerators was made by the American Ernest O. Lawrence, who devised a circular modification of the linear accelerator of Wideröe. In the Lawrence concept, particles moved along a spiral trajectory in a magnetic field, passing many times across the same accelerating space. The first device based on this principle, called a cyclotron, was built in 1932. It produced protons with energies of one MeV, while subsequent models were capable of producing energies of up to 25 MeV.

Developments in the 1940s

Until 1944 most of the work in nuclear physics was conducted with cyclotrons and electrostatic generators, but in the 1940s an electron accelerator, the betatron, actually invented by Wideröe in 1928, was developed in the United States. The betatron can produce electrons with energies approaching several hundred MeV.

A new stage in the development of accelerators began after the discovery (independently) by the Soviet physicist V.I. Veksler in 1944 and the American physicist E.M. McMillan in 1945 of the principle of automatic phase stability of particles. This principle initiated the creation of cyclic resonance accelerators: the synchrocyclotron (phasotron), the microtron, the synchrotron, and various other modifications. The significance of the principle of phase stability (see below *Proton linear accelerators*) lay in the fact that it eliminated the principal limitations on the ultimate energies attainable by accelerators. There appeared the possibility of obtaining beams of accelerated particles of any desired energy, thus increasing the yield of accelerators. Cost was the practical limit on energies that could be obtained.

The method of strong magnetic focussing (see below *Synchrotrons*) played a large role in accelerator technology. This technique was developed by several researchers in the early 1950s. The use of this focussing technique in cyclic accelerators allowed a significant reduction in the size of the beam of accelerated particles and in the weight of the accelerator's electromagnets and thus in the power required to operate the electromagnets. All contemporary cyclic accelerators of high and ultrahigh energies employ this principle. The method of strong focussing also finds use in linear accelerators. The discovery of new accelerator principles and general technological progress have brought a rapid increase in the energy capability of accelerators, which has grown on the average every six years, beginning with 1930, by a factor of 10.

Development of storage rings

In 1956 development began on an unusual type of accelerator, the so-called storage-ring device in which conditions are created for the motion of two accelerated particle beams in opposite directions but with similar momenta. When the particles collide, the total energy accumulated may be utilized for interactions as the centre of mass of colliding particles remains motionless. The storage rings made it possible to advance along the energy scale in investigating the microworld as compared with usual accelerators, where physical studies are performed with nuclei of motionless targets and the largest fraction of a particle's energy is wasted on the motion of centre of mass.

CONSTANT-VOLTAGE ACCELERATORS

The method of accelerating particles using constant voltage is the simplest of the various techniques in use. Acceleration takes place in a special vacuum accelerator tube between electrodes operating at a high voltage, V. The energy gained by the particle is equal to eV, where e is the charge of the particle. Forward-action accelerators are widely used because of their simplicity and the precision with which the energies and intensities of beams of accelerated ions can be controlled and the possibility of accelerating particles of any atomic number. More than half the accelerators existing in the early 1970s operated on the basis of the forward method of acceleration. The main problem is generating the high voltages required.

Voltage-multiplying devices. The first scheme for the multiplication of voltage was suggested in 1920 and perfected and applied in accelerator technology by Cockcroft and Walton in 1932. It involves a source of high alternating voltage and a network of rectifiers and filter capacitors. By suitable arrangement of the number and size of the elements in the network, a direct-current output voltage can be obtained that is any desired multiple of the original alternating voltage. The accelerator tube is connected to the constant output voltage of this network.

Cockcroft–Walton accelerator

The voltage available for the accelerating tube decreases as the current of accelerated ions increases, because of the increased load on the voltage-multiplier network. The drop of voltage can be decreased by increasing the values of the capacitors in the network and by increasing the frequency of the input voltage.

The limiting voltage for a setup of this kind depends on the ability of the surrounding medium to withstand high voltages. In the case of air, higher pressures permit higher voltages; therefore, to maintain greater voltages, cascade generators are frequently enclosed in special high-pressure containers.

Cascade generators can develop beams of ions with energies to four MeV and a beam power of up to 20 kilowatts.

Electrostatic generators. In these generators the power to the high-voltage electrode is conveyed on a moving belt of insulating material, which carries charged particles from the grounded to the high-voltage electrode. The first model of an electrostatic generator for accelerating charged particles was built by van de Graaff in 1929, although a belt conveyer of charges for another purpose was first carried out in 1925 by B.N. Urgimov, and the concept of charge-carrying drops of water had been proposed much earlier by Lord Kelvin. The basic scheme of the van de Graaff generator is shown in Figure 1. The belt is mounted between two pulleys, one of which is electrically grounded while the other is located within a hollow metallic chamber, for example, of spherical shape

The van de Graaff generator

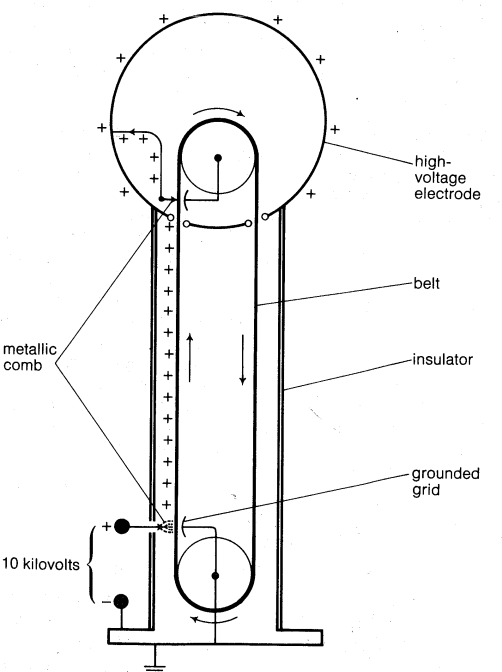


Figure 1: Simplified version of a van de Graaff high-voltage electrostatic generator (see text).

which serves as the high-voltage electrode. The sphere is mounted on the column, which is made of a good insulating material. The electrical charges build up from the side of the grounded pulley on the belt by means of corona discharge between a comb-like metallic needle structure, operated at about 10,000 volts, and the grounded grid located on the other side of the belt. Ionization of the air takes place under the effect of the high voltage near the tips of the teeth of the comb, and positive ions are directed against the surface of the grounded grid. Most of the ions are intercepted by the belt and settle on it. By changing the voltage polarity on the needles, both positive and negative charges can build up on the belt. Through the rotation of the pulley the charges are transported to the high-voltage electrode, where they are removed by the second comb of needles, which is electrically joined with the inner surface of the sphere. The charges are distributed on the sphere in such a way that the electrical field within the sphere is negligible and does not impede the runoff of charges on the sphere, regardless of the electrical potential. The potential of the sphere can increase to the point of electrical discharge through the air. The sparking field in dry air under indoor temperature and atmospheric-pressure conditions is approximately equal to 30 kilovolts per centimetre. On a sphere with a radius of one metre (3.3 feet) it is possible to obtain, under optimum conditions, a voltage of 3,000,000 volts.

Distortion of the symmetry of the electrical field of the sphere as a result of the comb and openings for the accelerating tube and transporting belt lowers the theoretically achievable voltage by more than a factor of 2. The voltage can be increased by mounting the device in a special container filled with high-pressure air or some other suitable mixture of gases. The achievable voltage is increased if a small proportion of carbon tetrachloride or Freon is added to the air. Sometimes a mixture of 80 percent nitrogen and 20 percent carbon dioxide under a pressure of 20 atmospheres is employed.

Contemporary electrostatic generators produce ion beams with energies to 20 MeV and with power capabilities to 1.5 kilowatts.

Tandem accelerators. The capabilities of the linear method of acceleration were significantly extended by the tandem method of accelerating heavy particles, developed in 1936 and based on the re-use of a single constant-high-voltage source.

In electrostatic accelerators, positive ions are accelerated from the high-voltage electrode to the grounded electrode, at which point they have acquired their maximum energy. In tandem accelerators, negative ions are first accelerated, moving from the grounded part of the device to the high-voltage electrode or positive potential. In transit across a channel inside the electrode, the negative ions are changed to positive ions and, moving further from the positive electrode to the opposite grounded part of the device, are accelerated a second time (from which derives the name tandem).

The two-stage tandem accelerator

The mechanism of acceleration is illustrated in the schematic of a two-stage tandem accelerator (Figure 2). Positive ions, generated at the point marked ion source at the right of Figure 2, leave the source with small velocities and are drawn into a boosting channel by electrons, where they pass through a tiny jet of gas (hydrogen, steam, mercury). Through reaction with the gas, some of the ions capture an excess number of electrons and become negatively charged ions. Then, by an analyzing magnet, the beam of selected negative ions is drawn into the main chamber and is accelerated by the positive high-voltage electrode in the first part of the device. The electrons are stripped away in the channel of the high-voltage electrode by a second jet of gas. The positive ions that have been generated continue to be accelerated in the other part of the device. The super-accelerated ions are then passed through a second analyzing magnet, which deflects in a predetermined direction only those particles that have the desired energy. These ions are then used to bombard various targets or for other experimental purposes.

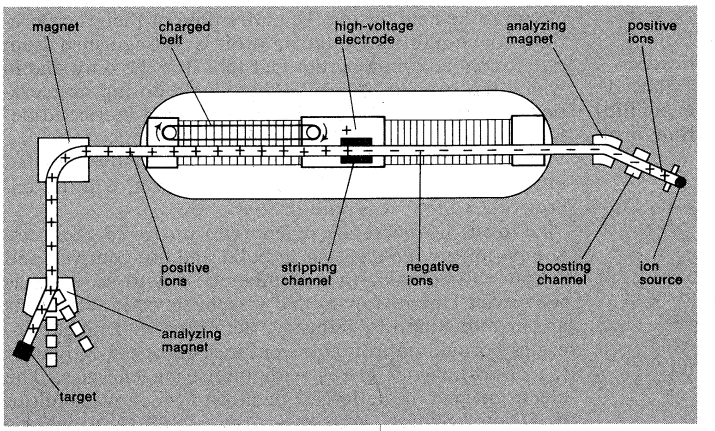


Figure 2: Two-stage tandem accelerator (see text).
From R.T. Van de Graaft, *Nuclear Instruments and Methods*, vol. 8, p. 195 (1960); North-Holland Publishing Co.

The voltage supply at the high-voltage electrode is created through the use of a charging belt or a high-voltage transformer and rectifier. The efficiency of altering the charge on both negative and positive ions approaches 75 percent.

Three- and four-stage variations of the tandem accelerator also exist. In electrostatic accelerators, the ultimate energy for protons is around 10 MeV, while the corresponding value for tandem generators is 30–40 MeV. Such accelerators have wide application for various purposes in many countries. In the early 1970s there were more than 50 tandem accelerators in use throughout the world.

BETATRONS

Betatrons are useful for accelerating electrons (beta particles, from which derives the name betatron). The basis of operation is that of a transformer in which the secondary winding is replaced by an electron beam that revolves in a toroidal (doughnut-shaped) vacuum chamber placed between the poles of the electromagnet. A cross section of a betatron is shown in Figure 3. An alternating magnetic field, which is concentrated mainly in the narrow gap in the electromagnet, induces in the circular chamber an electric field with closed force lines (electromagnetic induction). By axial symmetry of the magnetic current, the force lines of the electric field appear as concentric circles, the centres of which lie on the axis of symmetry. One of these circles coincides with the circular trajectory

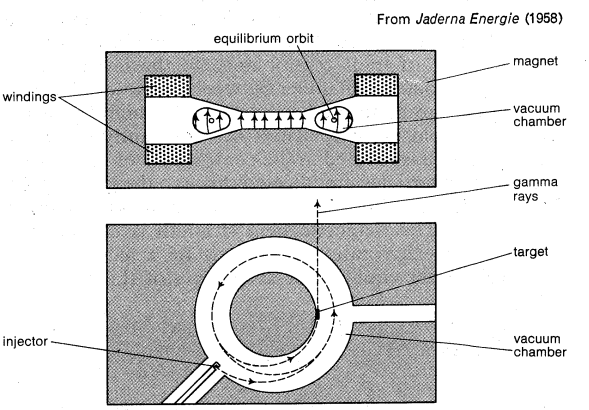


Figure 3: *Betatron.* (Top) Cross-section, (bottom) plan view (see text).

of the particle (its orbit). Thus, a force will act on the particle at every point of its orbit in such a direction as to produce the desired acceleration.

In true accelerators the increase in energy is around 100 eV per revolution, and particles make up to 1,000,000 revolutions during the time of acceleration. The betatron method of acceleration is possible only under the existence of a so-called equilibrium orbit, in which the parti-

Forces balanced in equilibrium orbit

cles must be held during the entire time of their acceleration. An equilibrium betatron orbit is distinguished from other circular orbits by the fact that it is the only one in which the increase of centrifugal force during acceleration will be automatically balanced by the increased deflecting force of the magnetic field.

The induction method of acceleration initially attracted attention in 1922. The first actual betatron, with an energy of 2.3 MeV, was built in 1940.

Betatrons in the range of energies up to 20 MeV are comparatively inexpensive to build and are compact and simple to maintain. In a number of countries, such as the Soviet Union, Japan, and Czechoslovakia, betatrons are manufactured by industry. Betatrons with higher energies become unprofitable ventures, since the weight of the electromagnet grows with increasing energies. The electromagnet in the largest betatron (which was built in the United States), with an energy of 300 MeV, weighs 330 metric tons (330,000 kilograms), for example, while that in a betatron of 20 MeV weighs about four tons. There is also a limit to attainable electron energies in betatrons. This limitation arises as a consequence of the emission of strong electromagnetic radiation by electrons due to their motion in a circular orbit (synchrotron radiation). The power of this radiation rapidly increases as the energy of the electrons increases. In betatrons the energy loss due to radiation cannot be compensated. At electron energies high enough to produce radiation, the radiation leads to disturbances of the particle motion along the equilibrium orbit with a resulting loss of particles. The ultimate energy for electrons in betatrons is about 500 MeV.

CYCLOTRONS

Cyclotrons (from the Greek word for "round" or "circle") are cyclic resonance accelerators that permit the acceleration of protons, deuterons, alpha particles, and multicharged ions. The motion of the particles proceeds along a spiral orbit in a fixed magnetic field. The acceleration of ions results from a high-frequency electric field.

Cyclotrons are used with a number of basic devices for the creation of fundamental nuclear phenomena and have found wide application in many countries. They are used in the technology of tracer atoms in medicine, biology, and chemistry. Development of the cyclotron followed the success in increasing the energies of ions and the intensities of accelerated beams. The deciding role in the development of cyclotron accelerators, however, was played by the principle of autophasing (automatic phase stabilization).

The classical cyclotron. A simplified schematic drawing of the cyclotron is shown in Figure 4. Two semicircular, hollow metallic electrodes of the same shape (called dees because of their resemblance to the letter *D*) are placed in a vacuum chamber between the poles of a magnet. These are connected to a source of high-frequency voltage so that their polarities are constantly changing. An ion source is located in the central part of the apparatus, and the particles are injected through the gap between the dees in the direction perpendicular to the magnetic force lines. The magnetic field strength is chosen such as to provide a rotation frequency of a particle along the orbit either equal to or a multiple of the frequency of the applied voltage. Under such conditions, a certain fraction of ions will cross the gap twice per revolution at the required polarity of voltage, thus each time increasing the energy of these ions. This process continues, giving the particle a higher and higher energy. The changes in polarity of the accelerating voltage take place when the particles move through the inner cavity of the dees where there is no electric field. An increase in the velocity of the particles leads to an increase in the radius of the circle, and the particles are moved by acceleration along a spiral from the centre to the periphery of the magnet. In classical cyclotrons there is a theoretical limit to the amount of energy that can be imparted to a charged particle, based on the fact of relativity that the mass of a particle increases as its velocity increases. The maximum energy of protons in classical cyclotrons does not exceed 25 MeV.

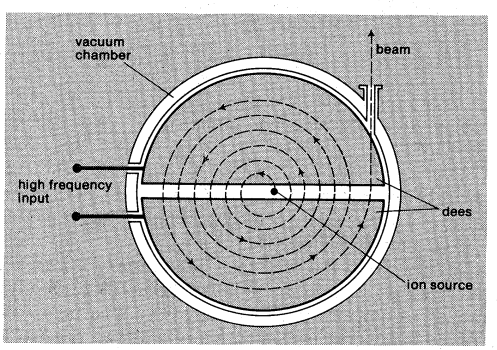


Figure 4: Plan view of the classical cyclotron.

Stability in the motion of a particle is achieved by creating a slight drop in the magnetic field from the centre to the edge of the magnet. In the central part of the accelerator the magnetic field does not depend on the radius, and magnet focussing does not exist.

Synchrocyclotrons (phasotrons). These devices differ from the classical cyclotron in that the frequency of the accelerating voltage changes. The motion of the particle proceeds as an uncoiling spiral, as in the classical cyclotron. Stability of the particle's motion is provided for by means of a decreasing magnetic field from the centre to the periphery. The frequency of the accelerating voltage decreases according to the growth of energy of the ion. This removes the energy limits characteristic of the cyclotron, but the arrangement becomes that of an impulse accelerator. The particle is engaged in the stabilized phasotron regimen of acceleration in the centre of the magnet, where the ion source is located, when the frequency of revolution of the ion is equal to the frequency of the accelerating voltage. After completion of the accelerating cycle for a given batch of particles, the frequency of the accelerating voltage is increased to its original value, and the next batch of particles from the source starts on the accelerating cycle. Impulse operation sharply reduces the average intensity of accelerated particles to that of classical cyclotrons.

Particle motion in the synchrocyclotron

Physics research utilizes beams of secondary particles (neutrons, pi mesons, and others) that are produced through the reaction of accelerated ions with target materials placed in the last loop of the orbit or accelerated ions that have been extracted from the synchrocyclotrons.

The energies of ions attainable by synchrocyclotrons are determined by the money available for construction. The practical limit on the diameter of the pole-pieces of an electromagnet with a magnetic field intensity in the upper usable limit is five to six metres (16 to 20 feet). This produces a maximum proton energy of about 1,000 MeV.

In 1948, in the synchrocyclotron at Berkeley, California, the first production under laboratory conditions of pi-meson nuclear particles was achieved through the bombardment of matter with alpha particles accelerated to 380 MeV. Since that time, synchrocyclotrons have been successfully used for nuclear research in many scientific centres: Berkeley (730 MeV); Dubna, Russian S.F.S.R. (680 MeV); Geneva, Switzerland (600 MeV); Pittsburgh (450 MeV); Chicago (450 MeV); and Liverpool (400 MeV). The average current of accelerated ions in the most powerful accelerators approaches five to ten microamperes.

Production of pi-meson particles

The sector-focussing cyclotron. This device creates a continuous regimen of acceleration. The synchronism between the period of revolution of the particles and an accelerating voltage with a fixed frequency is achieved through increasing the magnetic field from the centre to the outer perimeter of the cyclotron.

An increasing magnetic field focusses particles adequately in a radial direction but defocusses them in a vertical direction. In order to stabilize the motion in a vertical direction, an axially asymmetrical magnetic field is used, which is dependent not only upon the radius but

also upon the angular coordinate along the orbit. There exists a large number of configurations of the magnetic field that satisfy the condition for vertical focussing. The required magnetic field is created by means of fairly complex magnets.

Most sector-focussing cyclotrons are built to provide ion energies of up to 100 MeV. Unique devices with ion energies of up to one GeV and with intense proton beams, the average current of which approaches milliamperes, are planned in a number of meson factories. To this number belong, for example, the relativistic cyclotron of 700 MeV (Dubna), the Mc^2 cyclotron of 900 MeV (Oak Ridge, Tennessee), and the H^- cyclotron of 625 MeV (Los Angeles).

Separated-orbit cyclotrons. These combine the advantages of cyclotrons with sector focussing and linear accelerators. Separated-orbit cyclotron accelerators, including most strong-current cyclic proton accelerators, are capable of providing energies on the order of 1,000,000,000 eV. Schematically, the separated-orbit cyclotron can be presented in the manner of a linear accelerator that is coiled in a spiral with large separations. Ions are accelerated by a high, fixed-frequency resonator field. The particles pass the same resonator in each loop. The resonators are built in such a way that they affect all loops of the spiral orbit. The orbit is formed by the effect of fixed-field electromagnets, which are situated among the resonators. Focussing of the particles is provided by alternating magnets with positive and negative magnetic field gradients. The independence of the period of revolution of the particles of the number of loops of the spiral is achieved by changing the magnitude of the magnetic field along the orbit proportional to the increase in the total energy of the particles.

Uses of separated-orbit cyclotrons

The commercial value of accelerators of this type is seen in their use as sources of neutron radiation for creating isotopes, for producing and breeding nuclear fuels, and for performing experiments in nuclear physics that demand large intensities. Neutrons are produced by bombarding targets of heavy elements with beams of accelerated protons. The projected fluxes of thermal (low-speed) neutrons approach 10^{16} neutrons per square centimetre per second. The required power of proton beams, accelerated to energies of one GeV, approaches tens of megawatts.

LINEAR RESONANCE ACCELERATORS

From 1934 to 1946 work on linear accelerators practically ceased, since all of the attention of scientists was directed to cyclotrons. Major advances in the technology of short radio waves in the 1940s served for a time as a stimulus to further developments.

Linear resonance accelerators are divided into two groups: those used to accelerate lightweight particles (electrons, positrons) and those for heavy particles (protons, deuterons, multicharged ions). The main difference between the mechanisms lies in the fact that an energy of several MeV will accelerate light particles rapidly up to close to the speed of light. Further increases in energy, in accordance with the theory of relativity, result in increasing the mass of the particle with little increase in its velocity. Heavy particles, on the other hand (protons, for example), reach velocities approaching that of light only through applying energies of several GeV. This accounts for the difference in the construction of linear accelerators for light and heavy particles.

Proton linear accelerators. Accelerators in this class are frequently built according to a scheme presented by Luis W. Alvarez in 1946, in which resonator containers consisting of long cylinders, having walls of a good electrical conductor such as copper, are used. In resonators of this type a high-frequency electromagnetic oscillation is set up with a wavelength that is 1.3 times the resonator diameter. Such an oscillation produces a linear electrical field that has its maximum on the axis of the resonator. The technological limitation on the average intensity of the electric field is 15–20 kilovolts per centimetre. Therefore, accelerators with energies of tens and hundreds of MeV have considerable lengths. At reasonable diameters (one to two metres) the period of oscillation is so short that during the time of flight of the protons across the resonator the electric field changes its direction many times and will accelerate the particles at times while decelerating them at other times. To avoid this occurrence, a large number of metallic drift tubes, separated by intervals of space, are placed along the axis of the resonator (Figure 5). The length of the tubes and the distances between them are so calculated that the particles have passed through the tubes, where the electric field does not penetrate, when the electric field has a decelerating direction (decelerating phase) and are accelerated in the spaces between the tubes. Exact calculation of the spacing between drift tubes is possible only for the so-called equilibrium particles, which pass all the spaces at the same phase of the electric field (equilibrium phase), thus gaining each time the same energy. All the other particles cross the initial spaces at smaller or larger values of the electric field, being accelerated more or less rapidly than the equilibrium ones. The larger fraction of these particles will not go out of synchronism with the high frequency electric field and will be accelerated to the end. For the slower particles, the time of flight between spaces is greater and they will pass the following spaces at increasing electric field, thus catching up with the equilibrium one. Similarly, the faster particles will be slowed down until they reach equilibrium velocity and phase. This process of gathering particles around the equilibrium phase is called autophasing.

Use of drift tubes to improve operation

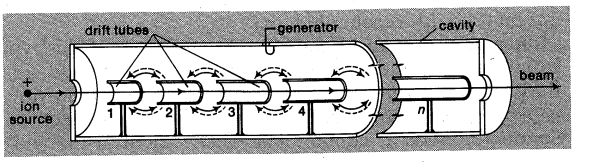


Figure 5: Proton linear resonance accelerator containing *n* metallic drift tubes (see text).

Beams of particles injected into the accelerator inevitably tend to spread out and can be lost in the drift tubes. Magnetic lenses in the device prevent such loss by minimizing spreading.

The proton linear accelerator provides a pulsed output. The duration of the impulses and the frequency of their repetition are restricted by the power available from the generators that feed the resonators. Power on the order of tens of megawatts per pulse is necessary to produce the required electrical field intensities.

Proton linear accelerators are in use that generate particle beams with average currents up to about one milliampere. They are widely used as injectors for cyclic proton accelerators of larger energies. Self-contained devices that will produce energies of 0.5–1 GeV and are up to 1,500 metres (4,900 feet) long are planned and are under construction. These devices are planned to be used for the generation of intensive beams of pi and mu mesons (the meson centre in Los Alamos, New Mexico) and for developing powerful neutron beams up to 10^{19} neutrons per second (Chalk River, Canada).

Heavy-particle linear accelerators. These are analogous to proton accelerators. The difference lies in the fact that the accelerated ions usually have a charge-to-mass ratio greater than that of the proton. Using suitable ion sources it is relatively easy to obtain heavy ions with charges of three to five times that of the proton. Before their introduction into the accelerator, ions are accelerated to an energy of several hundred MeV (in, for example, a Cockcroft–Walton accelerator). Frequently, linear accelerators of heavy ions are divided into two sections, between which the beam of particles passes through a thin metallic foil or mercury vapour for supplementary "stripping" of electrons from ions in order to increase the charge by several units. This permits higher energy gains in the second section. In the first section, ions are accelerated to an energy of approximately one MeV.

At Berkeley there is an accelerator of helium, carbon, and nitrogen ions and of heavier elements with ten MeV

per nucleon (a nucleon is an integral part of the nucleus —a proton or a neutron). The length of the first section is 4.5 metres (15 feet) and of the second, 27 metres (89 feet). The sections have 87 and 67 drift tubes, respectively. The resonators operate at a frequency of 70 megahertz (70,000,000 cycles per second). Fundamental research on the synthesis of new transuranium elements is conducted at many laboratories with the help of heavy-ion accelerators. New elements are formed by bombarding matter with accelerated ions.

Berkeley heavy-particle accelerator

Linear electron accelerators. Accelerators of this type capable of producing high energies are constructed similarly to wave guides. A travelling electromagnetic wave is propagated along a tube with walls of a good electrical conductor, such as copper. The electric field of the wave is directed along the axis of the tube. If a beam of electrons with a velocity equal to the phase velocity of the wave (the speed of translation of a crest of the wave) is directed into the wave guide, then the electrons, having hit the crests of the electric-field intensity, will be moved along with them and will be accelerated. The mechanism of acceleration is similar to that of surfing on the ocean, riding a board on the travelling waves of water. In a wave guide with smooth walls the phase velocity of a wave exceeds the velocity of light in free space. Lowering this value to equal the velocity of the electrons is achieved through the introduction of a series of metallic diaphragms, as shown in Figure 6.

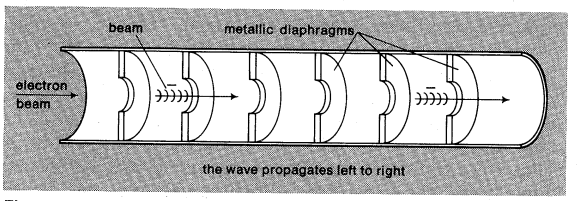


Figure 6: Waveguide construction in an electron linear accelerator showing introduction of metallic diaphragms to decrease phase velocity (see text).

Most of the bunching occurs at the beginning portions of the wave-guide accelerator, where the velocity of the electrons grows markedly because of acceleration. At this point the phase velocity must increase to match the increase in electron velocity, while the velocity of the electrons is much less than the speed of light. Farther along the tube, the design provides a constant phase velocity. Wave-guide accelerators usually operate at wavelengths in the range of tens of centimetres. A device at the end of the wave guide absorbs the energy of the travelling wave, preventing reflection of the wave back along the tube.

Increases in the amplitude of the electric field are limited by the availability of high-frequency power and are dependent upon the quality of the machining of the surfaces of the walls of the wave guide and the degree of vacuum. The maximum attainable energy increase for a wave guide of one metre length is about 15 MeV.

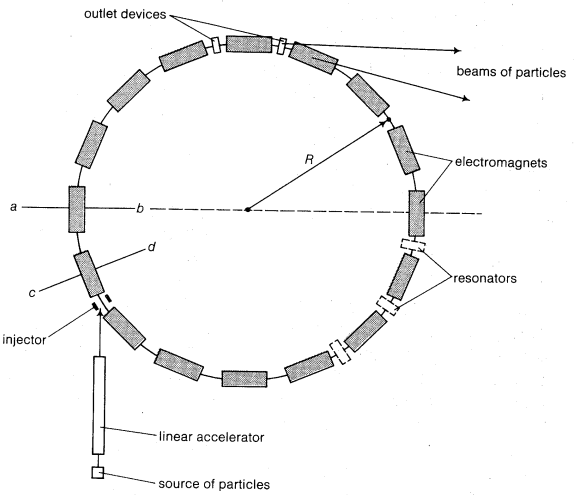
The main advantage of linear electron accelerators over cyclic accelerators lies in the almost complete absence of energy loss by the particles through radiation. Increasing the length of the accelerator can, in principle, produce electrons of as large an energy as desired.

Largest linear accelerator

The largest linear electron accelerator, with a 21 GeV energy range, was built at Stanford, California, in 1966. Its length is 3,050 metres (10,000 feet); the diameter of the wave guide, ten centimetres (four inches); the length of the travelling wave, 10.5 centimetres. The frequency of pulse repetition is 136 per second, with a duration of 10^{-6} second; the pulse current is 35 milliamperes. The accelerator can also accelerate positrons to energies of 15 GeV.

SYNCHROTRONS

The name synchrotron (from the Greek word meaning "simultaneous") is used for a number of cyclic resonance proton and electron accelerators. The orbit of the particles is formed by means of magnetic fields that increase over time proportional to the increased momentum of the particles, while the radius during acceleration remains constant. In this lies the main advantage of the synchrotron; *i,e.*, the magnetic field is needed only in a narrow circular area around the orbit. This allows a reduction in the total weight of the magnet and a significant reduction in the power required to operate the electromagnets. Focussing the particles in a direction lateral to the orbit is achieved by properly defining the magnetic field. Weak and strong focussing are different, however. In the first case, the field depends on the radius, much as in the betatron. Weak focussing acts simultaneously in both the horizontal and vertical directions. In the case of strong focussing, the orbit is formed by a large number of electromagnets, as shown in Figure 7. Two types of



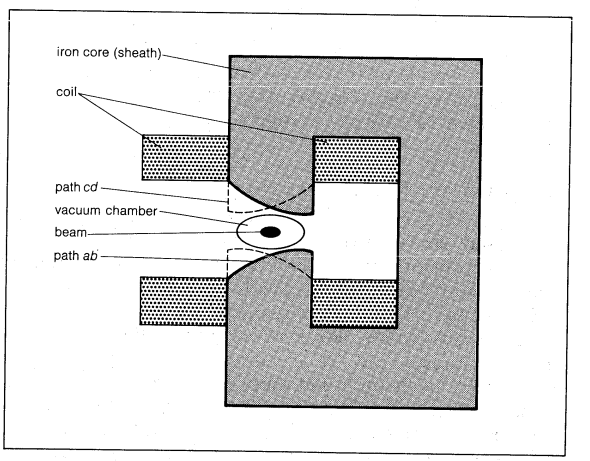


Figure 7: Synchrotron with strong focussing, showing placement of the two types of magnets. Shown at bottom is a cross section of a typical magnet for the two different pole structures used in a synchrotron (see text).

magnets alternate; in one the magnetic field sharply increases along the radius R and in the next sharply decreases. In magnets with increasing field, strong focussing arises in a radial direction, and strong defocussing arises in a vertical direction. In magnets with decreasing field, the opposite occurs. The two paths are indicated by *ab* and *cd* in Figure 7. The overall effect of the series of focussing and defocussing magnets on the beam of particles is to form a well-focussed beam.

Particles are accelerated in a linear accelerator prior to injection into the synchrotron. Farther along the channel, which consists of a series of magnetic lenses and deflecting magnets, the beam meets an injector device that directs the beam along a tangent to the orbit. The orbit presents itself as a closed line, consisting of a circular array of magnets with spaces between them. Filling

the perimeter of the orbit with particles by injection can be completed in the time of one or several revolutions.

Use of resonators to produce acceleration

Acceleration of the particles is accomplished by resonators placed in the gaps between the magnets. The motion of the particle along the orbit and the changing of the voltage in the resonators is arranged so that the frequency of voltage alternation is equal to or a multiple of the frequency of revolution of the particles. At first glance it seems that only particles of the same velocity and of the proper phase can be accelerated and that the intensity of the accelerated beam will be infinitesimally small. Under certain conditions, however, a large portion of the other particles will join the accelerated beam, the result of the phenomenon of autophasing, described previously.

The particles gain a small amount of energy each revolution, on the order of tens or hundreds of keV; the complete cycle of acceleration continues over the course of hundreds of thousands of revolutions. Particles are accelerated in a vacuum chamber made of nonmagnetic material. The chamber is placed in the gap between the poles of the magnets.

At the end of the acceleration cycle the beam of particles is withdrawn from the accelerator by means of deflecting devices or is directed against a special target placed in the vacuum chamber for the generation of various secondary particles (*e.g.*, gamma radiation, meson beams, neutrino radiation, and others).

In the relativistic case, the energy of the particle is proportional to the product of the magnetic field intensity and the radius of the particle orbit in this magnetic field. The maximum value for the magnetic-field strength of the electromagnet cannot exceed the saturation of the iron core. The higher the given energy for accelerated particles, therefore, the greater the radius needed for the ring electromagnet of the synchrotron. For an energy of 100 GeV the radius is 200 metres (660 feet).

Electron synchrotrons. The injection of electrons into an electron synchrotron takes place at relativistic energies on the order of ten MeV, at which energy the velocity is close to that of light. During the time of acceleration, the velocity of the particles—and consequently the frequency of revolution along the orbit—is increased by only a very small amount. This permits the introduction of a resonator operating at a fixed frequency.

The acceleration of electrons to large energies in synchrotrons is made difficult by the generation of strong electromagnetic radiation, the spectrum of which extends from radio waves to X-rays. The radiation is caused by the motion of the particle along a circular orbit. It is frequently called synchrotron radiation. The power of the radiation increases very rapidly as the energy increases. At energies of 10 GeV, each electron gives off radiation equivalent in energy to about 30 MeV per revolution, and a major portion of the power in the accelerating system goes toward compensating for this loss. For this reason, the use of synchrotrons for obtaining electrons with energies greater than 10–15 GeV is impracticable (see Table 1).

Table 1: Operating Electron Synchrotrons with Energies Greater than Five GeV

scientific centres	Cambridge, Mass.	Hamburg	Daresbury, Cheshire	Erevan, Armenian S.S.R.	Ithaca, N.Y.
Energy (GeV)	6.28	7.5	5	6.1	10
Intensity (electrons per second)	6×10^{12}	5.6×10^{12}	1.2×10^{13}	3.5×10^{12}	1.2×10^{12}
Radius of the orbit (metres)	37.5	25.2	35.1	34.5	125
Size of the chamber (centimetres)	15 × 4	11 × 5.4	13.5 × 6	12 × 4.2	5.5 × 2.5 (gap of magnet)
Frequency of repetition of the accelerating cycle (in seconds)	60	50	53	50	60
Weight of the magnet (tons)	336	647	400	427	226
Year of installation	1962	1964	1966	1967	1967

Source: Materials of the VII International Conference on Accelerators, 1969, Erevan.

Proton synchrotrons (synchrophasotrons). These devices are suitable for accelerating protons to very high energies. The ultimate limit on energy is determined by economic considerations. Injection of protons takes place at energies of 50–200 MeV and with velocities of 0.3 to 0.6 times the velocity of light. Unlike electron synchrotrons, therefore, the frequency of the accelerating voltage necessarily increases during the time of acceleration in order to match the increase in velocity of the protons. The velocity of the protons increases rapidly at the beginning of the acceleration and reaches a value of 0.98 times the velocity of light at an energy of two GeV. At the present time, only proton synchrotrons open the practical possibility of obtaining intense beams of particles with energies of several hundred GeV and in the near future with energies of 1,000 GeV and higher. In a number of the world's scientific centres, proton synchrotrons are basic tools for fundamental research on the interaction of particles at high energies (see Table 2).

Table 2: Operating Proton Synchrotrons with Energies Greater than Ten GeV

scientific centres	Dubna, Russian S.F.S.R.	Geneva	Brookhaven, N.Y.	Argonne, Ill.	Serpukhov, Russian S.F.S.R.
Energy (GeV)	10	28	33	12.7	76
Intensity (protons per second)	1.5×10^{10}	7.5×10^{11}	1.5×10^{12}	10^{12}	2×10^{11}
Frequency of repetition of the accelerating cycle (impulses per minute)	6.6	30	24	30	8
Radius of the orbit (metres)	36	100	128.5	27.4	236
Size of the vacuum chamber (centimetres)	120 × 35	15 × 7.4	17.4 × 7.9	82.2 × 14.6	19 × 11.5
Weight of the magnet (tons)	38,700	3,130	4,400	4,768	20,700
Focussing	weak	strong	strong	weak	strong
Year of installation	1957	1959	1960	1963	1967

Source: Materials of the VII International Conference on Accelerators, 1969, Erevan.

OTHER TYPES OF ACCELERATORS

Microtron. This is a cyclic resonance electron accelerator with a constant magnetic field and a fixed-frequency accelerating voltage. In these features it is similar to the cyclotron; sometimes it is called an electron cyclotron. Designed in 1944, it appeared historically as the first accelerator that was based on the principle of autophasing. The synchronism between the frequency of revolution of the particles in the magnetic field and the accelerating voltage is achieved by increasing the period of revolution (and thus the radius) of the particles each revolution by an integral number of accelerating voltage periods.

The microtron consists of a fixed-field electromagnet, a resonator that is placed between the poles of the electromagnet, and a generator that causes disturbances in the resonator. The source of electrons (for instance, a tungsten filament) is usually mounted in the resonator. There is an aperture in the resonator for the flight of electrons, whose orbits are circles with common tangents at a point inside the resonator (Figure 8). The radius of each subsequent orbit is greater than the preceding, depending upon the increase of energy the particles receive in the reso-

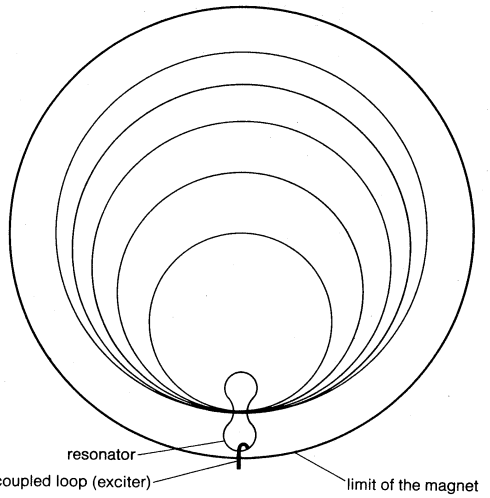


Figure 8: Typical orbits of particles in a microtron (see text).

From *Principles of Cyclic Particle Accelerators* © 1961 by Litton Educational Publishing, Inc. Reprinted by permission of Van Nostrand Reinhold Company

nator. To reduce the dimensions of the microtron, a strong magnetic field is chosen (the radius of the orbit is inversely proportional to this field). This requires an increase in the energy added during each revolution, making it necessary to excite the resonator with frequencies in the microwave range (from which the name microtron arises). In actual devices the dimensions of the microtron are relatively large. The microtron uses a pulse mechanism, since the needed voltages in the resonator cannot be maintained continuously. The duration of the pulses reaches 3.10^{-6} second at a repetition rate up to 2,000 times a second. Electrons make several dozen revolutions during the time of acceleration. A strong orbital deflection in the chamber permits easy removal of the particles from their final orbit. Microtrons with uniform magnetic fields do not focus in a vertical direction, so that in the final stages of acceleration focussing fields are developed at the periphery of the chamber by means of sectional magnets.

Microtrons are built with energies in the range of tens of MeV and are used as self-contained devices in physical research, such as injectors in a number of electron synchrotrons. Pulse currents (of electrons) in powerful microtrons reach 100–200 milliamperes.

Strong-focussing accelerators. Strong-focussing accelerators with fixed magnetic fields (ring phasotrons, fixed-field, alternating-gradient devices) have a characteristic feature in their use of strong focussing, where the magnetic field along the orbit changes in magnitude and direction. The principle was announced at Brookhaven National Laboratory in 1952 and developed in 1953 by the Soviet scientists A.A. Kolomieńsky, V.A. Petukhov, and M.S. Rabinovich, and T. Ohkawa of Japan.

The circular magnet is composed of separate sectors with a magnetic field that increases rapidly along the radius. This field has a different direction in neighbouring sectors. In sectors having a positive direction of magnetic-field, the radius of curvature of the trajectory of the particles is positive, and the magnetic field, which increases with the radius, creates focussing along the radius and defocussing along the vertical. In sectors with the opposite direction of magnetic-field, the trajectory curvature is negative, and the particles move, relative to the centre of the circle of their orbit, in a magnetic field that rapidly decreases with decreasing radius, thus changing the direction of focussing and defocussing. In this manner, alternating focussing and defocussing forces along the orbit are developed, resulting in increased focussing of the particles in both directions, such as in synchrotrons with strong focussing. In comparison with accelerators that do not have sectored magnetic fields with reversed signs, the dimensions of the fixed-field alternating-gradient devices are larger by a factor of about 5. This is the main disadvantage of these devices.

Injection of the particles takes place in an area with a weak magnetic field, in the interior part of the circular magnet. The accelerating high-frequency voltage (phasotron regimen of acceleration) is generated by resonators that are placed in the spaces between the magnetic sectors. The orbit expands toward the outer part of the magnet in the area of the stronger magnetic field, and the increase is proportional to the increase in energy. Both protons and electrons can be accelerated to relativistic energies in these accelerators.

Storage rings and colliding beams. Upon the collision of accelerated particles with the atomic nuclei of targets that are at rest in the laboratory coordinate system, not all of the energy of the particles goes into the reaction of the collision, including formation of new particles and excitation of the nuclei. Part of the energy is expended in altering the motion of the centre of mass of the bombarding particle and the one at rest. Such a pattern develops, for example, during collision of two absolutely inelastic balls with identical masses, one of which is moving while the other is at rest. After the collision the balls are joined and continue moving together with a new velocity. The momentum of the bombarding particle, according to the law of conservation of momentum, must be equal to the total momentum after the collision. From this can be derived the concept that half of the energy goes toward deformation of the balls and is dissipated in the form of heat, and it is just that energy which goes into the reaction upon collision of the accelerated and motionless particles. In the relativistic case during the collision of particles with identical masses, the energy available for the desired reaction is greatly reduced. During the collision of a proton that has been accelerated to 100 GeV with a proton at rest, for example, only 12 GeV can be utilized. The situation is radically different when the reaction of colliding particles that are bombarding each other is considered. During the collision of identical particles with identical energies travelling in opposite directions, the useful energy is equal to twice that of the energy of either particle, since in this case the centre of inertia of the two particles is at rest.

Energy lost in momentum changes

The equivalent energy that a particle must have upon collision with a motionless particle, in order for the useful energy to be the same as in a collision between moving particles, can be very large. For colliding protons with energies of 100 GeV, for example, the equivalent energy is 22,800 GeV. Thus, colliding beams provide a tremendous increase in useful energy and raise the possibility of conducting completely new physical research. The storage rings, however, cannot be an adequate replacement for ordinary accelerators, since most experiments must be performed with stationary targets.

The method of colliding beams was developed rapidly after the American physicist Donald W. Kerst showed, in 1956, that a sufficiently dense beam can be produced in strong-focussing accelerators. In a short time, a series of efficient devices with electron–electron, electron–positron, and proton–proton colliding beams were constructed; *e.g.*, at Stanford, California, and Cambridge, Massachusetts; Orsay, France; Novosibirsk, Russian S.F.S.R.; and Frascati, Italy. The largest device with a proton–proton colliding beam having 28 GeV energies was built at Geneva.

Production of colliding beams

Colliding beams are produced either in special storage rings or directly in the accelerator. One of the possible schemes of an electron–positron accumulating ring is shown in Figure 9. The circuit is essentially a synchrotron

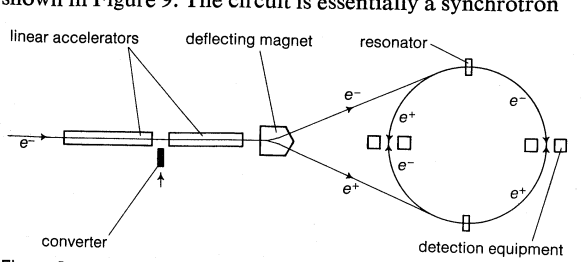


Figure 9: Accumulator with electron-positron colliding beams (see text).

with a fixed magnetic field. Electrons, e^-, and positrons, e^+, are sequentially injected. The injection of electrons is performed by the first linear accelerator (the second is disengaged during this time). The electrons are directed into the ring by deflecting magnets and are injected at a synchronized moment. After the completion of accumulation of the electrons, a converter (a tungsten plate, for instance) is positioned between the linear accelerators; in this converter electrons from the first linear accelerator produce positrons. A substantial portion of the positrons have an energy that is significantly less than the electrons, and these positrons are injected into the second linear accelerator. At the aforementioned synchronized moment the positrons are directed by deflecting magnets for injection into the ring in the direction opposite to that of electrons. Since the masses of the electron and positron are identical while their electric charge is opposite, their orbit in the magnetic field in the ring coincides. Revolving in opposite directions, the beams collide with each other.

Collision of the beams in rings proceeds with a frequency double that of revolution along the orbit (tens of millions of times a second). In ordinary accelerators particles pass the target under study only once. Therefore, in spite of the fact that the density of the motionless target (solid, liquid, or gas) is tens of millions times greater than the density of the particles in the colliding beams, the rate of the same reactions will be practically the same. In order to retain the particles in an orbit with constant radius and to compensate for the loss of energy due to synchrotron radiation, high-frequency resonators are placed along the orbit. Their action is analogous to that of the accelerating device in synchrotrons. The difference lies in the fact that in storage rings the energy of the high-frequency field of the resonators is used up only in radiation.

In storage rings conditions are created for the existence of beams of particles for periods of tens of hours. This is achieved primarily through the use of an ultrahigh vacuum in the storage ring chamber (10^{-9}–10^{-11} millimetre of mercury). This checks the loss of particles due to collision with atoms of residual gas in the chamber.

An apparatus that records the results of the reaction is located near the point of collision of the beams of particles.

FUTURE PROSPECTS

Future trends in accelerator design and development will be in the direction of obtaining still more intense beams of particles with still higher energies. Progress in this direction depends, on the one hand, on further improvements in traditional methods of acceleration and, on the other hand, on the discovery of new principles of acceleration.

Increasing the energy of accelerators that use the forward process of acceleration results in units of larger and larger sizes. In 1970 work was begun at Batavia, Illinois, on a proton synchrotron with an ultimate planned energy of 500 GeV. The radius of the orbit is one kilometre (0.6 mile). Injection of particles into the accelerator will take place at an energy of eight GeV. The particles will be accelerated first in a linear accelerator and then in an intermediate proton synchrotron. It is expected that the intensity of the accelerated beam will be 1.5×10^{13} protons per second or more. By late 1972, construction had advanced to the point that on December 14 a beam with an energy of 400 GeV was achieved.

In Geneva construction was under way in the early 1970s on a proton synchrotron with a planned energy of 300 GeV and an orbital radius of 1.1 kilometres. Injection of the particles will be by the now operating proton synchrotron with an energy of 28 GeV. The intensity of the beam is projected to be 10^{12}–10^{13} protons a second. Further increases in the energies of cyclic accelerators will probably require the creation of structures with radii of several kilometres. Hence, further progress in this area is impractical.

The only way to decrease the dimensions of cyclic accelerators is through abandoning the use of iron-core electromagnets and changing to a nonferrous system. The magnetic fields required for a given accelerator can be created by the use of electrical conductors wound in the desired configurations with a heavy current flowing through the conductors. A long-term goal involves the use of superconductors for the magnet coils. The remarkable feature of superconductors is that there is virtually no loss of energy in the circuit due to the passage of electrical current. This permits making the superconductors very thin and creating "fibres" of current with huge densities. With the help of superconductors, magnetic fields can be obtained that are several times as strong as those possible with ferromagnetic materials. Increasing the magnetic-field strength in cyclic accelerators results in proportionately decreasing the radius or, conversely, increasing the available particle energy.

Use of superconductors to improve performance

Superconductivity also finds application in linear resonance accelerators. If resonators are made of superconducting materials, practically none of the energy of the high-frequency field is dissipated in heating up the walls of the resonator, all of it going instead into acceleration of the particles. This facilitates the construction of strong-current linear accelerators with continuous operation.

The strength of the electric field in superconducting resonators can be made much higher than in ordinary accelerators. It is possible, for example, to consider field strengths as high as 500,000 volts per centimetre. The first experimental superconducting linear accelerator was built at Stanford in 1967. The ultimate gain in energy of the particles was three MeV per metre.

An important direction in the development of accelerators is research on increasing the intensity of accelerated beams of particles. The present threshold is 10^{13} particles per second. Further intensity increases in operating and in projected accelerators will substantially improve the physical-research potential and the application of accelerators in different areas of technology.

Included in a number of new ideas is one presented in 1956 involving the method of collective acceleration proposed by V.I. Veksler. In a special apparatus a stable dense cloud of electrons is formed, into which positively charged ions (protons, for example) are introduced. Acceleration of the electrons takes place in a magnetic field. Under certain conditions, the ions, trapped by electrical forces inside the cloud, are accelerated along with the electrons. Since the mass of an ion is significantly larger than the mass of an electron, at identical velocities the ions will have many times more energy than the electrons.

Collective acceleration

The realization of the method of collective acceleration will permit the acceleration of beams of heavy particles to large energies in very small distances. It will be possible, for example, to accelerate a proton to an energy of one GeV in an accelerator with a length of one metre. Conventional linear accelerators of such an energy require lengths of about 1,000 metres (3,300 feet). Experimental research on this new method is being conducted in a series of centres in the United States and the Soviet Union.

The search for means of creating accelerators of ultrahigh energies is caused by requirements of contemporary high-energy physics. Surmounting the next barrier in energies is always accompanied by new discoveries. It is expected that by increasing the energy of accelerated beams of particles and by studying processes that occur in very small distances (*e.g.*, 10^{-17} centimetre), new microworld phenomena will be discovered, such as fundamental particles, of which the famous "elementary" particles are formed, and fundamental length, which determines the applicability of the existing theory of fields and the theory of elementary particles.

BIBLIOGRAPHY. J.J. LIVINGOOD, *Principles of Cyclic Particle Accelerators* (1961); A.A. KOLOMENSKY and A.N. LEBEDEV, *Theory of Cyclic Accelerators* (1966; orig. pub. in Russian, 1962); M.H. BLEWETT, "Characteristics of Typical Accelerators," *A. Rev. Nucl. Sci.*, 17:427–468 (1967); J.D. COCKCROFT and E.T.S. WALTON, "Experiments with High Velocity Positive Ions," *Proc. R. Soc.*, Series A, 137:229–242 (1932), on the cascade generator; R.J. VAN DE GRAAFF, "A 1,500,000 Volt Electrostatic Generator," *Phys. Rev.*, 38:1919–1920 (1931); "Tandem Electrostatic Accelerators," *Nucl. Instrum.*

Meth., 8:195–202 (1960); D.W. KERST, "Development of the Betatron and Applications of High Energy Betatron Radiations," *Am. Scient.*, 35:56–84 (1947); E.O. LAWRENCE and N.E. EDELFSEN, "On the Production of High Speed Protons," *Science*, 72:376–377 (1930), on the classical cyclotron; L.H. THOMAS, "The Path of Ions in the Cyclotron," *Phys. Rev.*, 54: 580–598 (1938), on the cyclotron with classical focus; J.A. MARTIN, "The Separated-Orbit Cyclotron," *IEEE Trans. Nucl. Sci.*, New Series, 13:288–299 (1966); R. WIDEROE, "Über ein neues Prinzip zur Herstellung hoher Spannungen," *Arch. Elektrotech.*, 21:387 (1928), on the linear accelerator; V.I. VEKSLER, "A New Method for Acceleration of Relativistic Particles," *Dokl. Akad. Nauk SSSR*, 43:329 (1944), on the synchrotron; E.M. MCMILLAN, "The Synchrotron: A Proposed High Energy Particle Accelerator," *Phys. Rev.*, 68:143–144 (1945); D.W. KERST *et al.*, "Attainment of Very High Energy by Means of Intersecting Beams of Particles," *Phys. Rev.*, 102:590–591 (1956), on storage rings.

(Y.M.A.)

Acclimatization

Every organism is capable of adjusting to changes that occur in its environment. Some changes are gradual and elicit long-term responses—or the capability of responses—that are collectively grouped under the term acclimatization. Such responses are more or less habitual, and reversible, should environmental conditions revert to an earlier state. The numerous sudden changes that evoke rapid and short-term responses via the nervous and hormonal systems are not examples of acclimatization.

An individual organism can regulate its internal processes rapidly to sustain itself within the usual range of environmental changes that it encounters hourly or daily. But this rapid regulation, or homeostasis, is limited in its operation to a small range of environmental variations. Homeostatic regulation usually cannot meet effectively large environmental changes such as would allow a plant or animal living in the warmth of summer to operate in the cold of winter. As summer wanes, organisms change their substance and their habits in seeming anticipation of winter. This gradual adjustment to conditions is acclimatization.

GENERAL FEATURES

Range of acclimatization

Although acclimatization may appear to refer basically to adaptation to a climate, it can also be used to indicate the adjustments of a man to urban, social, or political conditions, or to the adaptation of a population of plants to conditions of cultivation, or of an animal to the unnatural conditions of captivity. (But adaptations to strange or artificial conditions are often difficult to describe in a useful manner, and only in a few examples can such adaptations be compared with acclimatization.) The following account is restricted to examples of acclimatization in which the environmental conditions that evoke the change can be measured and in which the intrinsic changes in the organism can be described.

Because organisms do not have unlimited combinations of adaptations, they may use a similar adaptive process with changes of different origins. In acclimatization to the low pressure of oxygen (hypoxia) in high mountains, animals, including man, improve the capacity of blood for the transport of oxygen by increasing the number of red blood cells (polycythemia); in the chronic disease emphysema, the inadequate supply of oxygen to the lungs is to some degree compensated by a similar polycythemia.

Acclimatization does not necessarily mean that a plant or animal is adapted to operate at its maximal rate. In the hot summer desert acclimatized birds and mammals often rest in the shade, and in winter cold some animals and all plants become dormant. Acclimatization often requires modification of activity, but the adaptive changes allow the individual to live normally and to utilize a greater variety of habitats than would otherwise be available. At extreme limits an organism may suffer some impairment of vigour, but it survives nevertheless; if the impairment is overt, acclimatization is considered inadequate.

In contrast to changes that occur during growth and development, acclimatization refers to an adaptive change that is reversible when conditions return to their former condition. Acclimatization of an individual does not leave a lasting impression upon its genetic mechanisms. The adaptation of populations to change that effects evolution by the selection of genetic capability is a different process from the acclimatization of an individual (see also EVOLUTION); nevertheless, acclimatization occurs commonly and is necessary in changing environments. Some of its manifestations, essential for the continuance of life, are especially interesting because they reveal the great flexibility of vital reactions.

INFLUENCES OF ENVIRONMENT

Environmental conditions change with the seasons, as from winter to summer or from periods of humid to dry weather in the tropics. Sedentary animals and plants adjust readily to the local seasons. But winds, rivers, currents, locomotion, and man have been operating for ages to disperse individual organisms into new environments. Those individuals that acclimatize to their newfound circumstances thrive and reproduce; those that fail to acclimatize perish.

In dealing with acclimatization the influence of climate upon life can be usefully divided under headings of adjustments for temperature, humidity, salinity, light, pressure, and certain chemical substances in the environment.

The hidden potential of organisms

Because animals and plants can be successfully introduced to new regions, it can be said that species do not necessarily thrive at their best potential in their native regions. Dandelions and Scotch broom, native to Europe, have spread successfully over large areas in North America, as have starlings and English sparrows. Muskrats from North America have extended widely through Europe and Asia. Monterey pine, a small tree native to a few thousand acres of California, grows into a large tree in New Zealand. Not every introduction succeeds, of course. Of those that do succeed, only a few are highly useful to man; many more become costly nuisances.

Adjustments to temperature. Considering that sunshine, moisture, and wind, among other conditions, greatly affect weather, it is remarkable that temperature alone can be a principal indicator of the suitability of climates for life. Rates of reactions in cells and tissues are critically affected by temperature. In addition to their obvious influence upon life, the effects of temperature are favoured for study by scientists because of the clear-cut relationship between reactions of organisms and changes in temperature.

Plants. As the weather warms in spring, plants that have been cultivated in the shelter of a greenhouse can be made to tolerate the daily changes in outdoor temperature by being gradually exposed to such conditions. This process, called hardening, readies them for planting in the field or garden. Pansies and cabbages are rather readily acclimatized to the variable weather of a temperate climate. More sensitive plants such as squashes and begonias need a longer time. Plants originating in the tropics commonly do not become acclimatized readily to the weather pattern in temperate regions.

Early in the history of agriculture, useful plants were transported to climates differing from their native one. In order to become established, the first transplants had to survive the transition and produce seeds. Thereafter natural or artificial selection could operate to favour variants among the progeny best adapted to the new climate.

Plant response to seasonal changes

In nature, the distribution of plants is obviously related to climate, in which temperature is the most easily recognizable factor. Many plants of temperate regions are able to thrive within a considerable range of seasonal temperatures. In more northward areas, many species of trees fail to acclimatize to the extreme seasonal changes. A few species, however, such as spruce, larch, and willow, can survive the coldest winters of the interior Arctic. They become acclimatized to endure very cold weather (−50° C or −58° F), provided that the short summer is warm enough to allow sufficient photosynthesis for growth and for the production of seeds.

The fertilization of ova (eggs) by pollen in flowers seems to be a part of the life history in many plants that can be

acclimatized only to a very small range of warm temperatures. Flowers of tomatoes do not set fruit without sufficient warmth. Wild Arctic and Alpine plants, however, flower abundantly in regions with intermittent freezing and warmer periods in summer. Acclimatization for cold may start even before seasonal cooling occurs. In temperate regions, as the days of late summer shorten, the leaves of some deciduous plants diminish their growth; but, before they drop, they transmit an influence, probably through regulating substances (*e.g.*, hormones), that induces cold tolerance in woody stems. Thus an influence other than cold can initiate hardening for winter.

The changing length of daylight is a more reliable indicator of average progress of the seasons than is the day-by-day fluctuation in temperature. Sensitivity to the duration of daylight or to the duration of darkness indicates that somehow time and thus season can be measured by organisms. By reference to the environmental signal of day length, organisms can prepare for the temperatures of the oncoming season. Doubtless, factors other than light are also involved in transmitting signals for acclimatization to approaching seasonal change.

Insects. Many insects pass through several stages in their life cycles. As adults and larvae they usually are active; as pupae they usually remain dormant. Many sub-Arctic insects, however, are now known to become dormant in their adult, or reproductive, forms. A very small carabid beetle winters in stumps in northern Alaska exposed to temperatures as low as −40° C (−40° F). As the weather cools in autumn, the beetle accumulates glycerol in its body and thereby acquires tolerance to cold, very much as antifreeze provides cold tolerance in automobiles. As the days of spring become warmer, the amount of glycerol diminishes, and cold hardiness is gradually lost.

Glycerol as protection against freezing

Insects are the most diverse class of animals in numbers, forms, and natural ranges. The tolerance of cold differs in each species, as does tolerance of heat. Possible acclimatization to heat differs with species and with stage of growth. The beetle example above, although not representative, illustrates the processes involved in acclimatization in some insects.

It is thought that glycerol prevents cell damage from freezing. Glycerol preparations of spermatozoa from bulls, of human red blood cells, and of tissues of the cornea from the eye of rabbits can be preserved for long periods in liquid carbon dioxide (−80° C or −112° F) and successfully restored for use. Several other substances with properties resembling glycerol also serve as cryoprotectants (*i.e.*, protectants against severe cold).

Fish. Fish of temperate and Arctic waters gradually adjust their activity and metabolism to yield peak effectiveness in the temperatures that prevail; some fishes of perpetually cold Antarctic seawater, however, can perform most effectively only at temperatures from −2° to +4° C (28° to 39° F). In temperate and Arctic zones, large bodies of water, near freezing in winter, may warm 10° C (18° F) in summer, whereas small bodies of water may warm more than 20° C (36° F). In the cold season fish tolerate colder temperatures than they do in summer. Conversely, in the warm season they tolerate warmer water than they do in winter. Not only does ability to survive a given warm or cold temperature change with experimental acclimatization (often called acclimation by physiologists) but maximal activity, measured by speed or metabolic use of oxygen, also changes.

The results of experiments on acclimatization of certain fishes to water temperatures generally agree with the natural temperatures of the seasonal habitats of the various species. There are, of course, hazards in referring to temperature as a single environmental factor for acclimatization—*i.e.*, temperature affects the amount of water that must be pumped over the gills for respiratory exchange, and, in the course of the year, growth and reproductive condition also exert their influences.

Birds. Birds maintain their bodily warmth in all climates and seasons. In the Arctic winter, small common redpolls keep warm in temperatures as low as −40° C (−40° F) by increasing their metabolic rate three or four times the normal rate at average winter temperatures. In summer, however, they cannot produce heat adequate for life in air at −40° C, even with an abundant supply of food. Evidently their metabolic processes become acclimatized for winter and summer conditions.

A striking illustration of acclimatization has been provided by an Arctic glaucous gull that was kept in a warm laboratory at Barrow, Alaska, during a midwinter month. It escaped outdoors in a temperature of −20° C (−4° F); although it was found a very short time later, its feet were frozen. Several other gulls kept in open outdoor cages and thus exposed to temperatures as low as −30° C (−22° F) survived without any freezing injury.

Mammals. The thickening of the fur of northern dogs, horses, and wild mammals in anticipation of winter is as familiar and useful an acclimatization as is the shedding of fur in summer. In nature, change in various properties of fur accompanies the change in seasonal temperature. Several species that live in snowy climates, such as weasels, varying hares, and Arctic foxes, assume white coats in winter. This transition to white can be unseasonably accelerated in weasels by experimentally shortening the duration of daylight. In some mammals, as in some plants, detection by the organism of the seasonal change of illumination seems to serve as a signal to initiate seasonal change. Several other species of mammals lengthen their fur on exposure to cold even when there is no noticeable change in illumination, as in laboratory experiments.

Northern varying hares change in spring from white to brown fur; if the snow cover melts early, the white hares may be conspicuous against the dark-coloured ground. Thus, when seasonal conditions are out of phase with the average state, the variation in light may not correctly signal a change of coat colour. As with other gradual processes of seasonal acclimatization, the change may not be in good relation to the momentary condition of the environment.

The economy of adaptation

Many adaptive processes seem to operate by reducing rather than by eliminating erroneous adaptation. When Australian dingoes were experimentally exposed to cold of about −20° C, their fur thickened noticeably; in addition, their measured metabolism in cold was less than for dingoes that had been kept warm. It was concluded that the ancestors of these dogs, introduced to Australia several thousand years ago, had remained in that warm climate long enough to become recognized as a distinct species, but they also retained the capability to acclimatize to temperatures far colder than their species had experienced in many centuries.

Seals acclimatize to ice water primarily by insulating their body with layers of fat instead of with fur. There is a gradual rise in temperatures through five or six centimetres of blubber before the internal body temperature of 37° C (99° F) is reached.

As temperatures rise above about 30° C (86° F), and depending upon the humidity, birds and mammals gradually lose their ability to dissipate sufficient heat to regulate body temperature and resort instead to evaporative cooling or sweating. Man is one of the most effective users of this mechanism, the effectiveness of which improves after a few days' exposure to hot weather. Acclimatization involves activation of more sweat glands, compensatory conservation of water excreted through kidneys, and some reduction in the outflow of salt through kidneys and sweat glands. The fullest acclimatization for heat probably involves alteration of the signals that evoke sweating and circulation of blood in the skin.

Adjustments to illumination. Plants react markedly to changing illumination. The most conspicuous reactions, which involve growth and development of new structures, are not properly acclimatizations. But the changing schedule of day length provides the signal that induces development of cold tolerance. Changing day length appears to be a signal that permits the acclimatization mechanisms to anticipate adaptation for extremes of temperatures that depart from average conditions.

Light as an environmental signal

Migratory birds prepare early in spring for the flight northward to their breeding grounds by accumulating fat

in amounts that may nearly double their body weight. Preparations of the reproductive organs are, likewise, initiated early and are complete soon after arrival on the nesting grounds. In the western American white-crowned sparrow, the signal that sets in motion this behaviour is the gradually increasing length of day during spring. The environmental signal acts through complex neural and hormonal paths centred in the bird's brain.

Later in summer, as the days shorten, the birds fatten again, but this time in preparation for the southward migration. In this highly adaptive behaviour of birds, day length provides a signal for activity appropriate to the usual seasonal habits both of species and of local populations. The whole system is so elaborate and is apparently such an effective guide in avian behaviour that it is not usually dealt with under the designation of acclimatization; nevertheless, it represents the best known of the complicated systems of reactions that keep each individual in a population in accord with the changing seasonal conditions of the environments upon which it so precisely depends.

Adaptations to meet changes of the seasons encounter the vagaries of weather, which at any date may differ widely from average conditions. In preparing to meet the coldness of November in the north, there is an almost even chance that the weather will be colder than average and a smaller chance that it may be much colder. Acclimatization for cold, therefore, must exceed the average of local conditions. In another sense, acclimatization must anticipate both the progress and degree of seasonal cold—that is, a plant or animal must change its outlook upon the oncoming seasons by referring to the seasons in its past experience. This would require reference to annual cycles of the past and expectations on which to predict the future. It is certain that seasonal change in length of day gives a clue as to the date, and hence time, for initiating the processes of acclimatization. Degrees of acclimatization for the future appear to require margins for error or departure from average environmental conditions.

How the length of day influences man's life is shown impressively during rapid travel by air westward, when the duration of daylight becomes prolonged. Customary times for meals and sleep become deranged, and for a few days the newly arrived traveller's internal condition is out of phase with the strange new program of environmental events. Even more disturbing to the rhythmical changes that occur according to time of day are the shortening of length of daylight and the shortening of corresponding schedules that result from travel eastward. After travelling from San Francisco to London, a person may require several days to become adapted to his normal efficiency.

Means of conserving water

Adjustments to humidity. Rain that falls on the desert promotes sudden germination of seeds and rapid new growth and flowering of plants that lay dormant through the dry period. Many of these transitions are concerned with growth of new structures and are not generally regarded as examples of acclimatization. Perennial stems and roots resume transport of water and dissolved material by reactivating functions that were inactive during the dry season. Perennial stems of cactus and some other desert plants are said to store water in humid seasons. During periods of high humidity, water available in the soil rises rapidly as a result of transpiration of water from leaves. Substances elaborated in leaves by photosynthesis are in turn distributed to all growing parts. It can be considered that activation of these transport systems during a humid period involves acclimatization of the systems of transport for special activity.

In winter, in northern regions, evaporation from twigs, buds, and needles of conifers can still occur. If the water in soil and transport systems is immobilized by freezing, the lost water is not replaced. If, furthermore, the stems, twigs, and needles are warmed by the late winter sun, the resulting drying out of buds and needles would be lethal; but the winter structures retain their ability to adjust to varying levels of water vapour. Buds of Arctic trees retain moisture well during winter; water resumes its movement into buds as adequate stores in large stems melt.

The lungfish (*Protopterus*) of the swampy overflows of the Nile River burrows in the mud and encases itself in an impervious covering as dry weather approaches. In this covering the fish can resist drying until the next flooding of the rivers. In a similar manner, numerous invertebrates evade drying out by special impervious structures in which they remain dormant during dry seasons.

Insects that breathe through respiratory systems consisting of air tubes guard against drying by changing the openings to their air tubes in such a way as to reduce loss of water vapour while still permitting the inward diffusion of oxygen and the outward escape of carbon dioxide waste.

Warm-blooded birds and mammals in climates with prolonged warm, dry periods must conserve water by reducing its use in sweating. Some kinds of cattle and sheep in hot desert lands are especially adapted for the economical use of water. The camel is one well-known example. On a hot, dry summer day, it allows its body temperature to rise several degrees above normal. During the night, when radiation through the clear atmosphere to the cold dark sky cools the air, the camel loses the heat stored by day and cools several degrees below its usual temperature. By utilizing its massive body in this manner, it loses less water by evaporative cooling. In the cooler season the camel does not show daily alterations of body temperature. A number of African and Australian herbivores regulate body temperature to conserve water in dry seasons. Man cannot manipulate his body temperature for significant economy of water but is well adapted for hot climates by his ability to sweat copiously.

Adjustments to chemicals. *Salts.* Closely related to acclimatization for humidity are the adjustments that must be made for salts dissolved in a watery environment. Mammals and birds regulate their intake of salts and water to a considerable extent by choice of food and drink. In addition, they excrete the excess salt and water mainly through kidneys and thus maintain the concentration of salts in their body fluids to an equivalent of about a 0.9 percent solution of sodium chloride. Because seasons affect the balance of nutritional intake of salts and water and the use of water for evaporative cooling, intricate adaptive regulatory processes are necessary. Organisms differ so widely that there is no single good example of a common process of this type of acclimatization.

Estuarine organisms

Animals of estuaries, where freshwater mixes with seawater, are exposed to changing concentrations of salt water as currents change seasonally or as the animals themselves move about. Trout and salmon swim from the sea, which contains dissolved salts equivalent to about 3.5 percent sodium chloride, into practically saltless freshwater. With the first encounter with diluted seawater, their blood becomes slightly diluted, but as they gradually reach freshwater, excess water is actively excreted to maintain their internal salt level at about 0.8 percent. Among the species of fishes that can become acclimatized to such changes are salmon, trout, sturgeon, and whitefish.

The ability to adjust to changes in salt concentration in the environment is found in many invertebrates in degrees varying from those that withstand very little change to a few, such as the brine shrimp (*Artemia*), that can endure saturated sodium chloride solutions or freshwater.

Other chemical substances. Mammals require various specific minimal quantities of proteins, fats, and carbohydrates, along with vitamins and inorganic elements (*e.g.*, calcium, iodine). There appears to be no adaptation that permits a species to thrive on nutritional deficiencies or excesses and hence no important acclimatization for variation in the essential composition of diet. Outright deficiencies or excesses cause overt defects. There is, however, vagueness regarding the upper and lower limits of nutrition for a species or for a population living in special circumstances. What was once generally regarded as acclimatization to diet is now known in part to be brought about by development of immunity to local infectious illnesses transmitted perhaps in foods and water.

Herbivorous animals such as domesticated cattle and wild deer accomplish preliminary digestion in the rumen by the action of micro-organisms that break down plants to substances that can be digested by the animal. In addition, the micro-organisms growing in the rumen furnish nitrogenous nutrients and vitamins. In spring the fresh green feed differs in digestibility from the dry feed of winter, and a dietary adjustment must be made in the digestive processes of the animal; part of this is accomplished by a gradual change in the microbial population of the animal's rumen. In a general way this adjustment represents acclimatization for diet.

Tolerance to pollutants

A long list of other substances, derived mainly from industrial processes and uses, now appears to be accumulated by plants and animals and transferred through chains of food consumption to man in dangerous amounts. Mercury and arsenic, for example, have been transferred through aquatic food chains and have caused illness and even fatalities in man. The fishes, however, that accumulate and transfer these poisons must presumably tolerate them long enough to reproduce stable populations. Their tolerance can be considered a form of acclimatization.

Adjustments to changes in pressure. The history of the conquests of the Inca Indians in the high Andes of South America included accounts of the suffering of the Spanish invaders and their horses from the altitude and its diminished atmospheric pressure. The residents native to the high mountains appeared to have no handicap. On the other hand, the Incas were acclimatized to the high altitudes and could not safely move to lower ones. For many years no births occurred among the Spaniards newly arrived in the high Andes, although the native people were prolific. Some European domesticated animals also were at first unsuccessful at breeding in the high Andes. The native Andean races of people and animals resident for many generations at high altitudes probably developed selectively inherited adaptive capabilities. As modern people become more mobile, and as scientific breeding makes possible transfers of stocks of domesticated animals, the heritable factors that allow adaptability to altitude in populations will be able to be contrasted with the acclimatization of individuals.

Aviators who rise rapidly to altitudes for which they cannot adequately compensate also suffer the discomfort of high altitudes. Some acclimatization is observed after repeated high flights, but provision of oxygen and pressurized cabins have now extended the limited ability of man. In very low atmospheric pressures, the entire body must be enclosed in a pressurized suit in order to maintain pressures that permit normal circulation of blood.

Because the pressure of oxygen in mountain air is reduced, oxygenation of the blood of a new arrival from the seacoast is diminished. In an attempt to compensate, the subject breathes rapidly, which, although somewhat increasing the oxygen content of blood, abnormally lowers the pressure of carbon dioxide and increases blood alkalinity, resulting in a condition called hypoxic hypoxia. Some persons suffer headaches and many symptoms of mountain sickness at high altitudes, the severity and duration of the discomfort varying among individuals. After a period of time (a week or months), newcomers lose the acute symptoms of mountain sickness and begin to regain their ability for activity.

Some of the early compensatory physiological adjustments to altitude illustrate the mechanisms involved in acclimatization. The number of red blood cells increases, thereby improving the capacity of the blood for transporting oxygen. Blood circulation increases, improving the blood supply to the tissues, and apparently even the intracellular utilization of oxygen is increased.

Limits of habituation to sound waves

Rapid pulsations of air pressure are detected by the ear as sound. Adaptations for exposure to sound can diminish the sensitivity of the ear; in fact, many kinds of sensory receptors show diminishing discharge as a stimulus continues to operate. This adaptation of sense organs is usually too sudden to be regarded as acclimatization. But the persistent disturbance of a noisy airport, factory, or building site may become less disturbing in time, and some tolerance develops that is perhaps a type of acclimatization. Such habituation to noise is seen among many animals, including numerous birds that nest and successfully raise young in the presence of almost unbearable noise. As the world becomes more noisy, the limits to which acclimatization can extend short of allowing physical damage and mental disturbance are of utmost importance to the well-being of man.

THE BIOLOGICAL SIGNIFICANCE OF ACCLIMATIZATION

Although increase in scientific knowledge has resulted in precise terms for many adjustments formerly considered under acclimatization, many other adjustments to environmental changes still remain obscure and are covered under that designation.

Acclimatization can gradually alter homeostatic capabilities for average circumstances that are altered by seasons or by movement to a new locality. If acclimatization cannot adjust to allow for active function in the altered circumstances, dormancy may be induced until suitable conditions recur.

Acclimatization thus permits an organism to exploit regions of great seasonal variation and, on occasion, to move in wholly new environments. Only the new settlers that acclimatize can survive to produce progeny from which a new population may become established. Ability to become acclimatized differs greatly among species of plants and animals. Some breeds of domesticated animals and cultivated plants are quite versatile in this ability, whereas others are narrowly restricted.

An interesting characteristic of seasonal acclimatization appears in animals and plants that become adjusted to cold beyond that which they are likely to encounter. Not only does acclimatization prepare them with a margin of safety but some micro-organisms, insects, and plants tolerate experimental exposure at temperatures far colder or warmer than ever occur in nature. It seems strange that adaptability enables these organisms to be prepared to encounter conditions beyond their natural experience.

The anticipation of change

Another surprising characteristic of acclimatization is its anticipatory nature—it can develop before the change occurs. It would seem that anticipation of the need for change would be required in order to make the slow physiological preparations for climatic changes that often set in very suddenly. Anticipation of acclimatization seems to require a sense of time by which the coming environmental conditions can be predicted. Length of day is one external signal, but it seems to impinge upon intrinsic rhythms that provide clues from within as to the passage of time (see also PERIODICITY, BIOLOGICAL). Consideration of acclimatization reveals many processes by which animals and plants prepare to function or to exist in dormant survival in the various and changing circumstances of the world.

BIBLIOGRAPHY. D.B. DILL, *Adaptation to the Environment*, sect. 4 of the *Handbook of Physiology* (1964), an authoritatively technical series of papers in the most comprehensive survey of adaptation to the environment available; "Physiological Adjustments to Altitude Changes," J.A.M.A., 205: 747–753 (1968), a consideration of the physiological changes leading to acclimatization of newcomers to high altitudes and the limits to which people become truly acclimatized; H.W. WOOLHOUSE (ed.), *Dormancy and Survival* (1969), a publication of the 23rd Symposium of the Society for Experimental Biology (British), with 22 papers by experts dealing with adaptations to severe climatic and seasonal conditions; A.U. SMITH, *Biological Effects of Freezing and Supercooling* (1961), original views of the natural and experimental tolerance to freezing and its employment for the preservation of cells and tissues; L. IRVING, "Adaptations to Cold," *Scient. Am.*, 214:94–101 (1966), a general comparison of the employment of insulation for the economy of heat in birds and mammals; K. SCHMIDT-NIELSEN, "The Physiology of the Camel," *Scient. Am.*, 201:140–151 (1959), an interesting account of the daily changes in body temperature that spare evaporative cooling in the hot desert; C. MONGE, *Influencia biológica del altiplano en el individuo, la raza, las sociedades y la historia de América* (1940; Eng. trans., *Acclimatization in the Andes*, 1948), a vivid historical account of the acclimatization of native high Andean peoples and the development of acclimatization in newcomers.

(L.I.)

Accounting, Principles of

The purpose of accounting is to provide information about the economic affairs of an organization. It is a means of recording facts or forecasts about the economic aspects of an organization's activities and translating them into a useful form for further action. The information thus dealt with may be used in a number of ways: (1) by the organization's managers to help them plan and control the organization's operations; (2) by owners and legislative or regulatory bodies to help them appraise the organization's performance and make decisions as to its future; (3) by owners, lenders, suppliers, employees, and others to help them decide how much time or money to devote to the organization.

Accounting provides information for all these purposes through the maintainance of files of data and the preparation of various kinds of reports. Most accounting information is historical—that is, the accountant observes the things that the organization does, records their effects, and prepares reports summarizing what he has recorded. In addition, however, his reports often summarize the anticipated results of the things the organization plans to do. Accounting information can be developed for any kind of organization, not just for privately-owned, profit-seeking businesses. One branch of accounting deals with the economic operations of entire nations. The remainder of this article, however, will be devoted primarily to business accounting.

COMPANY FINANCIAL STATEMENTS

Some accounting reports are issued only to the company's management (see below, *Managerial Accounting*); others are also sent to owners and others outside the management group. The reports most likely to go outside the company are called the company's financial statements, and their preparation is the province of the branch of accounting known as financial accounting. Four kinds of financial statements will be discussed: the balance sheet, the income statement, the statement of changes in retained earnings, and the statement of sources and uses of funds.

The balance sheet. A balance sheet describes the resources that are under the organization's control on a specified date and indicates where these resources have come from. It consists of three major sections: (1) the assets: valuable rights owned by the firm; (2) the liabilities: the funds that have been provided by outside lenders and other creditors in exchange for the company's promise to make payments or provide services in the future; (3) the owners' equity: the funds that have been provided by or on behalf of the company's owners.

The list of assets shows the forms in which the organization's resources are lodged; the lists of liabilities and the owners' equity indicate where these same resources have come from. The balance sheet, in other words, shows the organization's resources from two points of view, and the following relationship must always exist: total assets equals total liabilities plus total owners' equity.

This same identity is also expressed in another way: total assets minus total liabilities equals total owners' equity. In this form, the equation emphasizes that the owners' equity in the organization is always equal to the net assets (assets minus liabilities). Any increase in one will inevitably be accompanied by an increase in the other, and the only way to increase the owner's equity is to increase the net assets.

Assets and liabilities

Assets are ordinarily subdivided into current assets and noncurrent assets. The former include cash, amounts receivable from customers, inventories, and other assets that are expected to be consumed or can be readily converted into cash during the next operating cycle (production, sale, and collection). Noncurrent assets may include noncurrent receivables, fixed assets (such as the value of land and buildings), and long-term investments, usually shares of stock and bonds of other companies.

The liabilities are similarly divided into current liabilities and noncurrent liabilities. Most amounts payable to the firm's suppliers (accounts payable), to employees (wages payable), or to governments (taxes payable) will be included among the current liabilities. Noncurrent liabilities consist mainly of amounts payable to holders of the company's long-term bonds and such items as the firm's obligations to employees under company pension plans.

The difference between the total of the current assets and the total of the current liabilities is known as net current assets, or working capital.

The owners' equity of a United States company is divided between paid-in capital and retained earnings. Paid-in capital represents the amounts paid to the corporation in exchange for shares of the company's preferred and common stock. The major part of this, the capital paid in by the common shareholders, is usually divided into two parts, one representing the par value, or stated value, of the shares, the other representing the excess over this amount, often referred to as capital surplus. The amount of retained earnings is the difference between the amounts earned by the company in the past and the dividends that have been distributed to the owners. In other words, the amount of retained earnings shows how much of the company's assets have come from the reinvestment of the company's own earnings.

A slightly different breakdown of the owner's equity is used in most of continental Europe and in other parts of the world. The classification is a legal one, distinguishing between those amounts that cannot be distributed except as part of a formal liquidation of all or part of the company (capital and legal reserves) and those amounts that are not restricted in this way (free reserves and undistributed profits).

A simple balance sheet is shown in Table 1. Because the two sides of this balance sheet represent two different aspects of the same entity—the corporation's capital—the totals must always be identical. Thus a change in the amount for one item must always be accompanied by an equal change in some other item. For example, if the company pays $40 to one of its trade creditors, the cash balance will go down by $40, and the balance in accounts payable will go down by the same amount.

The income statement. The firm uses its assets to produce goods and services. Its success will depend on whether it is wise or lucky in the assets it chooses to hold and in the ways it uses these assets to produce goods and services.

Net income

The firm's success is measured by the amount of profit it earns—that is, the growth or decline in its stock of assets from all sources other than contributions or withdrawals of funds by owners and creditors. Net income is the accountant's term for the amount of profit that is reported for a particular time period.

The firm's income statement for a period of time shows how the net income figure for that period was derived. For example, the first line in the income statement in Table 2 shows the company's sales revenues for the period: the assets obtained from its customers in exchange for the goods and services that constitute the firm's ordinary stock in trade. The second line summarizes the company's revenues from other sources.

Next, the income statement shows the expenses of the period: the assets that were consumed while the revenues were being created. The expenses are usually broken down into several categories indicating what the assets were used for. In Table 2, six expense items are distinguished, starting with the cost of the merchandise that was sold during the period and continuing down through the provision for income taxes.

The bottom portion of the income statement is reserved for gains and losses from extraordinary events. In this case, the company sold some of its long-term investments for more than their original purchase price. Because this was not part of the company's normal business, the sale price, costs, and taxes on the sale were kept separate from the ordinary revenue and expense totals; the income statement shows only a single figure, the net gain on the sale.

Table 1: Any Company, Inc.: Balance Sheet as of Dec. 31, 19__

assets			liabilities and owners' equity		
Current assets			Current liabilities		
Cash		$100	Wages payable		$ 20
Marketable securities		50	Accounts payable		160
Accounts receivable		150	Total current liabilities		$180
Inventories		180	Long-term bonds payable		80
Total current assets		$480	Total liabilities		$260
Long-term investments		70	Owners' equity		
Plant and equipment			Common stock	$200	
Original cost	$300		Retained earnings	280	
Less: accumulated depreciation	110	190	Total owners' equity		480
Total assets		$740	Total liabilities and owners' equity		$740

Table 2: Any Company, Inc.: Income Statement for the Year Ending Dec. 31, 19__

Net sales revenues		$800
Interest and other revenues		14
Total revenues		$814
Expenses		
Cost of merchandise sold	$492	
Salaries of employees	116	
Depreciation	30	
Interest expense	4	
Other expenses	78	
Provision for taxes on ordinary income	47	767
Net earnings before extraordinary item		$ 47
Extraordinary item: gain on sale of investment (less applicable income taxes)		5
Net income		$ 52

The net income figure summarizes all of the gains and losses recognized during the period, including both the results of the firm's normal, day-by-day activities and any extraordinary events. If net income is negative, it is referred to as a net loss.

The statement of changes in retained earnings. Like the income statement, the statement of changes in retained earnings refers to a period of time rather than to a single date. It explains the change from one balance sheet date to the next in the amount shown as retained earnings.

The retained earnings statement is the link between the income statement and the balance sheet. This relationship can be demonstrated by rewriting the accounting equation in the following way: assets minus liabilities equals paid-in capital plus retained earnings. Whenever the firm reports net income, it means that more assets have come in as revenues than went out as expenses; more precisely, it means that the net asset total (assets minus liabilities) has increased. Therefore, if the accounting equation is to remain balanced, an amount equal to net income must be added to the retained earnings figure.

Similarly, the distribution of cash dividends reduces net assets, and this change again is usually balanced by a reduction in the retained earnings to keep the equation in trim. Thus the typical statement of changes in retained earnings will consist of the items shown in Table 3. Starting with retained earnings of $263 at the beginning of the year, the company added $52 in net income during the year (from Table 2). Dividends amounting to $35 were distributed to shareholders during the year, leaving a year-end balance of $280. This is the amount to be carried to the year-end balance sheet (Table 1).

Table 3: Any Company, Inc.: Statement of Changes in Retained Earnings for the Year Ending Dec. 31, 19__

Retained earnings, Jan. 1, 19__	$263
Net income earned during 19__	52
	$315
Less: dividends distributed to shareholders during 19__	35
Retained earnings, Dec. 31, 19__	$280

The statement of the sources and uses of funds. The statement of sources and uses of funds, or funds statement, identifies the major flows of funds during the period. In most corporate reports, "funds" means working capital. The idea is easier to understand, however, from an illustration in which the flow of funds means the flow of cash. If a salesman earns $18,000 in salary and commissions and pays $5,000 for airline tickets, hotel rooms, meals, and other travel expenses, his operations have provided him with funds amounting to $13,000. Further investigation reveals that he borrowed $2,000 from the bank during this same year, bought a car for $3,000, purchased stock for $4,000, and spent $9,000 on ordinary living expenses. At the end of the year, he had $1,000 less in his bank account than he had had at the beginning. A funds statement for the year would show the following:

The flow of funds

sources of funds		uses of funds	
from operations	$13,000	to purchase car	$3,000
from bank borrowing	2,000	to buy stock	4,000
from past savings	1,000	to pay living expenses	9,000
total sources	$16,000	total uses	$16,000

This is very different from an income statement. The sources of funds include amounts borrowed from the bank and taken from past personal savings. These are, of course, not thought of as current revenues. Similarly, the purchases of stock and of the car would not be thought of as current expenses: the purchases are made in the expectation of a future benefit, not mainly to produce current revenues.

The figure referred to as funds from operations ($13,000) comes from the income statement but it is not the same as net income. For example, the salesman may say that his expenses should include some recognition of depreciation on the automobile that he used in his business. If depreciation is estimated to be $600 for the year, his net income would be $12,400, not $13,000.

The purpose of the funds statement is to permit an examination of the firm's financial activities during the period. For example, Table 4 shows the funds statement of Any Company, Inc., for the year 19___. This shows that $77 flowed in from ordinary operations, $20 came from the issuance of the company's own bonds, and $5 came from the sale of securities that the company had held as a long-term investment. Almost half of these funds were used to buy a new plant and equipment. Another $35 went to stockholders in the form of dividends, and the remaining $17 was invested in working capital.

The statement itself serves merely to identify existing relationships. It reveals, for instance, that operations did not generate enough funds to cover the company's dividend and capital investment requirements. An observer may reasonably ask whether this seems to be the result of an aggressive and imaginative expansion program or the product of competitive weaknesses that limit the company's fund-generating capacity. Similarly, the amount of funds going into working capital is a statement of fact; whether it is good or bad depends on the reasons for it.

Disclosure and auditing requirements. A corporation's obligations to issue financial statements are prescribed in

Table 4: Any Company, Inc.: Statement of the Sources and Uses of Funds for the Year Ending Dec. 31, 19__

Sources of funds	
From ordinary operations	$ 77
From sale of investments	5
From sale of bonds	20
Total sources of funds	$102
Uses of funds	
To pay dividends	$ 35
To acquire plant and equipment	50
To increase working capital	17
Total uses of funds	$102

the company's own statutes or bylaws and in public laws and regulations. The financial statements of most large and medium-sized companies in the United States fall primarily within the jurisdiction of the Securities and Exchange Commission (SEC), a federal government agency. The SEC has a good deal of authority to prescribe the content and structure of the financial statements that are submitted to it.

A company's financial statements are ordinarily prepared initially by its own accountants and then reviewed, or audited, by outsiders. In most countries, including the United States, these outside auditors are selected by the company's shareholders. The audit of the statements of a company is ordinarily performed by professionally qualified, independent accountants who bear the title of Certified Public Accountants (CPA) in the United States, and Chartered Accountants in Great Britain and many other countries with British-based accounting traditions. Their primary task is to investigate the company's accounting figures and methods carefully enough to permit them to give their opinion that the financial statements are complete and that the figures are measured on the basis of acceptable measurement principles. This opinion is intended to protect the investor and others who must use the statements in decision making.

MEASUREMENT PRINCIPLES

In preparing financial statements, the accountant has many different measurement systems to choose from. Assets might, for example, be measured at what they cost in the past or what they could be sold for now, to mention only two possibilities. The choice must be based upon concepts of income and capital that are relevant to the purposes of measurement outlined above.

The concept of value. The economic value of an asset is the maximum amount that the firm would be willing to pay for it. This amount depends on what the firm expects to be able to do with it. For business assets, these expectations are usually expressed in terms of forecasts of the future inflows of cash that the company would receive if it owned the assets in the future. If, for example, the company believes that by spending $1 on advertising and other forms of sales promotion it can sell a certain product for $5, then this product is worth $4 to this firm.

When the cash inflows to be produced by an asset are expected to be delayed, value is less than the anticipated cash flow. For example, if the firm has to pay interest at the rate of 10 percent a year, an investment of $100 in a one-year asset today will not be worthwhile unless it will return at least $110 a year from now ($100 plus 10 percent interest for one year). In this example, $100 is the present value of the right to receive $110 one year later. Present value is the maximum amount the firm would be willing to pay for a future inflow of cash after deducting interest on the investment at a specified rate for the length of time the firm has to wait before it receives its cash.

Value, in other words, depends on three factors: (1) the amount of the anticipated future cash flows, (2) their timing, and (3) the interest rate. The lower the expectation, the more distant the timing, or the higher the interest rate, the less valuable the asset will be.

Value may also be represented by the amount the firm could obtain by selling its assets. This sale price is seldom a good measure of the assets' value to the firm, however, because few firms are likely to keep many assets that are worth no more to the firm than their market value. Continued ownership of an asset implies that its present value to the owner exceeds its market value, which is its apparent value to outsiders.

Cost and depreciation

Asset cost. Accountants are traditionally reluctant to accept value as the basis of asset measurement in the going concern. Although monetary assets such as cash or accounts receivable are usually measured by their value, all other assets are measured at cost. The reason for this difference is that the accountant finds it difficult to verify the forecasts upon which a generalized value measurement system would have to be based.

Even so, the cost of a nonmonetary asset is not always easily measurable. If for example, a company has built a special-purpose machine in one of its own factories for use in manufacturing other products, and the project required logistical support from all parts of the factory organization, from purchasing to quality control, then a good deal of judgment must be reflected in any estimate of how much of the costs of these logistical activities should be "capitalized" (*i.e.*, placed on the balance sheet) as part of the cost of the machine.

Some assets give up their services gradually rather than all at once. Such assets are reported on the balance sheet at their original cost, less an allowance for depreciation. The allowance for depreciation represents the cost of the portion of the asset's anticipated lifetime services that has already been used up. To measure this, the accountant must try to predict both how long the asset will continue to provide useful services and how much of its potential to provide these services will be used up in any one period.

Estimates of useful life are generally based on past experience with other assets of the same general type. This evidence is seldom wholly conclusive, but a far greater degree of uncertainty ordinarily attaches to the estimates of how much of the asset's services will be consumed each year. The depreciation charge is usually computed by some simple formula. The two most popular formulas in the United States are straight-line depreciation, in which the same amount of depreciation is recognized each year, and declining-charge depreciation, in which more depreciation is recognized during the early years of life than during the later years, on the assumption that the value of the asset's service declines as it gets older.

The role of the independent accountant is to see whether the company's estimates are based on formulas that seem reasonable in the light of whatever evidence is available and that these formulas are applied consistently from year to year. Again, what is "reasonable" is clearly a matter of judgment.

Income. One possible method of income measurement would be to measure the value of the firm's assets from time to time and measure income as the difference between successive measurements of value, adjusted for additions or withdrawals of investment capital during the period. Thus if value has gone from $1,000 to $1,200 and dividends of $100 have been distributed, income would be $300 ($1,200 minus $1,000, plus $100).

Accountants generally have rejected this approach for the same reason that they have found value an unacceptable basis for asset measurement: such a measure would rely too much on estimates of what will happen in the future, estimates that would not be readily susceptible to independent verification. Instead, the accountant has adopted what might be called a transactions approach to income measurement. He recognizes as income only those increases in wealth that can be substantiated from data pertaining to actual transactions that have taken place with persons outside the firm. In such systems, income is measured when work is performed for an outside customer, when goods are delivered, or when the customer is billed.

Recognition of income at this time requires two sets of estimates: (1) revenue estimates, representing the value of the cash that the company expects to receive from the

customer; and (2) expense estimates, representing the resources that have been consumed in the creation of the revenues. Revenue estimation is the easier of the two, but it still requires judgment. The main problem is to estimate the percentage of the gross sale price that will never be received, either because the customer will not pay his bills or because he will demand and receive credit for returned merchandise or defective work.

Expense estimates are generally based on the historical cost of the resources consumed. The net income figure, in other words, is the difference between the value received from the use of resources and the cost of the resources that were consumed in the process. As with asset measurement, the main problem is to estimate what portion of the cost of an asset has been consumed during the period in question.

The inclusion of holding gains

From an economic point of view, income can be defined as the change in the firm's wealth during a period of time, from all sources other than the injection or withdrawal of investment funds. Income is the amount that the firm could consume during the period and still have as much real wealth at the end of the period as it had at the beginning.

In concept, any increase in wealth increases the well-being of the firm and thus fits this definition, whether it comes from the firm's use of its assets or from a holding gain. A holding gain occurs whenever an asset's market value increases. Conversely, a holding loss occurs when the market value declines. For example, construction of a superhighway may increase the value of the company's land. This is a holding gain. Construction of a glue factory in the vicinity, on the other hand, may decrease land values. Such decreases are holding losses.

The accountant ordinarily reports holding gains only when assets are sold. Such gains are said to be realized. Unrealized holding gains are seldom reported to management and are reported even less frequently to the outside public. Holding losses, on the other hand, are ordinarily reported as soon as they can be identified, whether they are realized or not. Thus if land values have fallen below cost, the company will transfer a portion of the cost from the balance sheet to the income statement as a loss.

Accounting measurements of capital and income have been criticized on the grounds that they are not responsive enough to the dynamic influences in the modern economy. In some cases the strength and growth of the firm may depend as much or more on the wealth changes it excludes from reported income as on those it includes. To an oil producer, for instance, finding the oil may be a far more important economic event than bringing it out of the ground, refining it, or selling it. A measure of income that excludes wealth increments from this source is a poorer measure of overall economic performance than a broader measure that encompasses all wealth increases, no matter what their source.

To remedy this condition, proposals have been made to recognize all holding gains and losses, realized and unrealized, and report them in a separate section of the income statement. To accomplish this, all nonmonetary assets would have to be restated periodically, either by independent appraisal or by the application of price index numbers to the historical cost figures.

Price changes. Accounting reports have also been criticized on the grounds that they confuse monetary measures with the underlying realities when the prices of many goods and services have been changing rapidly. For example, if the wholesale price of an item has risen from $100 to $150 between the time the company bought it and the time it is sold, many accountants and businessmen claim that $150 is the better measure of the amount of resources consumed by the sale. The $50 is simply a holding gain that should not be classified as ordinary income.

This proposal to measure expense on the basis of current replacement cost rather than the original cost of the resources consumed is supported by the proponents of recognizing all holding gains as income. The support is not always reciprocated, however. Not all advocates of replacement costing would agree to the recognition of unrealized holding gains as current income, and many would exclude all holding gains that arise from increases in replacement prices.

This latter attitude often reflects a confusion between changes in the replacement prices of specific resources and changes in the average level of all prices. An increase in value that results solely from a rise in the general price level is not a real increase and therefore is not really income. The bridge from this to the belief that all holding gains are unreal and unearned is easy to understand but difficult to defend. Holding gains are real gains to the extent that they exceed the amount that would have resulted if the prices of the company's assets had advanced in proportion to the increase in prices generally.

Units of purchasing power

To adjust for the effects of an inflation in general price levels, many accountants advocate the use of units of constant purchasing power instead of ordinary monetary units. Upon acquisition, an asset would be stated as the equivalent of a given amount of purchasing power. Its replacement cost would then be estimated periodically, again in terms of the number of purchasing power units it represented. The holding gain or loss would then be measured by the change in the amount of purchasing power the asset represented.

If, for example a piece of land is purchased for $10,000 at a time when $10,000 can be used to buy the same basket of goods and services that could have been obtained for $8,000 in some base period, the general price level is $\$10{,}000/\$8{,}000$ or 125 percent of the level prevailing in the base period. The land will then be recorded as having cost 8,000 purchasing power units (designated $*).

If a year later an appraisal indicates that the land could be replaced at a price of $11,700 but the general price level has risen by 5 percentage points to 130 percent of the base level, then the $11,700 price is equivalent to $\$11{,}700/1.3$, or 9,000 purchasing power units. The monetary gain is $1,700 ($11,700 minus $10,000), but the gain in purchasing power units is only $*1,000 ($*9,000 minus $*8,000).

In countries that have experienced severe and prolonged inflation, companies have been allowed or even required to restate their assets to reflect the more recent and higher levels of purchase prices. The increment in the asset balances in such cases has not been reported as income, but depreciation thereafter has been based on these higher figures.

Individual companies in the United States, the Netherlands, and some other countries have made similar adjustments in their financial statements, but the practice has been slow to spread. The main reason for this has been the accountant's reluctance to agree that appraisals and price index numbers are objective, reliable, and verifiable enough to replace the historical cost figures with which he is familiar.

MANAGERIAL ACCOUNTING

Although published financial statements are the most widely visible outputs of business accounting systems and the ones with which the public is most concerned, they are only the tip of the iceberg. Most accounting data and most accounting reports are generated solely or mainly for the company's managers. Reports to management may be either summaries of past events, forecasts of the future, or a combination of the two. Preparation of these reports is the focus of managerial accounting.

The varied work of the managerial accountant can be classified into three broad functions: (1) cost finding, (2) budgetary planning and performance reporting, and (3) cost and profit analysis.

Cost finding. Cost finding is the process by which the company obtains estimates of the costs of producing a product, providing a service, performing a function, or operating a department. Some of these estimates are historical—how much did it cost?—while others are predictive of what it will cost in the future.

The most fully developed methods of cost finding are used to estimate the costs that have been incurred in a

factory to manufacture different products. The simplest of these methods is known as process costing. In this method, the accountant first accumulates the costs of each separate production operation or process for a specified period of time. The total of these costs is then restated as an average by dividing it by the total output of the process during the same period.

Process costing can be used whenever the output of individual processes is reasonably uniform or homogeneous, as in cement manufacturing, flour milling, and other relatively continuous production processes.

Job order costing

A second method, job order costing, is used when individual production centres or departments work on a variety of products rather than just one during a typical time period. Two categories of factory cost are recognized under this method: prime costs and factory overhead costs. Prime costs are those that can be traced directly to a specific batch, or job lot, of products. These are the direct labour and direct materials costs of production. Overhead costs, on the other hand, are those than can be traced only to departmental operations or to the factory as a whole and not to individual job orders. The salary of a departmental foreman is an example of an overhead cost.

Direct materials and direct labour costs are recorded directly on the job order cost sheets, one for each job. Because overhead costs cannot be traced to individual jobs, they are accumulated separately, with one set of overhead accounts for each factory department.

Although not traceable to individual jobs, overhead costs are generally assigned to them by means of burden rates; *i.e.*, the ratio of total overhead cost to total production volume for a given time period. A separate burden rate is usually calculated for each production department, and if the operations of a department are very varied, it is often subdivided into a set of more homogeneous cost centres, each with its own burden rate. Separate burden rates are sometimes even used for individual processing machines within a department if the machines differ widely in such factors as power consumption, maintenance cost, and depreciation.

Because output within a cost centre is not homogeneous, production volume must be measured by something other than the number of units of product, such as the number of machine hours or direct labour hours. Once the burden rate has been determined, a provision for overhead cost can be entered on each job order cost sheet on the basis of the number of direct labour hours or machine hours used on that job. For example, if the burden rate is $3.00 a machine hour and Job No. 7128 used 600 machine hours, then $1,800 would be shown as the overhead cost of this job.

Most burden rates are predetermined—that is, they represent the average planned overhead cost at some production volume. The main reason for this is that actual overhead cost averages depend on the total volume and efficiency of operations and not on any one job alone. The relevance of job order cost information will be impaired if these external fluctuations are allowed to change the amount of overhead cost assigned to a particular job.

The methods of cost finding described in the preceding paragraphs are known as full costing, or absorption costing, methods, in that the burden rates are intended to include provisions for all manufacturing costs. Both process and job order costing methods can also be adapted to variable costing, or direct costing, in which only variable manufacturing costs are included in product cost. Variable costs are those that will be greater in total in the upper portions of the company's normal range of volumes than in the lower portion of this range. Total fixed costs, in contrast, are the same at all volume levels within the normal range.

Variable costing

Unit cost under variable costing represents the average variable cost of making the product. The main argument for the variable costing approach is that average variable cost is more relevant to most managerial decisions than average full cost. In deciding whether to manufacture goods in small or large lots, for example, management needs to estimate the cost of carrying different amounts of finished goods in inventory. (A large lot size leads to a larger average inventory.) More variable costs will have to be incurred to build up the inventory to a higher level; fixed manufacturing costs presumably will be unaffected.

Furthermore, when a management decision changes the company's fixed costs, the change is unlikely to be proportional to the change in volume; therefore, average fixed cost is seldom a valid basis for estimating the cost effects of such decisions. Variable costing eliminates the temptation to assume without question that average fixed cost can be used to estimate changes in total fixed cost.

Product cost estimates

Estimates of product costs can be built up in exactly the same way as historical averages. Once the route that the product will take through the factory is known, estimates of the amounts of materials, labour, and overhead required at each stage can be written down, totalled, and averaged.

Many systems go even farther than this. Estimates of the average costs of each type of material, each operation, and each product are prepared routinely and identified as standard costs. These standard cost figures are then readily available whenever estimates are needed and also serve as an important element in the company's performance reporting system, as described below.

Distribution cost analysis. Similar methods of cost finding can be used to determine or estimate the cost of providing services rather than physical goods. Most advertising agencies and consulting firms, for example, maintain some form of job cost records, either as a basis for billing their clients or as a means of estimating the profitability of individual job or accounts.

Estimates of the unit cost of distributing the company's products or of providing administrative and supporting services are more difficult to determine, however, and few companies attempt to derive unit costs in this area on a regular basis. Instead, estimates are prepared from time to time on the basis of special studies, mostly centring on observations of how employees spend their time during selected test periods. Providing these estimates is the task of distribution cost analysis.

Budgetary planning and performance reporting. The second major component of internal accounting systems for management's use is the company's budgetary planning and performance reporting system. This consists of two highly interrelated elements: (1) establishing budgetary plans and setting performance standards and (2) measuring actual results and reporting differences between actual performance and the plans.

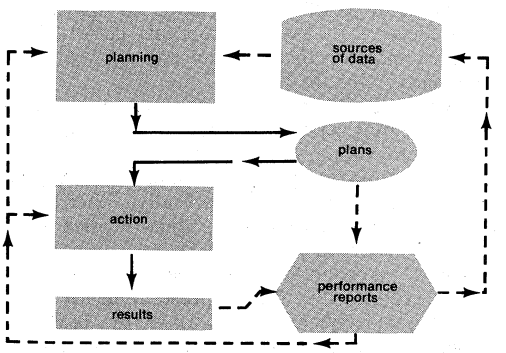


Figure 1: Budget planning and performance reporting.

The simplified diagram in Figure 1 illustrates the interrelationships between these two elements. Planning leads to plans, which then are translated into action. The results of these actions are compared with the plans and reported to management in comparative form. Management can then respond to substantial deviations from plan, either by taking corrective action, or if outside conditions differ from those predicated in the plans, by preparing revised plans.

Although plans can be either broad, strategic outlines of

the company's future, or schedules of the inputs and outputs associated with specific independent programs, most business plans are periodic plans—that is, they refer to company operations for a specified period of time. These periodic plans are summarized in a series of projected financial statements, or budgets.

The profit plan

The two principal budget statements are the profit plan and the cash forecast. The profit plan is an estimated income statement for the budget period. It summarizes the planned level of selling effort, reflected as selling expense, and the results of that effort, shown as sales revenue and the accompanying cost of goods sold. Separate profit plans are ordinarily prepared for each major segment of the company's operations.

The details underlying the profit plan are contained in departmental sales and cost budgets, each part identified with the executive or group responsible for carrying out that part. Figure 2 shows the essence of this relationship. The diagram in Figure 2, for example, shows that the company's profit plan is really the integrated product of the plans of its two major product divisions. The arrows connecting the two divisional plans represent the coordinative communications that tie them together on matters of mutual concern.

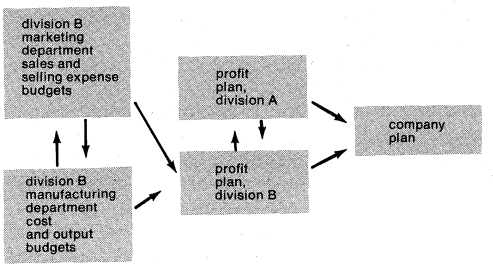


Figure 2: Relationship of company profit plan to responsibility structure.

The exhibit also goes one level farther down, showing that division B's profit plan is really a coordinated synthesis of the plans of the division's marketing and manufacturing departments. Arrows again emphasize the necessary coordination between the two. Each of these departmental plans, in turn, is a summary of the plans of the major offices, plants, or other units within the division. A complete representation of the company's profit plan would require an extension of the diagram to encompass every single responsibility centre in the entire company.

If standard costs have been established, they are shown in the factory cost budgets and in the standard cost of goods sold on the summary profit plans, with anticipated deviations, or variances, from standard cost budgeted separately. Many companies also prepare alternative budgets for operating volumes other than the volume anticipated for the period. A set of such alternative budgets is known as the flexible budget. The practice of flexible budgeting has been adopted widely by factory management to facilitate evaluation of cost performance at different volume levels and has also been extended to other elements of the profit plan.

The cash forecast

The second major component of the annual budgetary plan, the cash forecast or cash budget, summarizes the anticipated effects on cash of all of the firm's activities. It lists the anticipated cash payments, cash receipts, and the amount of cash on hand, month by month throughout the year. In most companies, responsibility for cash management rests mainly in the head office rather than at the divisional level. For this reason, divisional cash forecasts tend to be less important than divisional profit plans.

Company-wide cash forecasts, on the other hand, are just as important as company profit plans. Preliminary cash forecasts are used in deciding how much money will be made available for the payment of dividends, for the purchase or construction of buildings and equipment, and for other programs that do not pay for themselves immediately. The amount of short-term borrowing or short-term investment of temporarily idle funds is then generally geared to the requirements summarized in the final, adjusted forecast.

Other elements of the budgetary plan, in addition to the profit plan and the cash forecast, include capital expenditure budgets, manpower budgets, production budgets, and budgeted balance sheets. They all serve the same purpose: to help management decide upon a course of action and to serve as a point of reference against which to measure subsequent performance.

Planning is a management responsibility, not an accounting function. To plan is to decide, and only the manager has the authority to choose the direction the company is to take. Accounting personnel are nevertheless deeply involved in the planning process. First, they administer the budgetary planning system, establishing deadlines for the completion of each part of the process and seeing that these deadlines are met. Second, they analyze data and help management at various levels compare the estimated effects of different courses of action. Third, they are responsible for collating the tentative plans and proposals coming from the individual departments and divisions and then reviewing them for consistency and feasibility and sometimes for desirability as well. Finally, they must draw up the final plans and see that these plans are understood by the operating executives.

Once the budgetary plan has been adopted, accounting's next task is to prepare information on the results of company activities and make it available to management. The manager's main interest in this information centres on three questions: Have his own actions had the results he expected, and if not, why not? How successful have his subordinates been in managing the activities entrusted to them? What problems and opportunities seem to have arisen since the budgetary plan was prepared? For these purposes, the information must be comparative, relating actual results to the level of results that management regards as satisfactory. In each case, the standard for comparison is provided by the budgetary plan, adopted by the manager himself and accepted by his superiors.

Much of this information is contained in periodic financial reports. At the top management and divisional levels, the most important of these is the comparative income statement, one of which is illustrated in Table 5. This shows the profit planned for this period, the actual results, and the differences between the two. It also gives an explanation of some of the reasons for the difference between planned and actual income.

Table 5: Any Company, Inc.: Comparative Income Statement for the Month of October 19__

(in $000,000)

	actual	planned	difference
Divisional profit contribution			
Division A	$100	$120	$(20)
Division B	70	65	5
Division C	140	150	(10)
Total profit contribution	$310	$335	$(25)
Head office expenses	110	105	(5)
Net income before taxes	$200	$230	$(30)

Comments

(1) An industry-wide strike in customer factories caused a slowdown in division A deliveries, reducing profit by $24,000,000.

(2) Demand continued strong in division B, reflected in increased volume and a stronger price structure.

(3) Division C's unfavourable results were the result of production losses due to the model change and heavy introductory costs for the new line. These losses should be recovered by the end of the year.

(4) Salary increases account for most of the variance in head office expense. The methods department is seeking ways to offset this.

The report in this exhibit employs the widely used profit contribution format, in which divisional results reflect sales and expenses traceable to the individual divisions, with no deduction for head office expenses. Company net income is then obtained by deducting head office expenses as a lump sum from the total of the divisional profit contributions.

Measuring variances

A similar format can be used within the division, reporting the profit contribution of each of the division's product lines, with divisional headquarters expenses deducted at the bottom. By far the greatest number of reports, however, are cost or sales reports, mostly on a departmental basis. Departmental sales reports usually represent comparisons of actual sales with the volumes planned for the period. Departmental cost performance reports, in contrast, typically compare actual costs with standards or budgets that have been adjusted to correspond to the actual volume of work done during the period. This practice reflects a recognition that volume fluctuations generally originate outside the department and that the department head's responsibility is ordinarily to minimize cost while meeting the delivery schedules imposed on him by higher management.

For example, a factory department's output consists entirely of a single product, with a standard materials cost of $3 a unit and standard labour cost of $6. Materials cost represents three pounds of raw materials at $1 a pound; standard labour cost is two hours of labour at $3 an hour. Overhead costs in the department are budgeted at $10,000 a month plus $2 a unit. Under normal conditions, volume is 7,000 units a month, but during October 19___, only 6,000 units were produced. The cost standards for the month would be as follows:

materials
6,000 units × 3 lb per unit × $1.00 per lb = $18,000
labour
6,000 units × 2 hr per unit × $3.00 per hr = $36,000
overhead
6,000 units × $2.00 per unit + $10,000 = $22,000

The actual cost this month was $17,850 for materials (17,000 pounds at $1.05), $38,750 for labour (12,500 hours at $3.10 an hour), and $23,000 for overhead. A summary report would show the following:

	actual	standard	variance
materials	$17,850	$18,000	$150
labour	38,750	36,000	(2,750)
overhead	23,000	22,000	(1,000)
total	$79,600	$76,000	$(3,600)

These variances are usually analyzed further in attempt to identify the underlying causes. The labour variance, for example, can be seen as a result of both high wage rates ($3.10 instead of $3.00) and high labour usage (12,500 hours instead of 12,000). The factory accountant ordinarily would measure the effect of the rate change in the following way:

actual labour cost = 12,500 × $3.10 = (actual hours × actual wage rate)	$38,750
actual labour hours at standard wage rate = 12,500 × $3.00 =	37,500
labour rate variance	$(1,250)

The labour usage variance would then be obtained as follows:

actual labour hours × standard wage rate = 12,500 × $3.00	$37,500
standard labour cost (standard labour hours × standard wage rate) = 6,000 × 2 × $3.00	36,000
labour usage variance	$(1,500)

In most cases, the labour rate variance would not be reported to the department head, because it is not subject to his control.

Cost and profit analysis. The accountant shares with many others the task of analyzing cost and profit data to provide guidance in managerial decision making. Even if the analytical work is done largely by others, he has an interest in analytical methods because the systems he designs must collect data in forms suitable for analysis.

Managerial decisions are based on comparisons of the estimated future results of the alternative courses of action that the decision maker is choosing among. Recorded historical accounting data, in contrast, reflect conditions and experience of the past. Furthermore they are absolute, not comparative, in that they show the effects of one course of action but not whether these were better or worse than those that would have resulted from some other choice.

For decision making, therefore, historical accounting data must be examined, modified, and placed on a comparative basis. Even estimated data, such as budgets and standard costs, must be examined to see whether the estimates are still valid and relevant to managerial comparisons. To a large extent, this job of review and restatement is an accounting responsibility. Accordingly, a major part of the accountant's preparation for his profession is devoted to the study of methods and principles of analysis for managerial decision making.

OTHER PURPOSES OF ACCOUNTING SYSTEMS

Accounting systems are designed mainly to provide information that managers and outsiders can use in decision making. They also serve other purposes: to produce operating documents, to protect the company's assets, and to provide data for company tax returns.

Document preparation

The accounting organization is responsible for preparing documents that contain instructions for a variety of tasks, such as payment of customer bills or preparing employee payrolls. It also must prepare documents that serve what might be called private information purposes, such as the employee's own record of his salary and wages. Many of these documents also serve as sources of data for other accounting purposes, but they would have to be prepared even if no information reports were necessary. Measured by the number of people and the amount of time required, document preparation is one of accounting's biggest jobs.

Accounting systems must provide means of reducing the chance of losses of assets due to carelessness or dishonesty on the part of employees, suppliers, and customers. Asset protection devices are often very simple; for example, many restaurants use numbered meal checks so that waiters will not be able to submit one check to the customer and another, with a lower total, to the cashier. Other devices entail a partial duplication of effort or a division of tasks between two individuals to reduce the opportunity for unobserved thefts.

These are all part of the company's system of internal controls. Another important element in the internal control system is internal auditing. The task of the internal auditor is to see whether prescribed data handling and asset protection procedures are being followed. To accomplish this, he usually observes some of the work as it is being performed and examines a sample of past transactions for accuracy and fidelity to the system. He may insert a set of fictitious data into the system and see whether the resulting output meets a predetermined standard. This last technique is particularly useful in testing the validity of the programs that are used to process data through electronic computers.

Finally, the accounting system must provide data for use in the completion of the company's tax returns. This function is the concern of tax accounting. In some countries financial accounting must obey rules laid down for tax accounting by national tax laws and regulations, but no such requirement is imposed in the United States, and figures prepared for tax purposes often diverge from those submitted to shareholders and others. "Taxable income" is a legal concept rather than an accounting concept. Tax laws include incentives to encourage companies to do certain things and discourage them from doing others. Accordingly, what is "income" or "capital" to a tax agency may be far different from the accountant's measures of these same concepts.

BIBLIOGRAPHY

Classics: R.G. BROWN and K.S. JOHNSTON, *Paciolo on Accounting* (1963), an annotated translation of the treatise, published in 1494, that is widely accepted as the foundation stone of modern accounting and bookkeeping systems, with biographical notes and a reproduction of the original manuscript; A.C. LITTLETON, *Accounting Evolution to 1900* (1933), a scholarly, well-written analysis of the development of accounting from the Renaissance to modern times; J.B. CANNING, *Economics of Accountancy* (1929), an early landmark in the development of 20th-century accounting thought and one of the first systematic attempts to build a structure of accounting on the basis of economic theory; E. SCHMALENBACH, *Dynamic Accounting* (1959), a translation of the 12th edition of the most influential accounting book published in Germany in the first half of the 20th century; Schmalenbach's ideas have had a measurable impact on the accounting systems in use in most continental European countries and underlie many of the uniform accounting schemes that have been proposed or adopted in individual European countries.

Current references: E.O. EDWARDS and P.W. BELL, *The Theory and Measurement of Business Income* (1961), a critical review of conventional accounting measurements, with a detailed examination of possible alternative measurement systems; R.N. ANTHONY, *Essentials of Accounting* (1964), an easy-to-follow, self-teaching guide to accounting fundamentals, using the "programmed learning" approach; E.S. HENDRIKSEN, *Accounting Theory*, rev. ed. (1970), a current textbook, providing a historical survey and a critical examination of the measurement systems underlying business financial statements; G. SHILLINGLAW, *Cost Accounting: Analysis and Control*, rev. ed. (1967), a current textbook providing a thorough coverage of managerial and cost accounting.

(G.Sh.)

Accra

Accra, capital of Ghana, is in population and area the largest city in the country. Architecturally and otherwise, it piquantly blends modern and traditional West African characteristics.

History. When the Portuguese first settled on the coast of what is now Ghana in 1482, the present site of Accra was occupied by several villages of the Ga tribe, ruled from a parent settlement, Ayaso or Ayawaso, located about 15 miles north. Subsequently, between 1650 and 1680 two forts and a castle were built by Europeans on promontories to seaward of the villages. These fortified trading posts were, from west to east, Ft. James (English), Ft. Crevecoeur (Dutch), and Christiansborg Castle (Danish). At about the same time as these European posts were being built, Ayaso was destroyed in a tribal war, and its population, together with that of the other major Ga towns on the Accra plains, was drawn to the coast by the prospect of profitable trade with the Europeans. As a result, three coastal villages—Osu (Christiansborg), Dutch Accra (later known as Ussher Town), and James Town—sprang up, becoming the nuclei of what was to be Accra. The name Accra itself is a corruption of the Akan word *nkran*. It refers to the black ants, which abound in the vicinity. It came to be applied to the inhabitants of this part of the Accra plains.

Two major factors accounted for the subsequent growth of Accra—firstly trade and secondly the transfer in 1877 of the capital of the British Gold Coast colony from Cape Coast to the village of Osu (Christiansborg). (The Danes and the Dutch had left in 1850 and 1872, respectively.) Accra was made the capital for three reasons. It was considered to be healthier, being close to the sanatorium at Aburi in the nearby Akwapim Mountains; it was a prosperous trading centre, and the surroundings were judged to be aesthetically pleasing. At this time James Town and Ussher Town formed a compact mass of buildings separated by narrow twisting lanes. Osu was another huddle of buildings, separated from the James Town–Ussher Town complex by a stretch of open country. It was also independent politically, with its own chief.

Population growth

In 1898 the Accra Municipal Council was formed to improve the town but could achieve little at first. The insanitary conditions that prevailed led to an outbreak of bubonic plague in 1907. Later, under the successful Ten Year (1920–30) Development Plan launched by Governor Gordon Guggisberg, Accra was systematically laid out. This and later development was accomplished by rapid population increase—from 60,726 in 1931, to 133,192 in 1948, to 337,828 in 1960, and then to about 634,000 in 1970.

Physical geography. Backed by the Accra Plains, the city's site is low-lying—elevation varies from 10 to 100 feet above sea level—and is flat or gently undulating. Inland the elevation increases gradually, and several hills stand out abruptly from the plain to the north of the town. On one of the hills, Legon, is the University of Ghana. Geologically, the flatness of the land is due to the weathering of underlying rocks followed by the deposition of clays, loose sand, and horizontally laid sandstones by the sea. The rocks have undergone much faulting. Topography and geological history thus pose problems for Accra. In many parts of the city the ground is not firm enough, because of the sands and clays, to provide an adequate natural foundation for buildings. Also, the land, being flat, floods readily, and its susceptibility to faulting is the cause of frequent earthquakes. The most violent recorded so far was in 1939.

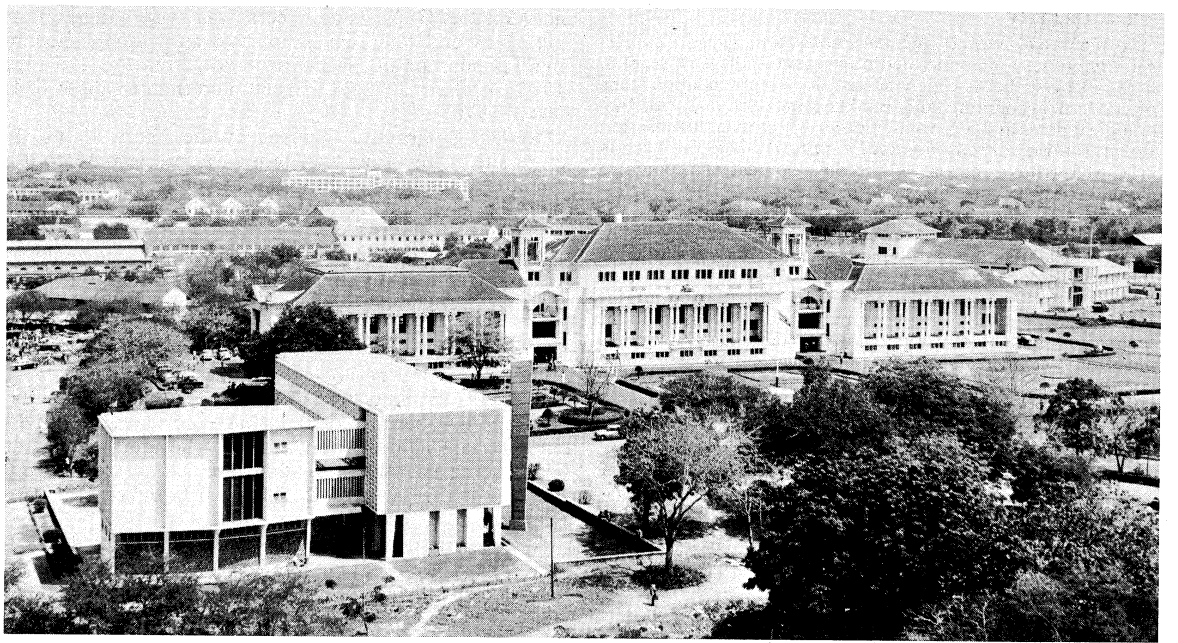
A seasonal river, the Odaw, flows through the city to enter the Korle Lagoon—the largest lagoon in the city—which forms the western boundary of James Town–Ussher Town. Elsewhere smaller seasonal streams enter the sea via lagoons. The most important river, however, is the Densu, close to the western boundary, the source of Accra's water supply. The Densu rises in the high rain forest region of the interior and is perennial.

Accra has a peculiarly low annual rainfall of about 29 inches. Except for this anomaly, the climate is typically equatorial, with monthly noon temperatures deviating little from the annual mean 84.5° F (29.2° C). Daily variations in temperature are slightly greater than the monthly variations. Relative humidities are lower than elsewhere on the coast, as a drier atmosphere attenuates the "stickiness" of the weather. Associated with this dry climate is a grassland vegetation dotted with clumps of bush. There is evidence that the vegetation was formerly more lush.

Accra is roughly 15 miles from west to east along the coast and 8½ miles at its widest north-south extent. Its population is mixed. Before 1969, when a number of African residents from other countries were deported, the local Ga formed roughly 40% and the Akan and Ewe immigrants together 38% of the total population. The rest were non-Ghanaian Nigerians and Togolese forming 40% and 30% respectively of the non-Ghanaian population. The extent of change in the ethnic composition since the mass exodus of non-Ghanaian Africans in 1969 remains to be determined. About 4.5% of the total population is rural, engaged in fishing and farming, 28% is in manufacturing, construction, electricity etc., 44% in commerce, transport, and related activities, and the rest in other services. Christians form about 62% of the total population, Muslims about 15%, and those in traditional religion 12%. The remaining 11% apparently do not profess any religion.

The people and the city. The population lives in clearly demarcated residential areas, where the architectural styles indicate the age of the buildings. For example, the oldest parts of Osu have a mixture of African buildings (single-story, rectangular, and gable roofed, and often grouped around a courtyard), and large two-storied buildings of European style with wide verandas on pillars, and rainwater storage tanks in walled-in courtyards in the rear. In the new residential areas, on the other hand, zoning policy permits only expensive modern houses to be built.

The area of uninterrupted settlement stretches in a broad arc, concave to the sea, from Labadi in the east through Tesano to Mamprobi-Chorkor in the west. Teshi, Nungua, Achimota, and other residential areas outside this arc are true suburbs. All the residential areas are being developed into neighbourhood units with markets, shops, police stations, elementary schools, clinics, cinemas, etc.



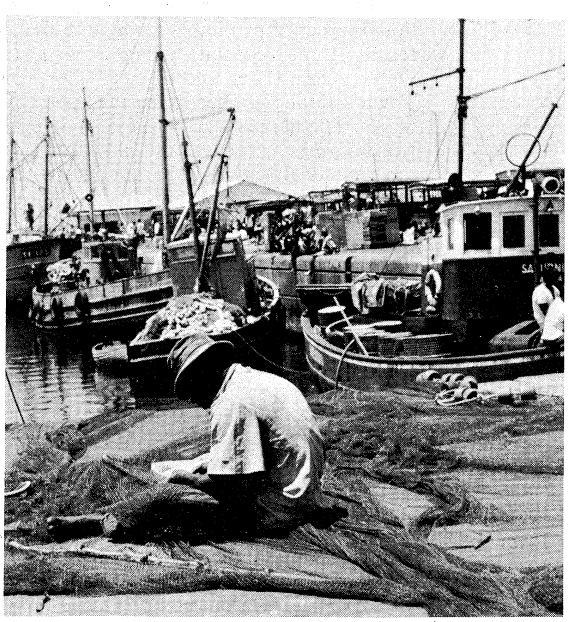
Public buildings of Accra. In the left foreground is the library, and in the centre and far right are the Supreme Court and Parliament buildings.
Ewing Galloway

Living conditions

There are marked variations in living conditions in different areas. At one extreme is Nima, the worst slum in Accra and the source of many social problems. Nima is overcrowded, and the population poorly housed, with densities reaching 400 persons an acre. In 1968 a mild outbreak of smallpox in the city originated there. Sabon Zongo, originally a quarter for northern Ghanaians only, is another such slum, with conditions little better than those in Nima. Other poor residential areas are former outlying villages—now suburbs—such as Okponglo and Odorkor, as well as those areas that were the original nuclei of the city. Ussher Town, for example, has population densities approaching those in Nima. At the other extreme are such well-to-do residential areas as the Ridge (built originally for Europeans and now occupied by Ghanaian senior civil servants), the Airport Residential Area, the Cantonments, and Tesano (where most of the houses are privately owned). There are also housing developments, of which some, like Ringway Estate and North Labone Estate, are designed for persons with high incomes, while others, such as Korle Gonno, Mamprobi, and Labadi, are more modest. Population densities in the wealthier areas average up to 10 persons per acre, and in the middle income areas average between 50 to 100 persons per acre.

The central business district, in which most of the multistoried buildings in the city are found, lies between James Town and Osu and has prolongations along Liberty Avenue and Kojo Thompson Road as far as Adabraka. Houses increase in frequency away from the core area. In the district are located the head offices of all of the large banks in the country (Bank of Ghana, Ghana Commercial Bank, Standard Bank of West Africa), the large trading firms (mostly foreign owned), the insurance agencies, the electricity corporation, the general post office, the large open markets to which most of the food supplies come, the Community Centre, the Supreme Court, and the Accra Central Library buildings.





(Left) Osu (Christiansborg) Castle, Accra, built by the Danes in the 17th century, and now headquarters of Ghana's chief of state. (Right) Fisherman repairing nets at Tema, near Accra.
(Left) Camera Press—Publix, (right) Authenticated News International

Half way between Ussher Town and Christiansborg is the administrative district, containing the government ministries and the offices of the Volta River Authority, as well as the City Council, which is directly responsible for the running of the city.

At Labadi, on Brewery Road, Ring Road West, and near the Korle Lagoon are the industrial estates, where vehicle assembly plants, workshops, and garment factories are located. The disposal of industrial waste does not cause serious pollution.

West of the Korle Lagoon from Ussher Town is Korle Bu General Hospital, the largest and best staffed in the country. (The protein deficiency disease, common in the tropics, known as kwashiorkor—a word meaning "golden boy" or "red boy" in the Ga language—was first given its distinctive name at Korle Bu in 1933.) It also houses the Ghana Medical School. Elsewhere, hospitals and polyclinics, both privately and state owned, are dispersed throughout the city. Similarly scattered are the numerous night clubs and hotels (the Ambassador, Continental, and Star hotels are the three largest), elementary and secondary schools, and churches. There are three cathedrals—Holy Spirit (Roman Catholic), Holy Trinity (Anglican), and Methodist. Close to the Holy Spirit Cathedral stand the national archives and the national museum. Also located in the city, and close to the Airport Residential Area, are the offices of the Council for Scientific and Industrial Research and the Ghana Academy of Arts and Sciences.

Recreational areas

Throughout the city are to be found many large open spaces, including undeveloped areas, cemeteries, car and lorry parks, landscape parks, gardens, and football parks. The football stadium and the race course are both close to the government ministries. Also nearby is Black Star Square, in which stands Independence Arch, used for ceremonial parades.

Accra is well paved. Its most important streets radiate inland from the central business district and are connected to one another by Ring Road. In each part of the city the streets form a clear pattern (usually concentric or gridiron). Public transport along them is provided by City Council buses (the bus service was inaugurated in 1927) but buses are so few that greater reliance is placed on privately owned taxis and passenger motor vehicles known locally as "tro-tro" (from "tro," the Ga word for threepence, which was once the fare).

A focus for converging roads, Accra is also connected directly by rail to Kumasi, as well as to the new port city of Tema, 17 miles to the east, which has taken over Accra's port function. Accra is a terminus of the internal air service and has Ghana's only international airport, Kotoka International Airport. Also located in Accra are the Ghana Broadcasting System (Radio Ghana) studios, the headquarters of all radio and television services, and the offices of the principal newspapers.

BIBLIOGRAPHY. GHANA, TOWN AND COUNTRY PLANNING DIVISION, *Accra, Plan for the Town: The Report for the Minister of Housing* (1958), an invaluable background study on the problems of physical planning in the city; K.B. DICKSON, *A Historical Geography of Ghana* (1969); I. ACQUAH, *Accra Survey* (1958), a sociological study.

(K.B.D.)

Acid–Base Reactions and Equilibria

Acids are chemical compounds that show, in water solution, a sharp taste, a corrosive action on metals, and the ability to turn certain blue vegetable dyes red. Bases are chemical compounds that, in solution, are soapy to the touch and turn red vegetable dyes blue. When mixed, acids and bases neutralize one another and produce salts, substances with a salty taste and none of the characteristic properties of either acids or bases.

Early ideas about acids and bases

The idea that some substances are acids, whereas others are bases, is almost as old as chemistry; and the terms acid, base, and salt occur very early in the writings of the medieval alchemists. Acids were probably the first of these to be recognized, apparently because of their sour taste. The English word acid, the French *acide*, the German *Säure*, and the Russian *kislota* are all derived from words meaning sour (Latin *acidus*, German *sauer*, Old Norse *sūur*, and Russian *kisly*). Other properties associated at an early date with acids were their solvent, or corrosive, action; their effect on vegetable dyes; and their effervescence when applied to chalk (production of bubbles of carbon dioxide gas). Bases (or alkalies) were characterized mainly by their ability to neutralize acids and form salts, the latter being typified rather loosely as crystalline substances soluble in water and having a saline taste.

In spite of their imprecise nature, these ideas served to correlate a considerable range of qualitative observations; and many of the commonest chemical materials that early chemists encountered could be classified as acids (hydrochloric, sulfuric, nitric, and carbonic acids), bases (soda, potash, lime, ammonia) or salts (common salt, salammoniac, saltpetre, alum, borax). The absence of any apparent physical basis for the phenomena concerned made it difficult to make quantitative progress in understanding acid–base behaviour, but the ability of a fixed quantity of acid to neutralize a fixed quantity of base was one of the earliest examples of chemical equivalence: the idea that a certain measure of one substance is in some chemical sense equal to a different amount of a second substance. In addition, it was found quite early that one acid could be displaced from a salt with another acid, and this made it possible to arrange acids in an approximate order of strength. It also soon became clear that many of these displacements could take place in either direction according to experimental conditions. This phenomenon suggested that acid–base reactions are reversible—that is, that the products of the reaction can interact to regenerate the starting material. It also introduced the concept of equilibrium to acid–base chemistry—this concept states that reversible chemical reactions reach a point of balance, or equilibrium, at which the starting materials and the products are each regenerated by one of the two reactions as rapidly as they are consumed by the other.

Apart from their theoretical interest, acids and bases play a large part in industrial chemistry and in everyday life. Sulfuric acid and sodium hydroxide are among the products of chemical industry manufactured in largest amount, and a large percentage of chemical processes involve acids or bases as reactants or as catalysts. Almost every biological chemical process is closely bound up with acid–base equilibria in the cell, or in the organism as a whole, and the acidity or alkalinity of the soil and water are of great importance for the plants or animals living in them. Both the ideas and the terminology of acid–base chemistry have permeated daily life, and the term salt is especially common.

THEORETICAL APPROACHES TO THE DEFINITION OF ACIDS AND BASES

Definitions in terms of hydrogen and hydroxide ions. The first attempt at a theoretical interpretation of acid behaviour was made by the French chemist Antoine-Laurent Lavoisier at the end of the 18th century. Lavoisier supposed that all acids must contain oxygen, and this idea was incorporated in the names used for this element in the various languages; the English oxygen, from the Greek *oxys* (sour) and *genna* (production); the German *Sauerstoff*, literally acid material; and the Russian *kislorod*, from *kislota* (acid). Following the discovery that hydrochloric acid contained no oxygen, the English chemist Sir Humphry Davy in about 1815 first recognized that the key element in acids was hydrogen. Not all substances that contain hydrogen, however, are acids, and the first really satisfactory definition of an acid was given by the German chemist Justus von Liebig in 1838. According to Liebig, an acid is a compound containing hydrogen in a form in which it can be replaced by a metal. This definition held the field for about 50 years and is still considered essentially correct, though somewhat outmoded. At the time of Liebig's proposal, bases were still regarded solely as substances that neutralized acids with the production of salts, and nothing was known about the constitutional features of bases that enabled them to do this.

The oxygen theory of acids

The whole subject of acid–base chemistry acquired a new look and a quantitative aspect with the advent of the electrolytic dissociation theory propounded by the German chemist Wilhelm Ostwald and the Swedish chemist Svante August Arrhenius (both Nobel laureates) in the 1880s. The principal feature of this theory is that certain compounds, called electrolytes, dissociate in solution to give ions—atoms bearing positive and negative charges resulting from the transfer of electrons from one atom to another of different kind. With the development of this theory it was realized that acids are merely those hydrogen compounds that give rise to hydrogen ions (H^+) in aqueous solution. It was also realized at that time that there is a correspondence between the degree of acidity of a solution (as shown by effects on vegetable dyes and other properties) and the concentration of hydrogen ions in the solution. Correspondingly, basic (or alkaline) properties could then be associated with the presence of hydroxide ions (OH^-) in aqueous solution, and the neutralization of acids by bases could be explained in terms of the reaction of these two ions to give the neutral molecule water ($H^+ + OH^- \rightarrow H_2O$.) This led naturally to the following simple definitions: Acids and bases are substances that give rise, respectively, to hydrogen and hydroxide ions in aqueous solution. This definition was generally accepted for the next 30 or 40 years. In purely qualitative terms, it does not offer many advantages over Liebig's definition of acids, but it does provide a satisfactory definition for bases.

Hydrogen and hydroxide ion definitions of acids and bases

Nevertheless, there is a great advantage in the definition of acids and bases in terms of hydrogen and hydroxide ions and this advantage lies in its quantitative aspects. Because the concentrations of hydrogen and hydroxide ions in solution can be measured, notably by determining the electrical conductivity of the solution (its ability to carry an electrical current), a quantitative measure of the acidity or alkalinity of the solution is provided. Moreover, the equations developed to express the relationships between the various components of reversible reactions can be applied to acid and base dissociations to give definite values, called dissociation constants. These constants can be used to characterize the relative strengths (degrees of dissociation) of acids and bases and, for this reason, supersede earlier semiquantitative estimates of acid or base strength. As a result of this approach, a satisfactory quantitative description was given at an early date of a large mass of experimental observations, a description that remains essentially unaffected by later developments in definitions of acid–base reactions.

The success of these quantitative developments, however, unfortunately helped to conceal some ambiguities and logical inconsistencies in the qualitative definitions of acids and bases in terms of the production of hydrogen and hydroxide ions, respectively. For example, it was not clear whether a substance like anhydrous hydrogen chloride, which would not conduct electricity, should be regarded as an acid, or whether it should be considered an acid only after it had come in contact with water. A modified definition of bases also seemed to be required that would be applicable to nonaqueous solutions, in which the anion (negatively charged ion) produced is not the hydroxide ion, as is the case in water, but varies from solvent to solvent, the methoxide ion (CH_3O^-) acting as the basic anion in methanol (CH_3OH), for example, and the amide ion (NH_2^-) playing the same role in liquid ammonia (NH_3). Even with acids the solvent is involved, since there is much evidence to show that the so-called hydrogen ion in solution does not exist as H^+, but always contains at least one molecule of solvent, as H_3O^+ in water, $CH_3OH_2^+$ in methanol, and NH_4^+ in liquid ammonia. These considerations led to the development of definitions of acids and bases that depended on the solvent (see below). In spite of this change, however, the difficulty still remained that typical acid–base properties, such as neutralization, indicator (vegetable-dye) effects, and catalysis, often took place in solvents such as benzene or chloroform in which free ions could barely be detected at all (by conductivity measurements). Even for aqueous solutions a particular ambiguity arises in the definition of bases, some of which (for example, metallic hydroxides) contain a hydroxyl group, whereas others (such as amines) do not. The latter produce hydroxide ions in solution by reacting with water molecules.

The Brønsted–Lowry acid–base definition. In order to resolve the various difficulties in the hydrogen–hydroxide ion definitions of acids and bases, a new, more generalized definition was proposed in 1923 almost simultaneously by the Danish chemist Johannes Nicolaus Brønsted and the English chemist Thomas Martin Lowry. Although the pursuit of exact verbal definitions of qualitative concepts is usually not profitable in physical science, the Brønsted–Lowry definition of acids and bases has had far-reaching consequences in the understanding of a wide range of phenomena and in the stimulation of much experimental work. The definition is as follows: an acid is a species having a tendency to lose a proton, and a base is a species having a tendency to gain a proton. The term proton means the species H^+ (the nucleus of the hydrogen atom) rather than the actual hydrogen ions that occur in various solutions; the definition is thus independent of the solvent. The use of the word species rather than substance or molecule implies that the terms acid and base are not restricted to uncharged molecules, but apply also to positively or negatively charged ions. This extension is one of the important features of the Brønsted–Lowry definition. It can be summarized by the equation $A \rightleftarrows B + H^+$ in which A and B together are a conjugate acid–base pair. In such a pair A must obviously have one more positive charge (or one less negative charge) than B, but there is no other restriction on the sign or magnitude of the charges. Several examples of conjugate acid–base pairs are given in Table 1.

Proton definition of acids and bases

Table 1: Examples of Conjugate Acid-Base Pairs

acid	base
Acetic acid, CH_3CO_2H	acetate ion, $CH_3CO_2^-$
Bisulfate ion, HSO_4^-	sulfate ion, SO_4^{2-}
Ammonium ion, NH_4^+	ammonia, NH_3
Ammonia, NH_3	amide ion, NH_2^-
Water, H_2O	hydroxide ion, OH^-
Hydronium (oxonium) ion, H_3O^+	water, H_2O

A number of points about the Brønsted–Lowry definition should be emphasized:

1. As mentioned above, this definition is independent of the solvent. The ions derived from the solvent (H_3O^+ and OH^- in water, and NH_4^+ and NH_2^- in liquid ammonia) are not accorded any special status, but appear as examples of acids or bases in terms of the general definition. On the other hand, of course, they will be particularly important species for reactions in the solvent to which they relate.

2. In addition to the familar molecular acids, two classes of ionic acids emerge from the new definition. The first comprises anions derived from acids containing more than one acidic hydrogen; *e.g.*, the bisulfate ion (HSO_4^-, as in Table 1) and primary and secondary phosphate ions ($H_2PO_4^-$ and HPO_4^{2-}) derived from phosphoric acid (H_3PO_4). The second and more interesting class consists of positively charged ions (cations), such as the ammonium ion (NH_4^+), which can be derived by the addition of a proton to a molecular base, in this case ammonia (NH_3). The hydronium ion (H_3O^+), which is the hydrogen ion in aqueous solution, also belongs to this class. The charge of these ionic acids, of course, always must be balanced by ions of opposite charges, but these oppositely charged ions usually are irrelevant to the acid–base properties of the system. For example, if sodium bisulfate ($Na^+HSO_4^-$) or ammonium chloride ($NH_4^+Cl^-$) is used as an acid, the sodium ion (Na^+) and the chloride ion (Cl^-) contribute nothing to the acidic properties and could equally well be replaced by other ions, such as potassium (K^+) and perchlorate (ClO_4^-), respectively.

3. Molecules such as ammonia and organic amines are bases by virtue of their tendency to accept a proton. With metallic hydroxides such as sodium hydroxide, on the other hand, the basic properties are due to the hydroxide ion itself, the sodium ion serving merely to preserve electrical

neutrality. Moreover, not only the hydroxide ion, but also the anions of other weak acids (for example the acetate ion) must be classed as bases because of their tendency to reform the acid by accepting a proton. Formally, the anion of any acid might be regarded as a base, but for the anion of a very strong acid (the chloride ion, for example) the tendency to accept a proton is so weak that its basic properties are insignificant and it is inappropriate to describe it as a base. Similarly, all hydrogen compounds could formally be defined as acids, but in many of them (for example, most hydrocarbons, such as methane, CH_4) the tendency to lose a proton is so small that the term acid would not normally be applied to them.

Amphoteric substance

4. Some species, including molecules as well as ions, possess both acidic and basic properties; such materials are said to be amphoteric. As shown in Table 1, both water and ammonia are amphoteric, a situation that can be represented by the schemes H_3O^+–H_2O–OH^- and NH_4^+–NH_3–NH_2^-. Another example is the secondary phosphate ion, HPO_4^{2-}, which can either lose or accept a proton, according to the following equations: $HPO_4^{2-} \rightleftarrows PO_4^{3-} + H^+$ and $HPO_4^{2-} + H^+ \rightleftarrows H_2PO_4^-$. The amphoteric properties of water are particularly important in determining its properties as a solvent for acid–base reactions.

5. The equation $A \rightleftarrows B + H^+$, used in the Brønsted–Lowry definition, does not represent a reaction that can be observed in practice, since the free proton, H^+, can be observed only in gaseous systems at low pressures. In solution, the proton always is attached to some other species, commonly a solvent molecule. Thus in water the ion H_3O^+ consists of a proton bound to a water molecule. For this reason all observable acid–base reactions in solution are combined in pairs, with the result that they are of the form $A_1 + B_2 \rightleftarrows B_1 + A_2$. The fact that the process $A \rightleftarrows B + H^+$ cannot be observed does not imply any serious inadequacy of the definition. A similar situation exists with the definitions of oxidizing and reducing agents, which are defined respectively as species having a tendency to gain or lose electrons, even though one of these reactions never occurs alone and free electrons are never detected in solution (any more than free protons are).

Alternative acid–base definitions. Although the Brønsted–Lowry concept of acids and bases as donors and acceptors of protons is still the most generally accepted one, other definitions are often encountered. Certain of these are adapted for special situations only, but the most important of these other definitions is in some respects more general than the Brønsted–Lowry definition. This definition was first proposed by the U.S. chemist Gilbert N. Lewis in 1923.

The Lewis definition of acids and bases

According to Lewis, an acid is a species that can accept an electron pair from a base with the formation of a chemical bond composed of a shared electron pair (covalent bond). This classification includes as bases the same species covered by the Brønsted–Lowry definition, since a molecule or ion that can accept a proton does so because it has one or more unshared pairs of electrons, and therefore it also can combine with electron acceptors other than the proton. On the other hand, the typical Lewis acids need not (and usually do not) contain protons, being species with outer electron shells that are capable of expansion, such as boron trifluoride (BF_3), sulfur trioxide (SO_3), and silver ion (Ag^+). Lewis originally based his ideas on the experimental fact that these nonprotonic acids often exhibit the properties regarded as typical of acids, such as neutralization of bases, action on indicators, and catalysis. Such substances often are electron acceptors, but this is not always the case; carbon dioxide (CO_2) and nitrogen pentoxide (N_2O_5), for example, contain completed octets of electrons and, according to usual valence theory, cannot accept any more. In addition, hydrogen-containing substances that have always been regarded as acids, acetic acid, for example, are not obviously electron acceptors, being rather adducts of the proton (a true Lewis acid) and a base such as the acetate ion. They can only be brought logically into the Lewis scheme by appealing to the fact that the reaction between a proton acid, which may be designated as XH, and a base, denoted by B, passes through an intermediate hydrogen-bonded state, X-H . . . B (in which the dotted line indicates a hydrogen-bond, a relatively weak secondary attractive force).

Numerous lengthy polemical exchanges have taken place regarding the relative merits of the Brønsted–Lowry and Lewis definitions of acids and bases. The difference is essentially one of nomenclature and has little scientific content. In the remainder of this article the term acid will be used to denote a proton donor (following the Brønsted–Lowry terminology) whereas the term Lewis acid will be employed exclusively to refer to electron-pair acceptors. This choice is based partly upon the logical difficulties mentioned in the last paragraph, and partly on the fact (see below) that the quantitative description of acid–base reactions is much simpler when it is confined to proton acids. This terminology also represents the commonest usage of the terms.

The definition of Lewis acids and bases in terms of the gain or loss of electrons should not be confused with the definition of oxidizing and reducing agents in similar terms. In oxidation–reduction reactions one or more electrons is transferred completely from the reducing agent to the oxidizing agent, whereas in a Lewis acid-base reaction an electron pair on the base is used to form a covalent link with the acid.

Certain other acid–base definitions have been based upon reactions occurring in specific solvent systems. For proton acids in amphoteric solvents these are equivalent to the Brønsted–Lowry definition. It is sometimes convenient to have general terms for the cation and anion derived from the solvent molecule by the addition and removal of a proton, respectively. The terms lyonium and lyate ions are occasionally used in this way. In water, the lyonium and lyate ions are H_3O^+ and OH^-; in ethanol, they are $C_2H_5OH_2^+$ and $C_2H_5O^-$; and in liquid ammonia, NH_4^+ and NH_2^-. For a given solvent, an acid can then be defined as a substance that increases the lyonium ion concentration (and correspondingly decreases the lyate ion concentration), whereas a base increases the lyate ion concentration (and decreases the lyonium ion concentration). This kind of definition, to be sure, really does not add anything to the concept of acids and bases as proton donors and proton acceptors.

Generalized solvent definitions

The idea that an acid is a solute that gives rise to cations characteristic of the solvent, and a base is a solute that gives rise to anions characteristic of the solvent, has sometimes been extended to solvents where no protons are involved at all, for example, liquid sulfur dioxide, SO_2. In this example, the solvent is supposed to ionize according to the equation $2SO_2 \rightleftarrows SO^{2+} + SO_3^{2-}$. Thionyl chloride, regarded as $SO^{2+} + 2Cl^-$, then can be considered an acid, and potassium sulfite, $2K^+ + SO_3^{2-}$, a base. The species SO^{2+} and SO_3^{2-} can certainly be regarded as Lewis acids and bases, but it is doubtful that they exist to any appreciable extent in liquid sulfur dioxide, a situation that makes the discussion somewhat artificial. Although this view of acids and bases has been useful in stimulating work in unusual types of solvent (for example in carbonyl chloride, selenium oxychloride, antimony trichloride, and hydrogen cyanide) it has not met with general acceptance.

ACID–BASE REACTIONS

Proton-transfer reactions. As already mentioned, the reaction expressed by the Brønsted–Lowry definition, $A \rightleftarrows B + H^+$, does not actually occur in any solution processes. This is because H^+, the bare proton, has an enormous tendency to add to almost all chemical species and cannot exist in any detectable concentrations except in a high vacuum. Apart from any specific chemical interaction, the very small size of the proton (about 10^{-15} metres) means that it exerts an extremely powerful electric field, which will polarize and therefore attract any molecule or ion it comes into contact with. It has been estimated that the dissociation of 19 grams of the hydronium ion H_3O^+ to give 1 gram of protons and 18 grams of water would require the expenditure of about 1,200,000 joules (290,000 calories) of energy, and thus it is an extremely unlikely process indeed.

Combined pairs of acid–base reactions

Typical acid–base reactions may be thought of as the combination of two reaction schemes $A_1 \rightleftarrows B_1 + H^+$ and $H^+ + B_2 \rightleftarrows A_2$, leading to the combined form: $A_1 + B_2 \rightleftarrows B_1 + A_2$. This represents a proton-transfer reaction from A_1 to B_2, producing B_1 and A_2. A large number of reactions in solution, often referred to under a variety of names, can be represented in this way. This is illustrated by the following examples, in each of which the species are written in the order A_1, B_2, B_1, A_2.

Dissociation of molecular acids in water. In this instance, water acts as a base. The equation for the dissociation of acetic acid, for example, is as follows: $CH_3CO_2H + H_2O \rightleftarrows CH_3CO_2^- + H_3O^+$.

Dissociation of bases in water. In this case, the water molecule acts as an acid and adds a proton to the base. An example, using ammonia as the base, is $H_2O + NH_3 \rightleftarrows OH^- + NH_4^+$. Older formulations would have written the left-hand side of the equation as ammonium hydroxide, NH_4OH, but it is not now believed that this species exists, except as a weak, hydrogen-bonded complex.

Dissociation of acids and bases in nonaqueous solvents. These situations are entirely analogous to the comparable reactions in water. For example, the dissociation of acetic acid in methanol may be written as $CH_3CO_2H + CH_3OH \rightleftarrows CH_3CO_2^- + CH_3OH_2^+$, and the dissociation of ammonia in the same solvent as $CH_3OH + NH_3 \rightleftarrows CH_3O^- + NH_4^+$.

Self-dissociation of amphoteric solvents. In this case, one solvent molecule acts as an acid, and the other as a base. Self-dissociation of water and liquid ammonia may be given as examples:

$$H_2O + H_2O \rightleftarrows OH^- + H_3O^+$$
$$NH_3 + NH_3 \rightleftarrows NH_2^- + NH_4^+.$$

Neutralization. For a strong acid and a strong base in water, the neutralization reaction is between hydrogen and hydroxide ions; *i.e.*, $H_3O^+ + OH^- \rightleftarrows H_2O$. For a weak acid and a weak base neutralization is more appropriately considered to involve direct proton transfer from the acid to the base. For example, the neutralization of acetic acid by ammonia may be written as $CH_3CO_2H + NH_3 \rightarrow CH_3CO_2^- + NH_4^+$. This equation does not involve the solvent; it therefore also represents the process of neutralization in an inert solvent, such as benzene, or in the complete absence of a solvent. (If one of the reactants is present in large excess, the reaction is more appropriately described as the dissociation of acetic acid in liquid ammonia, or of ammonia in glacial acetic acid.)

Acidic and basic salt solutions

Hydrolysis of salts. Many salts give aqueous solutions with acidic or basic properties. This is termed hydrolysis, and the explanation of hydrolysis reactions in classical acid–base terms was somewhat involved. In terms of the Brønsted–Lowry concept, however, hydrolysis appears to be a natural consequence of the acidic properties of cations derived from weak bases and the basic properties of anions derived from weak acids. For example, hydrolysis of aqueous solutions of ammonium chloride and of sodium acetate is represented by the following equations:

$$(Cl^-) + NH_4^+ + H_2O \rightleftarrows (Cl^-) + NH_3 + H_3O^+$$
$$(Na^+) + H_2O + CH_3CO_2^- \rightleftarrows (Na^+) + OH^- + CH_3CO_2H$$

The sodium and chloride ions take no part in the reaction and could equally well be omitted from the equations.

The acidity of the solution represented by the first equation is due to the presence of the hydronium ion (H_3O^+) and the basicity of the second comes from the hydroxide ion (OH^-). The reverse reactions simply represent, respectively, the neutralization of aqueous ammonia by a strong acid and of aqueous acetic acid by a strong base.

A superficially different type of hydrolysis occurs in aqueous solutions of salts of some metals, especially those giving multiply charged cations. For example, aluminum, ferric, and chromic salts all give aqueous solutions that are acidic. This behaviour also can be interpreted in terms of proton-transfer reactions if it is remembered that the ions involved are strongly hydrated in solution. In a solution of an aluminum salt, for instance, a proton is transferred from one of the water molecules in the hydration shell to a molecule of solvent water. The resulting hydronium ion (H_3O^+) accounts for the acidity of the solution:

$$Al(H_2O)_6^{3+} + H_2O \rightleftarrows Al(H_2O)_5OH^{2+} + H_3O^+.$$

Reactions of Lewis acids. In the reaction of a Lewis acid with a base the essential process is the formation of an adduct in which the two species are joined by a covalent bond; proton transfers are not normally involved. If both the Lewis acid and base are uncharged, the resulting bond is termed semipolar or coordinate, as in the reaction of boron trifluoride with ammonia.

Formation of adducts

$$\underset{\text{acid}}{BF_3} + \underset{\text{base}}{NH_3} \rightarrow \underset{\text{adduct.}}{F_3\bar{B}\text{–}\overset{+}{N}H_3}$$

Frequently, however, either or both species bears a charge (most commonly a positive charge on the acid or a negative charge on the base), and the location of charges within the adduct often depends upon the theoretical interpretation of the valences involved. Examples of such reactions are:

$$\underset{\text{silver ion}}{Ag^+} + \underset{\text{ammonia}}{2NH_3} \rightarrow \underset{\text{adduct}}{[Ag(NH_3)_2]^+}$$
$$\underset{\text{boron trifluoride}}{BF_3} + \underset{\text{hydroxide ion}}{OH^-} \rightarrow \underset{\text{adduct}}{[BF_3(OH)]^-}$$
$$\underset{\text{cadmium ion}}{Cd^{2+}} + \underset{\text{cyanide ion}}{4CN^-} \rightarrow \underset{\text{adduct.}}{[Cd(CN)_4]^{2-}}$$

In another common type of process, one acid, or base, in an adduct is replaced by another:

$$\underset{\text{adduct (1)}}{Cl_4\bar{S}n\text{–}\overset{+}{N}H_3} + \underset{\text{acid (2)}}{BF_3} \rightarrow \underset{\text{adduct (2)}}{F_3\bar{B}\text{–}\overset{+}{N}H_3} + \underset{\text{acid (1)}}{SnCl_4}$$
$$\underset{\text{adduct (1)}}{Al(H_2O)_6^{3+}} + \underset{\text{base (2)}}{6F^-} \rightarrow \underset{\text{adduct (2)}}{AlF_6^{3-}} + \underset{\text{base (1).}}{6H_2O}$$

In fact, reactions such as the simple adduct formations above often are formulated more correctly as replacements. For example, if the reaction of boron trifluoride with ammonia is carried out in ether as a solvent, it becomes a replacement reaction:

$$(C_2H_5)_2\overset{+}{O}\text{–}\bar{B}F_3 + NH_3 \rightarrow F_3\bar{B}\text{–}\overset{+}{N}H_3 + (C_2H_5)_2O.$$

Similarly, the reaction of silver ions with ammonia in aqueous solution is better written as a replacement reaction:

$$Ag(H_2O)_2^+ + 2NH_3 \rightarrow Ag(NH_3)_2^+ + 2H_2O.$$

Furthermore, if most covalent molecules are regarded as adducts of (often hypothetical) Lewis acids and bases, an enormous number of reactions can be formulated in the same way. To take a single example, the reaction of methyl chloride with hydroxide ion to give methanol and chloride ion (usually written as $CH_3Cl + OH^- \rightarrow CH_3OH + Cl^-$) can be reformulated as replacement of a base in a Lewis acid–base adduct, as follows: (adduct of CH_3^+ and Cl^-) + $OH^- \rightarrow$ (adduct of CH_3^+ and OH^-) + Cl^-. Opinions differ as to the usefulness of this extremely generalized extension of the Lewis acid–base-adduct concept.

Reactions of anhydrous oxides

The reactions of anhydrous oxides (usually solid or molten) to give salts may be regarded as examples of Lewis acid–base-adduct formation. For example, in the reaction of calcium oxide with silica to give calcium silicate, the calcium ions play no essential part in the process, which may be considered therefore to be adduct formation between silica as the acid and oxide ion as the base:

$$\underset{\text{acid}}{SiO_2} + \underset{\text{base}}{O^{2-}} \rightarrow \underset{\text{adduct.}}{SiO_3^{2-}}$$

A great deal of the chemistry of molten-oxide systems can be represented in this way, or in terms of the replacement of one acid by another in an adduct.

Acid–base catalysis. Acids (including Lewis acids) and bases act as powerful catalysts for a great variety of chemical reactions, in the laboratory, in industry, and in processes occurring in nature. Historically, catalytic action was regarded as one of the essential characteristics of acids, and the parallel occurrence of catalytic action and electrical conductivity was one of the compelling pieces of evidence in establishing the theory of electrolytic dissociation as the basis of acid–base behaviour at the end of the 19th century.

Acid–base catalysis was originally thought of in terms of

a mysterious influence of the acid or base, but it is now generally believed to involve an actual acid–base reaction between the catalyst and the reacting substance, termed the substrate, with the catalyst being regenerated at a later stage of the reaction. Moreover, knowledge of reaction mechanisms is now sufficient to suggest detailed sequences of reactions for many acid- or base-catalysis reactions, most of which sequences are at least plausible and in many instances are well established.

In most acid–base reactions the addition or removal of a proton does not bring about any drastic change in the structure of the molecule, or in its stability or reactivity. It is a characteristic of reactions catalyzed by acids or bases, however, that the addition or removal of a proton either makes the substrate unstable, so that it decomposes or rearranges, or that it causes the substrate to become reactive toward some other species present in the system. In cases of rearrangement, the regeneration of the catalyst often involves the removal or addition of a proton at a site other than that at which the initial addition or removal took place. It is not necessary that the substrate in an acid- or base-catalyzed reaction should itself have marked acid–base properties, since even a very small extent of initial acid–base reaction may be enough to bring about the subsequent change.

Regeneration of the catalyst

Instances of acid–base catalysis are numerous indeed; a few examples are given here, as follows:

Isomerization of olefins, acid-catalyzed. Unsaturated compounds frequently rearrange reversibly under the influence of acids to give products in which the double bond occurs in a new location. The interconversion of 2-butene and 1-butene is shown here:

$$CH_3CH=CHCH_3 + H_3O^+ \rightleftarrows CH_3CH_2\overset{+}{C}HCH_3 + H_2O$$
$$CH_3CH_2\overset{+}{C}HCH_3 + H_2O \rightleftarrows CH_3CH_2CH=CH_2 + H_3O^+.$$

Reversible dehydration of alcohols, acid-catalyzed. Under the influence of acids, alcohols generally undergo loss of water to give olefinic products. The dehydration of ethanol to ethylene occurs as follows:

$$CH_3CH_2OH + H_3O^+ \rightleftarrows CH_3CH_2OH_2^+ + H_2O$$
$$CH_3CH_2OH_2^+ \rightleftarrows CH_3CH_2^+ + H_2O \rightleftarrows$$
$$CH_2=CH_2 + H_3O^+.$$

Keto–enol tautomerism, acid- and base-catalyzed. Acids and bases both bring about the establishment of an equilibrium between ketones (or aldehydes) and their enol forms, which consist of a hydroxyl group directly attached to a doubly bonded carbon atom. (–C(OH)=C–). The interconversion between the two forms is called keto–enol tautomerism. The reaction cannot always be observed directly, since the enol form may not reach measurable concentrations, even at equilibrium, but the highly active enol may be detected by its reaction with various reagents, notably the halogens (bromine, for example). Keto–enol tautomerism of acetone can be brought about by acid or base catalysis, as follows:

$$CH_3\underset{\underset{\text{keto form}}{\|}}{\overset{}{C}}CH_3 + H_3O^+ \rightleftarrows CH_3\overset{+OH}{\overset{|}{C}}CH_3 + H_2O \rightleftarrows$$
$$CH_3C(OH)=CH_2 + H_3O^+ \quad \text{(enol form)}$$
$$CH_3\underset{\underset{\text{keto form}}{\|}}{\overset{}{C}}CH_3 + OH^- \rightleftarrows CH_3\overset{O^-}{\overset{|}{C}}=CH_2 + H_2O \rightleftarrows$$
$$CH_3C(OH)=CH_2 + OH^- \quad \text{(enol form)}$$

Aldol condensation, base-catalyzed. Self-condensation of aldehydes, the so-called aldol condensation, occurs readily, when catalyzed by bases, to give β-hydroxyaldehydes. The prototype of this reaction is the conversion of acetaldehyde to β-hydroxybutyraldehyde, or aldol. The first step of this reaction is the production of an enolate ion (as in formation of the keto–enol tautomeric mixture), but this anion then reacts with a second molecule of acetaldehyde to give the product as shown below:

$$CH_3CHO + OH^- \rightleftarrows [CH_2CHO]^- + H_2O$$
$$CH_3CHO + [CH_2CHO]^- \rightleftarrows CH_3CH(O^-)CH_2CHO$$
$$CH_3CH(O^-)CH_2CHO + H_2O \rightleftarrows \underset{\text{aldol}}{CH_3CHOHCH_2CHO} + OH^-$$

These examples illustrate the importance of acid–base catalysis in organic reactions. The equations have been written in terms of H_3O^+ and OH^- as the acid and base catalysts, respectively, and these are certainly the most important catalysts in aqueous solution. For many of these reactions (especially isomerization of olefins and dehydration of alcohols), there is ample evidence that other acids or bases also can act as catalysts. This behaviour is known as general acid–base catalysis. It appears particularly clearly in inert solvents such as benzene, in which catalysis by molecular acids and bases is frequently observed despite the absence of detectable quantities of ions derived from the solvent. Acidic groups, such as sulfonic acid ($-SO_3H$) and carboxylic acid groups ($-CO_2H$), attached to a solid molecular framework (as in some ion exchange resins) also act as heterogeneous catalysts for many chemical reactions.

Importance of acid–base catalysis

The above examples show that proton-transfer processes can play a specific part in reaction mechanisms and, in these and similar instances, it is doubtful whether any uncatalyzed or spontaneous reaction of the same type can take place. Apparent evidence to the contrary can usually be explained by catalysis by solvent molecules or by adventitious acidic or basic impurities.

Lewis acids can exert a catalytic effect in two different ways. In the first of these they interact with hydrogen-containing compounds present in the system to assist the release of a proton to the substrate. For example, the polymerization of olefins by Lewis acids, such as boron trifluoride (BF_3), aluminum chloride ($AlCl_3$), and titanium tetrachloride ($TiCl_4$), is believed to be caused by their interaction with proton acids (for example, traces of water) and the olefin to give a carbonium ion, which then reacts further with more olefin:

Lewis-acid catalysis

$$BF_3 + H_2O + CHR=CHR' \rightarrow CH_2R\text{–}\overset{+}{C}HR' + [BF_3OH]^-$$
$$CH_2R\text{–}\overset{+}{C}HR' + CHR=CHR' \rightarrow CH_2RCHR'CHR\overset{+}{C}HR', \text{ etc.}$$

In the second mode of action, the Lewis acid acts directly on the substrate, and by withdrawing electrons converts it into a reactive form. A typical example is the action of catalysts like aluminum chloride on alkyl halides to produce carbonium ions: $RCl + AlCl_3 \rightarrow R^+ + [AlCl_4]^-$. The carbonium ion can then react further with other substances, for example, aromatic hydrocarbons. The same type of catalysis probably occurs with many solid oxide catalysts (for example, aluminosilicates), although it is often difficult to decide whether the catalytic action of these materials is due to centres with a deficiency of electrons or to acidic hydroxyl groups.

QUANTITATIVE ASPECTS OF ACID–BASE EQUILIBRIA

General considerations. Certain general principles apply to any solvent with both acidic and basic properties, for example, water, alcohols, ammonia, amines, and acetic acid. Denoting the solvent molecule by SH, it can give rise to the ions SH_2^+ and S^-, sometimes called lyonium and lyate ions, respectively (see above). In the pure solvent these are the only ions present, and they must be present in equal concentrations to preserve electrical neutrality. The equilibrium involved, therefore, is as follows: $2SH \rightleftarrows SH_2^+ + S^-$. The equilibrium constant (K_s') for this reaction (the mathematical quantity that expresses the relationships between the concentrations of the various species present at equilibrium) would normally be given by the equation $K_s' = [SH_2^+][S^-]/[SH]^2$, in which the square brackets denote the concentrations of the species within the brackets. In a given solvent, however, the concentration of the

Equilibrium constants

solvent, [SH], is a large and constant quantity, and it is therefore usual to eliminate this term and express the self-dissociation of the solvent by the equation $K_s = [SH_2^+]\,[S^-]$. In this equation, K_s is termed the ion product or the autoprotolysis constant of the solvent. The concentrations are usually expressed in moles per litre, a mole being the molecular weight of the compound in grams. Since a solvent that is a good proton donor is normally a poor proton acceptor, and vice versa, the degree of ionization is generally low and K_s is usually a small quantity. It is about 10^{-14} for water at ordinary temperatures, and one of the largest K_s values known is 1.7×10^{-4} for 100 percent sulfuric acid. The above equation applies not only to the pure solvent, but also (with the same value of K_s) to any dilute solutions of acids, bases, or salts in the solvent in question. In these solutions $[SH_2^+]$ and $[S^-]$ need not be equal, since the condition of electrical neutrality involves the concentration of other ions as well, and it is obvious from the equation that a high value of $[SH_2^+]$ must imply a low value of $[S^-]$, and vice versa.

If an acid A is added to the solvent SH it will be at least partly converted into the conjugate base B according to the reaction $A + SH \rightleftarrows B + SH_2^+$, which would be characterized formally by an equilibrium constant $[B]\,[SH_2^+]/[A]\,[SH]$. Again, however, it is usual to omit the term for the constant concentration of the solvent, [SH], from this expression, and to define a constant K_a by the equation

$$\frac{[B]\,[SH_2^+]}{[A]} = K_a.$$

Dissociation constants

K_a is known as the dissociation constant of the acid A in the solvent SH. Any acid–base reaction $A_1 + B_2 \rightleftarrows B_1 + A_2$ will proceed from left to right almost completely if A_1 is a much stronger acid than A_2. It is a natural extension of this idea to use the equilibrium constant as a measure of the strength of A_1 relative to A_2. The dissociation constant is thus (apart from the constant factor [SH], which has been omitted) a measure of the acid strength of A relative to that of the lyonium ion SH_2^+.

In some instances reaction goes so completely from left to right that it is not possible to measure the equilibrium constant. A is said then to be a strong acid in the solvent SH; similarly, acids with readily measurable dissociation constants (in practice less than about 0.1) are known as weak acids.

Similar considerations apply to solutions of bases. The reaction involved in this case is $SH + B \rightleftarrows S^- + A$ and the equilibrium constant K_b defined by

$$\frac{[S^-]\,[A]}{[B]} = K_b$$

is termed the dissociation constant of the base B. Apart from the omitted constant factor [SH], K_b represents the basic strength of B relative to that of the lyate ion S^-. Bases are termed strong and weak in the same way that acids are.

The values of K_a and K_b for a conjugate acid–base pair A–B in a given solvent are not independent, since consideration of the dissociation constants of the solvent, acid, and base show that $K_aK_b = [SH_2^+]\,[S^-] = K_s$ in which K_s is the ion product of the solvent. It is therefore unnecessary to specify both K_a and K_b, and it has become common practice to characterize an acid–base pair by K_a only, which may be termed the acidity constant of A–B in the solvent SH. If the value of K_b is required it is readily obtained from K_a and K_s. Since readily accessible values of K_a are always much less than unity, it is often convenient to introduce a quantity pK_a, sometimes called the acidity exponent, and

Acidity exponent

defined by the relation $pK_a = -\log_{10} K_a$. Values of pK_a are generally of a more convenient magnitude.

The above expressions for the various equilibrium constants depend only on the concentrations of the species concerned, which are tacitly assumed to exist in solution independently of one another. This is not always the case, and in exact treatments of these equilibria two modifications are frequently necessary. In the first place, some or all of the reacting species are ions and, because of the electrical forces between them, the law expressing their concentrations at equilibrium is not always valid. Corrections may be applied by multiplying the concentrations by certain factors called activity coefficients, the values of which can be calculated theoretically or derived from other measurements. Furthermore, ions of opposite charge may attract one another so strongly that they no longer exist independently, but are partly present as ion pairs, thus altering the forms of the equilibrium equations. For many purposes, however, the simple equations given here are adequate, especially with regard to reactions in aqueous solutions.

Acid–base equilibria in aqueous solutions. Since aqueous solutions are of particular importance in the laboratory and in the physiology of animals and plants, it is appropriate to consider them separately. The ion product of water,

Ion product of water

$K_w = [H_3O^+]\,[OH^-]$, has the value 1.0×10^{-14} mole^2 litre^{-2} at 25° C, but it is strongly temperature dependent, becoming 1.0×10^{-15} at 0° C and 7×10^{-13} at 100° C. In principle the value of K_w can be determined by measuring the electrical conductance of very pure water, in which $[H_3O^+] = [OH^-] = 10^{-7}$ at 25° C, but in practice it is derived from other measurements, for example, measurements of the degree of hydrolysis of salts.

For an uncharged acid, such as acetic, the dissociation constant is given by the following expression:

$$\frac{[CH_3CO_2^-]\,[H_3O^+]}{[CH_3CO_2H]} = K_a.$$

For this example (acetic acid) K_a has the value 1.76×10^{-5} at 25° C. It is often convenient to express the dissociation constant in terms of the degree of dissociation of the acid. This quantity, which is represented in equations by the Greek letter alpha, α, is equal to the fraction of the acid that appears in dissociated form—in this case as the ions $CH_3CO_2^-$ and H_3O^+. If the initial concentration of acid is designated by c, then the concentrations of the ions are each equal to αc, that is $[H_3O^+] = [CH_3CO_2^-] = \alpha c$, and the concentration of undissociated acid is equal to $c(1 - \alpha)$, or $[CH_3CO_2H] = c(1 - \alpha)$.

Substituting these expressions into the equation giving the value of the dissociation constant gives the following result: $\frac{\alpha^2 c}{1 - \alpha} = K_a$. From this equation it can be inferred that the degree of dissociation (α) increases with decreasing concentration (c). For small degrees of dissociation ($\alpha << 1$) the equation becomes $\alpha = \frac{K_a^{\frac{1}{2}}}{c^{\frac{1}{2}}}$ whereas at sufficiently low concentrations ($c << 1$) α tends to unity ($\alpha \rightarrow 1$).

Discussions exactly analogous to this apply to a number of other acid–base equilibria, for example, (1) the dissociation of ammonia in water, (2) the hydrolysis of ammonium salts, and (3) the hydrolysis of an acetate.

(1) $H_2O + NH_3 \rightleftarrows OH^- + NH_4^+$
(2) $NH_4^+ + H_2O \rightleftarrows NH_3 + H_3O^+$
(3) $H_2O + CH_3CO_2^- \rightleftarrows OH^- + CH_3CO_2H$

For reaction (1) α is the degree of dissociation of ammonia, and the dissociation constant is K_b, the basic dissociation constant. In reaction (2), the hydrolysis of an ammonium salt (for example, ammonium chloride), α would be termed the degree of hydrolysis and K the hydrolysis constant. In terms of the general definition of acids and bases, however, K could equally be called the acidity constant for the acid–base pair NH_4^+–NH_3, and this is a more rational way of describing the process. Finally, reaction (3) represents the hydrolysis of an acetate (for example, sodium acetate); the resulting equilibrium constant is termed the hydrolysis constant and can be seen to equal K_w/K_a, where K_w is the ion product of water and K_a the acidity constant for the acid–base pair CH_3CO_2H – $CH_3CO_2^-$; *i.e.*, the dissociation constant of acetic acid. The investigation of equilibria such as this is, in fact, one of the methods for determining the value of K_w (see above).

Dissociations in water

The equilibria considered so far arise when one component of an acid–base pair is dissolved in water—if necessary, along with an ion, such as Na^+ or Cl^-, having negligible acid–base properties. The direct consequence of this is that the two new species produced (for example, those on the right-hand sides of the equations [1–3] above) have equal

concentrations (αc), and hence the previously given equation $K = \frac{\alpha^2 c}{1 - \alpha}$ is applicable.

Buffered solutions

A solution of a more generally useful type can be obtained by deliberately varying the proportions of acid and base present; such a solution is called a buffered solution or, somewhat more colloquially, a buffer. A buffered solution containing various concentrations of acetic acid and acetate ion, for example, can be prepared by mixing solutions of acetic acid and sodium acetate, by partially neutralizing a solution of acetic acid with sodium hydroxide, or by adding less than one equivalent of a strong acid to a solution of sodium acetate. Similarly, a buffer based on the pair $NH_4^+ - NH_3$ can be prepared by mixing solutions of ammonia and an ammonium salt, by partially neutralizing a solution of ammonia with a strong acid, or by adding less than one equivalent of sodium hydroxide to a solution of an ammonium salt. The hydrogen ion concentration in a buffer solution is, of course, still given by the usual equation, which is conveniently written as

$$[H_3O^+] = K_a \frac{[A]}{[B]}.$$

Hydrogen-ion concentration and *p*H

Since hydrogen ion concentrations are usually less than unity, and cover an extremely wide range, it is often convenient to use instead the negative logarithm of the actual concentration, a figure that varies usually only in the range 1–13. This figure is termed the *p*H, and its definition is expressed by the equation $pH = -\log_{10}[H_3O^+]$. For example, in pure water $[H_3O^+] = 1 \times 10^{-7}$, with the result that the *p*H = 7.0. The same term can be applied to alkaline solutions; thus in 0.1 molar sodium hydroxide $[OH^-] = 0.1$, $[H_3O^+] = K_w/[OH^-] = 1 \times 10^{-14}/0.1 = 10^{-13}$, *p*H = 13.0.

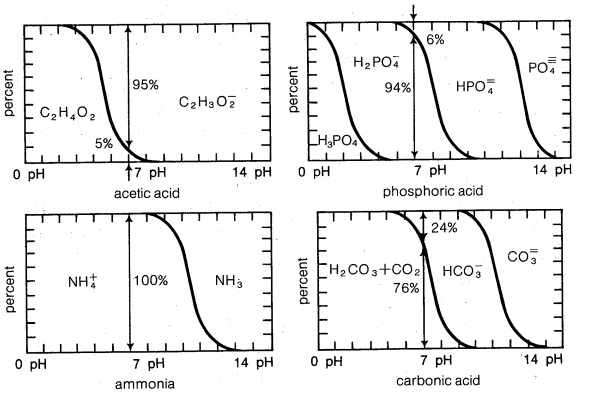
Applying the concept of *p*H to buffered solutions gives the following equation: $pH = pK_a - \log_{10}\frac{[A]}{[B]}$ in which $\frac{[A]}{[B]}$, known as the buffer ratio, can be calculated from the way in which the solution is prepared. According to this equation, the *p*H of the buffered solution depends only on the pK_a of the acid and on the buffer ratio. Most particularly it does not depend on the actual concentrations of A and B. Therefore, the *p*H of a buffered solution is little affected by dilution of the solution. It is also insensitive to the addition of acid or alkali, provided that the amounts added are much smaller than both [A] and [B]. This so-called buffering action will be impaired if either [A] or [B] becomes too small; hence, buffer ratios must not deviate too far from unity, and the effective buffering range of a given acid–base system is roughly from $pH = pK_a + 1$ to $pH = pK_a - 1$, corresponding to buffer ratios from 0.1 to 10.

Buffer action

The figure shows the relation between *p*H and composition for a number of commonly used buffer systems. Effective buffer action is confined to the central, steep portion of each curve, where the *p*H is least sensitive to the composition. The figure shows that an acid bearing several acidic hydrogens, such as phosphoric acid, can be used to prepare buffer solutions in several different *p*H ranges. Buffer action plays an important part in controlling the *p*H of many biological fluids; for example, the *p*H of the blood is controlled at about 7.4 by the carbonic acid–bicarbonate system shown in the figure. Buffers are widely used to control the *p*H in chemical or biological experiments. For the latter, the system $H_2PO_4^- - HPO_4^{2-}$ is particularly useful, being effective in the physiological *p*H range, 6–8.

The same principles can be applied for the quantitative treatment of systems containing larger numbers of acid–base pairs; for example, in an aqueous solution of ammonium acetate, the following acid–base pairs must be considered: $NH_4^+ - NH_3$, $CH_3CO_2H - CH_3CO_2^-$, $H_3O - H_2O$, and $H_2O - OH^-$. The situation is much more complicated in many solutions that are important in industry or in nature, but it is always possible to make a complete prediction of the state of the system in terms of the acidity constants K_a of each acid–base pair (provided, of course, that reactions other than proton transfers do not interfere).

Acid–base equilibria in non-aqueous solvents. Although acid–base properties have been investigated most thoroughly in aqueous solutions, partly because of their practical importance, water is in many respects an abnormal solvent. In particular, it has a higher dielectric constant (a measure of the ability of the medium to reduce the force between two electric charges) than most other liquids, and it is able itself to act either as an acid or as a base. The behaviour of acids and bases in several other solvents will be described briefly here.



Relation between *p*H and composition for a number of commonly used buffer systems.
From N. J. Bjerrum, *Inorganic Chemistry* (1936); Heinemann Educational Books Ltd.

Effects of solvent on dissociation

The effect of the solvent on the dissociation of acids or bases depends largely upon the basic or acidic properties of the solvent, respectively. Since many acid–base reactions involve an increase or decrease in the number of ions, they are also influenced by the dielectric constant of the solvent, for a higher dielectric constant favours the formation of ions. Finally, the specific solvation (or close association with the solvent) of particular ions (excluding the solvation of the proton to give SH_2^+, which is already included in the basicity of the solvent) may be important. It is usually not easy to separate these three effects and, in particular, the effects of dielectric constant and solvation merge into one another. These points are illustrated with examples of several of the more important solvents. In this discussion the solvents are classified as amphoteric (both acidic and basic), acidic (in which the acidic properties are much more prominent than the basic), basic (in which the reverse is true), and aprotic (in which both acidic and basic properties are almost entirely absent). Finally, concentrated aqueous acids are mentioned as an example, and a particularly important one, of mixed solvents.

Amphoteric solvents. The most important nonaqueous solvents of this class are the lower alcohols methanol and ethanol. They resemble water in their acid–base properties but, because of their lower dielectric constants, facilitate processes producing ions to a much smaller extent. In particular, the ion products of these solvents are much smaller ($K_s = 10^{-17}$ for CH_3OH and 10^{-19} for C_2H_5OH, compared with 10^{-14} for water), and the dissociation constants of molecular acids and bases are uniformly lower than in water by four to five powers of ten. Nitric acid, for example, which is almost completely dissociated in water (K_a about 20), has $K_a = 2.5 \times 10^{-4}$ in methanol. On the other hand, the equilibrium constants of processes such as $NH_4^+ + ROH \rightleftarrows NH_3 + ROH_2^+$ and $CH_3CO_2^- + ROH \rightleftarrows CH_3CO_2H + RO^-$ are similar in all three solvents, since they do not involve any change in the number of ions.

Sulfuric acid as solvent

Acidic solvents. The most important strongly acidic solvent is sulfuric acid, which is able to protonate a wide variety of compounds containing oxygen or nitrogen. Thus, water, alcohols, ethers, ketones, nitro compounds, and sulfones all act as bases in sulfuric acid. This solvent must also possess some basic properties, because its ionic product is high ($[H_3SO_4^+]\,[HSO_4^-] = 1.7 \times 10^{-4}$), but the basicity of the solvent is obscured normally by its very high acidity. For example, carboxylic acids behave as strong bases in sulfuric acid, reacting almost completely according to the equation $RCO_2H + H_2SO_4 \rightarrow RCO_2H_2^+ + HSO_4^-$. Many substances undergo reactions in sulfuric acid that are more complicated than simple proton transfers, often

yielding species important because of their chemical reactivity. Thus, some alcohols produce carbonium ions in sulfuric acid; with triphenylcarbinol, for example, the reaction is $(C_6H_5)_3COH + 2H_2SO_4 \rightarrow (C_6H_5)_3C^+ + H_3O^+ + 2HSO_4^-$. Nitric acid gives the nitronium ion, NO_2^+, according to the equation ($HNO_3 + 2H_2SO_4 \rightarrow NO_2^+ + H_3O^+ + 2HSO_4^-$). This ion frequently is the active agent in the nitration of organic compounds. Hydrogen fluoride has solvent properties resembling those of sulfuric acid, but is less acidic and has negligible basic properties.

Acetic acid is another acidic solvent that has been extensively studied. Because of its low dielectric constant, ions exist in it largely in the form of ion pairs, and more complex associates are frequently formed. For this reason a quantitative interpretation of acid–base equilibria in acetic acid is often difficult, but some general conclusions can be drawn. In particular, it can be seen that all substances more basic in water solution than aniline react completely with acetic acid according to the equation $B + CH_3CO_2H \rightarrow BH^+ + CH_3CO_2^-$. All such bases therefore give solutions with indistinguishable acid–base properties; this is often referred to as a levelling effect of the solvent. The converse is true for acids; for example, the strong mineral acids, nitric, hydrochloric, sulfuric, hydrobromic, and perchloric (HNO_3, HCl, H_2SO_4, HBr, and $HClO_4$) are "levelled" in aqueous solution by complete conversion to the hydronium ion, but in acetic acid they are differentiated as weak acids with strengths in the approximate ratio 1:9:30:160:400.

Basic solvents. The only basic solvent that has been investigated in any detail is liquid ammonia, which has the very low ion product $[NH_4^+]\,[NH_2^-] = 10^{-33}$. As might be expected, this solvent has a marked levelling effect upon acids; thus, for example, acetic, benzoic, nitric, and hydrochloric acids all give solutions with identical acidic properties, due to the ion NH_4^+—although, of course, in water they behave very differently.

Aprotic solvents. Strictly aprotic solvents include the hydrocarbons and their halogen derivatives, which undergo no reaction with added acids or bases. Acid–base equilibrium in these solvents can be investigated only when a second acid–base system is added; the usual reaction $A_1 + B_2 \rightleftarrows B_1 + A_2$ then takes place. Most such investigations have employed an indicator as one of the reacting systems, but the results are often difficult to interpret because of association of both ions and molecules in these media of low dielectric constant.

The term aprotic has been extended recently to include solvents that are unable to lose a proton, although they may have weakly basic properties. Some of these aprotic solvents have high dielectric constants (for example, N, N-dimethylformamide, dimethyl sulfoxide, and nitrobenzene) and are good solvents for a variety of substances. They have a powerful differentiating effect on the properties of acids and bases. In particular, basic anions are poorly solvated in these solvents, and thus behave as very strong bases; for example, it has been estimated that sodium methoxide dissolved in dimethyl sulfoxide gives a solution 10^9 times as basic as in methanol.

Concentrated aqueous acids. Dilute solutions of strong acids, for example, hydrochloric, sulfuric, and perchloric (HCl, H_2SO_4, $HClO_4$), in water behave essentially as solutions of the ion H_3O^+, and their acidity increases in proportion to their concentration. At concentrations greater than about one molar (that is, containing one mole of acid per litre of solution), however, the acidity, as measured by action on indicators or by catalytic ability, increases much more rapidly than the concentration. For example, a 10 molar solution of any strong acid is about 1,000 times as acidic as a 1 molar solution. This behaviour is undoubtedly largely due to the depletion of water with increasing concentration of acid; the hydronium ion, H_3O^+, is known to have a strong tendency to further hydration, probably mainly to the ion $H_3O^+(H_2O)_3$ (*i.e.*, $H_9O_4^+$), and a decrease in water content increases the proton-donating power of the solution. The acidity of these concentrated solutions is commonly measured by the acidity function, H_o, a quantity measured by the effect of the solvent on a basic indicator I. It is defined by $H_o + pK_{IH^+} - \log_{10} [IH^+]/[I]$ and becomes equal to the *p*H in dilute solution. The acidity function H_o is often found to be independent of the nature of the indicator and to give an approximate measure of the catalytic power of the acid solution. Sulfuric acid–water mixtures ranging from 10 to 100 percent sulfuric acid have H_o values between -0.3 and -11.1, corresponding to an acidity range of nearly 11 powers of ten.

Acidity function

Acid–base equilibria involving Lewis acids. Much less information is available about Lewis acid–base equilibria than about ordinary acid–base equilibria, but it is clear that the situation is less simple for the former than for the latter. When a given Lewis acid reacts with a series of similarly constituted bases the equilibrium constants often vary in parallel with the conventional basic strengths. This is the case when a zinc halide, ZnX_2, for example, reacts with a series of amines. In general, however, it is not possible to arrange Lewis acids and bases in a unique order that will predict the extent to which a given pair will react. Thus, although the hydroxide ion (OH^-) is always a much stronger base than ammonia (NH_3) in reactions with proton acids, in reactions with the Lewis acid Ag^+, the complex $Ag(NH_3)_2^+$ is fairly stable, whereas AgOH is completely dissociated. Similarly, for some metal cations complex formation increases in the order fluoride < chloride < bromide < iodide, whereas for other metal cations the order is the reverse of this.

This kind of behaviour has led to a classification of Lewis acids and bases into "hard" and "soft" categories; as a rule, hard acids react preferentially with hard bases and, similarly, soft acids react with soft bases. The terms hard and soft are chosen to suggest that the atomic structures associated with hard acids and bases are rigid and impenetrable, whereas those associated with soft acids and bases are more readily deformable. Hard acids include: the proton; sodium, calcium, and, aluminum ions; and carbonium ions. The soft acids include: cuprous, silver, mercurous, and the halogen cations. Typical soft bases are iodide, thiocyanate, sulfide, and triphenylphosphine; whereas hard bases include hydroxide, fluoride, and many oxyanions. The dividing line between the hard and soft categories is not a sharp one, and its theoretical interpretation is obscure. Nevertheless, a surprising amount of factual information can be coordinated on the basis of preferential reactions of hard acids with hard bases and soft acids with soft bases.

Hard and soft acids and bases

The effect of molecular structure on acid–base equilibria. Regularities and trends in the properties of the elements are best understood in terms of the periodic table, an orderly pattern seen when the elements are arranged in order of increasing atomic number. Comparing the hydrides of the various elements in the table reveals appreciable acidity only among those of the elements on the right-hand side of the table, especially the halogen elements—fluorine, chlorine, bromine, and iodine. This generalization is borne out when the elements across the first period of the table are examined in order; as the right-hand side of the table is approached the elements encountered are carbon, nitrogen, oxygen, and fluorine. The hydrides of these elements show increasing acidity. Methane, CH_4, a hydride of carbon, has no detectable acidic properties, and the pK_a decreases sharply in the series ammonia (NH_3) 35; water (H_2O), 16; and hydrogen fluoride (HF), 4. In any given group of the periodic table, the acidity of the hydrides increases as the group is descended. For example, the two groups at the right-hand side of the table include, respectively, and in descending order, the elements oxygen, sulfur, selenium, and tellurium; and fluorine, chlorine, bromine, and iodine. The pK_a's of the hydrides of the first group are as follows: water (H_2O), 16; hydrogen sulfide (H_2S), 7; hydrogen selenide (H_2Se), 4; and hydrogen telluride (H_2Te), 3. Similarly, hydrogen fluoride (HF) is a weak acid, whereas hydrogen chloride (HCl), hydrogen bromide (HBr), and hydrogen iodide (HI) are all completely dissociated (are strong acids) in aqueous solution. These trends are due to variations in bond strength, electronegativity (attractive power of the atomic

nucleus for electrons), and ionic solvation energy, of which the first is the most important. When a hydride is able to lose two or more protons, the loss of the second is always more difficult because of the increased negative charge on the base—*e.g.*, $H_2S - HS^-$ (pK 7), $HS^- - S^{2-}$ (pK 15); similarly, $NH_4^+ - NH_3$ (pK 9.5), $NH_3 - NH_2^-$ (pK 35).

A simple rule applies to the strengths of the oxyacids, which can be given the general formula $XO_n(OH)_m$, in which X is any nonmetal. In these compounds, the pK decreases with increasing n, but does not depend significantly upon m. When n = 0 (*e.g.*, ClOH, $Si(OH)_4$), pK_a is between 8 and 11; n = 1 (*e.g.*, HNO_2, H_2SO_3) gives pK_a 2–4; whereas with n = 2 or 3 (*e.g.*, H_2SO_4, $HClO_4$), the acids are completely dissociated in water ($pK_a < 0$). These regularities are probably attributable to the sharing of the negative charge of the anion between n + 1 equivalent oxygen atoms; the more extensive the charge spread, the lower is the energy of the anion and hence the stronger the acid.

Strengths of oxyacids

The most important groups of organic acids are the alcohols (including the phenols) and the carboxylic acids. The simple alcohols are very weak acids (pK 16–19); the phenols are considerably stronger ($pK \sim 10$); and the carboxylic acids stronger still ($pK \sim 5$). The strength of the carboxylic acids is due to the sharing of the negative charge between two equivalent oxygen atoms in the ion RCO_2^-. The most important organic bases are the amines, RNH_2, R_2NH, or R_3N. Most of these are stronger bases than ammonia; *i.e.*, their cations are weaker acids than the ammonium ion.

The effect of substituents on the acid–base properties of organic molecules has been very extensively studied and is one of the main methods of investigating the nature of the electron displacements produced by substitution in these molecules. The simplest classification is into electron-attracting substituents (halogens, carbonyl, nitro, and positively charged groups) and electron-repelling groups (alkyl groups, negatively charged groups). The electron-attracting groups make acids stronger and bases weaker, whereas electron-repelling groups have the opposite effects. There are, however, often more specific electronic effects, especially in aromatic and unsaturated compounds, for which special explanations are needed.

THE EXPERIMENTAL STUDY OF ACID–BASE REACTIONS AND EQUILIBRIA

The determination of dissociation constants in aqueous solution. The classical method for determining the dissociation constant of an acid or a base is to measure the electrical conductivity of solutions of varying concentrations. From these the degree of dissociation (α, see above) can be determined and K_a calculated from the equation

Measurements of electrical conductivity

$$K_a = \frac{\alpha^2 c}{1 - \alpha}.$$

This method is unsuitable for acids with pK less than 2 because α is then close to unity and the value $1 - \alpha$ is therefore subject to error. It also is unsuitable for acids of $pK > 7$, because impurities in the solvent may affect the conductivity or displace the dissociation equilibrium.

It is often preferable to use a more specific method for determining the concentration of one of the species in the scheme $A + H_2O \rightleftarrows B + H_3O^+$. For example, a hydrogen electrode (or more commonly a glass electrode, which responds in the same way) together with a reference electrode, commonly the calomel electrode, serves to measure the actual hydrogen ion concentration, or the pH, of the solution. If E is the electromotive force (in volts) observed by the electrode, the equation giving the pH is as follows:

$$E = E_o + 0.059 \log_{10} [H_3O^+] = E_o - 0.059\ pH.$$

In this equation, the value of E_o depends on the nature of the reference electrode and is usually obtained by calibration with a solution of known pH. Measurements can be made on aqueous solutions of the acid, in which case $[B] = [H_3O^+]$, but it is better to use a series of buffer solutions with known ratios [A]/[B], since these are less sensitive to the presence of impurities. Such a series is obtained by successive additions of alkali to a solution of the acid (or of a strong acid to a solution of the base) and the procedure is then often termed a pH titration.

If A and B have different optical properties—for example, if they differ in colour or in the absorption of ultraviolet light—this property can be used to measure the ratio [A]/[B], commonly by using an instrument called a spectrophotometer. Since $[H_3O^+]$ must also be known, the commonest procedure is to measure [A]/[B] in a solution made by adding a small quantity of A or B to a standard buffer solution. If A and B do not have convenient optical properties—as is commonly the case—an indicator; that is, an acid–base system that does show a difference in colour in changing from A to B, is used. If a small quantity of indicator A_I–B_I, with acidity constant K_I, is added to a buffer solution A–B, it is easily shown that the following relation holds:

$$K_A = K_I \frac{[B]}{[A]} \cdot \frac{[A_I]}{[B_I]},$$

in which $[A_I]/[B_I]$ is measured spectrophotometrically, and all the other quantities on the right hand side of the equation are known.

If accurate values of K are required, it is necessary in all the above methods to take into account the effect of interionic forces upon the equation and the quantities measured. This factor can induce a considerable degree of complexity into the problem.

Selected values of acidity constants. Table 2 contains acidity constants for selected substances. These are listed as acids or bases according to the nature of the uncharged species, but in each case the value given is (pK_a) for the acid form (pK_a and pK_b for a conjugate acid–base pair being related by the equation

$$pK_a + pK_b = pK_w = 14.00$$

for aqueous solutions at 25° C). For instances in which several values of pK_a are given, these relate to successive dissociations; *e.g.*, for phosphoric acid they correspond to dissociations of H_3PO_4, $H_2PO_4^-$, and HPO_4^{2-}. All values given refer to aqueous solutions at or near 25° C; parentheses indicate values that have been estimated indirectly, or are uncertain for other reasons.

Table 2: pK_a's of Representative Acids and Bases

	pK_a		pK_a
Inorganic acids		**Carboxylic acids**	
Boric acid	9.1 (20° C)	Formic	3.75 (20° C)
Hypochlorous acid	7.53 (18° C)	Acetic	4.75
Hydrogen sulfide	7.0, 11.9 (18° C)	Chloroacetic	2.85
Carbonic acid	6.4, 10.3	Dichloroacetic	1.48
Phosphoric acid	2.1, 7.2, 12.8	Trichloroacetic	0.70
Sulfurous acid	1.8, 6.9 (18° C)	Oxalic	1.23, 4.19
Nitric acid	−1.6	Malonic	2.83, 5.69
Sulfuric acid	(−3), 1.9	Benzoic	4.19
Hydrogen chloride	(−7)	**Nitrogen bases**	
Perchloric acid	(−8)	Methylamine	10.66
Inorganic bases		Dimethylamine	10.73
Ammonia	9.25	Trimethylamine	9.81
Hydrazine	−0.9, 8.23 (20° C)	Piperidine	11.12
Hydroxylamine	6.03 (20° C)	Aniline	4.63
Alcohols and phenols		Pyridine	5.25
Methanol	15.5	Quinoline	4.90 (20° C)
Trifluoroethanol	12.37	Pyrrole	−0.27
Phenol	9.89		
o-Nitrophenol	7.17		
m-Nitrophenol	8.28		
p-Nitrophenol	7.15		
o-Chlorophenol	8.49		
m-Chlorophenol	8.85		
p-Chlorophenol	9.18		
Picric acid	0.38		

BIBLIOGRAPHY. For a general discussion, see P. WALDEN, *Salts, Acids and Bases* (1929), an account of the early development of concepts in this field; R.P. BELL, *Acids and Bases*, 2nd ed. (1969), an elementary account, *The Proton in Chemistry* (1959); and E.J. KING, *Acid-Base Equilibria* (1965), two books at a somewhat more advanced level. Experimental methods are described in V. GOLD, *pH Measurements* (1956), elementary; R.G. BATES, *Determination of pH; Theory and Practice*, rev. ed. (1964), more advanced; and A. ALBERT and

E.P. SERJEANT, *Ionization Constants of Acids and Bases*, 2nd ed. (1970), a laboratory manual containing tables of ionization constants. Special topics covered in this article are discussed in W.F. LUDER and S. ZUFFANTI, *The Electronic Theory of Acids and Bases* (1946); in collected papers reprinted from the *Journal of Chemical Education, Acids and Bases* (1941), and *More Acids and Bases* (1944), three publications dealing mainly with Lewis acids; and in R.G. PEARSON, "Acids and Bases," *Science*, 151:172–177 (1966), an article on hard and soft acids and bases; L.F. AUDRIETH and J. KLEINBERG, *Non-Aqueous Solvents* (1953); J.J. LAGOWSKI (ed.), *The Chemistry of Non-Aqueous Solvents*, 2 vol. (1966–67); and C.H. ROCHESTER, *Acidity Functions* (1970). For tables of dissociation constants, see G.F.A. KORTUEM, W. VOGEL, and K. ANDRUSSOW, *Dissociation Constants of Organic Acids in Aqueous Solution* (1961); and D.D. PERRIN, *Dissociation Constants of Organic Bases in Aqueous Solution* (1965).

(R.P.B.)

Acoustical Engineering

Acoustical engineering is the field of engineering concerned with the control of sound and noise in the environment and with the provision of good hearing conditions in buildings. Although there has been concern about noise and hearing from early times, modern pioneering work in the field started about 1900. Since then, a considerable body of knowledge has developed, making it possible to solve a wide variety of acoustical problems in advance of construction.

Acoustics often has been regarded as a science dealing only with hearing conditions in auditoriums, but there are other aspects of the environment in which it plays an important role. Every office building, airport terminal, hospital, and even the private house has sound and noise-control problems. If these problems are not recognized and dealt with during the design of the structure, they must be corrected later.

Acoustical engineering is also concerned with the control of noise from external sources, such as transportation systems. In a building near an airport, heavier construction is usually employed, special precautions must be taken with windows, and the building must be sealed and air-conditioned.

The nature of sound. A sound source produces pressure variations in the air, compressions and rarefactions moving outward from the source in a progressive, spherical wave at a speed of about 1,100 feet (335 metres) per second. In free space, the energy intensity in the wave decreases as the distance from the source increases. For each doubling of distance between source and receiver, the perceived loudness of the sound decreases about one third. Sound waves indoors are reflected many times by the enclosing surfaces, the intensity (except near the source) remaining more or less uniform throughout a room. The sound wave carries energy from the source to the ear or other receiver.

The table shows the enormous range of intensity of sounds to which the ear is exposed. Using a threshold value of one arbitrary unit for a sound so faint that it can just be detected, the loudest sound heard is 1,000,000,000,000 times that volume. A more convenient means of expressing this vast range has been developed by using the unit known as the decibel. The number of decibels between two intensity levels is obtained by taking ten times the logarithm of the ratio of the intensities. Although the tolerable range is great, the absolute energy of sound is small, an average voice having only about 1/1,000,000 of the energy needed to operate an ordinary electric lamp (see also SOUND).

The decibel sound rating

Intensity of Sound

type of noise	units	decibels
Airplane engine nearby	1,000,000,000,000	120
	100,000,000,000	110
Airplane cabin (normal flight)	10,000,000,000	100
Heavy traffic, pneumatic drill	1,000,000,000	90
	100,000,000	80
Noisy office, telephone conversation, ordinary traffic	10,000,000	70
	1,000,000	60
Average office	100,000	50
Ordinary conversation	10,000	40
Quiet home, quiet conversation	1,000	30
	100	20
Rustle of leaves, whisper	10	10
	1	0

Criteria for sound-control design. The amount of sound energy that reaches the listener depends on the paths by which the sound has travelled, on the distance from the source, the power of the source, and the nature of such intervening barriers as walls, doors, windows, and sound-absorbing ducts. Whether the acoustical engineer's objective is to make it possible to hear more or less, the characteristics of the source, of the path, and of the receiver must be considered.

Tolerable noise levels

Qualities that characterize a desirable acoustic environment vary widely, depending on how the space is to be used, how particular the users are, and how the specific space involved relates to other parts of the building. A library reading room, for example, should be free of distraction. This objective can be achieved either by keeping the reading room silent or by providing a moderate, continuous, background sound level of unobtrusive, unrecognizable character, sometimes called white noise. Because the background sound hides or masks the minor noises that are unavoidable, the second approach is considered more realistic. In a large business office, a higher degree of background noise might be acceptable, up to a point at which it interfered with conferences and telephone conversations. In general, people usually tolerate a noise conveying no information better than they do one that tells them something, and an expected noise is usually more tolerable than an unexpected one of the same magnitude.

The acoustical character of a space is as important as the amount of sound generated in it or transmitted into it. If a room is finished in hard, sound-reflecting materials, sounds in the room not only persist but seem to come from all directions.

There are optimum ranges for the reverberant character of occupied spaces. An aid to the acoustical engineer is the fact that the acoustical character does not need to be uniform throughout a building; on the contrary, variations may provide a pleasant contrast, from a large, reverberant, monumental space to a smaller, quieter, more intimate space in which communication is easy.

CONTROL OF NOISE

Outdoors. The most effective method of noise control outdoors is to separate the source and receiver in space, but because this is not always feasible, some type of barrier is often introduced. Walls and earth banks can be used to control many types of outdoor noise, but the effectiveness of any barrier depends on how well it cuts off the line-of-sight path between source and receiver. A wall to control traffic noise, for example, can be effective at ground level, provided it is near either the source or the receiver, but a low wall constructed in front of a multistory apartment house will do nothing to control noise for those above the level of the wall. Trees and shrubs have little effect.

Indoors. When a sound wave strikes a sizable obstacle such as a wall or ceiling, part of it is reflected, part of it is absorbed, and part of it is transmitted to some adjoining space. The relative magnitudes of these three parts of the original sound are determined by the physical properties of the obstacle. A hard-surfaced, dense plaster reflects a great deal of the sound that strikes it, while a soft surface such as a heavy carpet absorbs most sound and reflects little. Absorption and reflection are important only in connection with the space in which the sound originates. More often the problem is the portion of the sound that is transmitted to an adjoining space. Although the mechanisms governing sound absorption and sound isolation are complex, they must be understood in order to predict the resulting acoustical environment in a finished building.

Sound absorption. Useful sound absorption is provided by porous materials such as carpets, draperies.

Sound-absorbing materials

glass-fibre blankets, clothing, and the many specially made, sound-absorbing materials. An essential property of a sound-absorbing material is that it have a porous structure into which the molecules in the air carrying sound energy can dissipate, thus losing energy in the form of heat. Sound energy so dissipated is never recovered as sound. Because it is essential that the interstices of sound-absorbing materials be intercommunicating, a closed-cell foam material is not a good sound absorber. Porous sound-absorbing materials may sometimes be faced with hard-surfaced, refinishable perforated panels, wood strips, or the like, but these facings must be so designed that the sound can penetrate into the porous material behind the facings. In general, the solid parts of the facing material should not be larger than two or three inches (five to eight centimetres) and the openings between these elements to the sound-absorbing material should represent 10–25 percent of the area.

When sound-absorbing materials are placed within a room, the reflection of energy is greatly reduced and the sound dies away rapidly instead of being reflected many times by the enclosing surfaces. When the sound dies away quickly, the room has what is called a short reverberation time. If surfaces are highly reflective and sound continues to reflect from surface to surface with almost no loss, the room has a long reverberation time. The "feel" or quality of a space as sensed by the occupants is determined largely by the amount and type of sound-absorbing finish.

Dead and live rooms

When a room has a short reverberation time, sounds seem to come from their point of origin rather than from everywhere, surrounding the listener with many reflections. There is actually some reduction in the level of sound from a source in such a "dead" room when contrasted with a hard, reverberant, sound-reflecting "live" room. A nonreverberant space seems more comfortable, and speech communication is easier because the jumble of successive sounds is absent. The proper use of sound-absorbing finishes can improve the quality of almost any space in which a number of people must work, live, or play together. Large open-plan offices and schools work satisfactorily only when both the floor and ceiling are covered with a sound-absorbing finish.

Sound-absorbing ceilings can be very useful for noise control in factories, but the source of the noise must also be considered. Since a sound-absorbing ceiling affects only the noise level in the overall space, it has no effect on the level of sound reaching the ear of a worker near a noisy machine. In areas where noise is less intense, as in restaurants and college dining halls, a sound-absorbing ceiling is more effective. The potential of sound-absorbing materials in private residences has been largely neglected. A sound-absorbing ceiling in a dining room, for example, can reduce the din at a dinner party (whether the room is carpeted or not) to a pleasant level that permits conversation at ordinary voice level. The kitchen, with its customarily hard finishes, can also be a more pleasant space in which to work if it has a sound-absorbing ceiling.

Sound-absorbing materials are used as linings for air ducts to reduce the transmission of sound from room to room and from the fan. Also, mufflers for internal combustion engines incorporate either resonant or dissipative sound-absorbing materials to reduce the noise transmitted with the exhaust gases.

Sound isolation. While sound-absorbing treatments are useful in the control of the character of occupied spaces and in reducing the magnitude of noise in an enclosure, they are useless as noise barriers. A curtain hung in a doorway or over a wall provides no improvement in sound isolation. Acoustic tiles nailed to the underside of floor joints do not improve the isolation between the living room and a downstairs playroom.

The sound isolation provided by a barrier such as a wall, floor, ceiling, or window, is determined almost entirely by its weight or inertia. Since a sound wave consists of the back-and-forth oscillating motion of molecules in the air, these moving particles, when they encounter an obstacle, tend to make the obstacle oscillate slightly. If the obstacle has a great deal of weight or inertia, it resists this attempt at being moved, but if it is relatively lightweight it may move quite easily. When the obstacle is moved by action of the sound wave on one side, it radiates a new sound wave on the other side. This new sound wave is the transmitted sound, the sound heard in the adjoining space. The sound energy is transmitted from one side of the partition to the other by the direct motion of the partition. The heavier the construction, the less sound gets through. A great deal less sound can be heard through an eight-inch brick wall, for example, than through a lightweight plasterboard partition.

Any construction must be completely airtight if it is to be useful as a sound isolator. An opening in a barrier (such as an open door) transmits much more sound energy than an area of wall of equal size. A wall of porous concrete block or construction with loose joints allows sound energy to be transmitted through the leaks and cracks. Doors and windows work at their best as sound barriers only when they are tightly weatherstripped. Probably the most difficult problem in modern building construction is to make airtight connections between the many dry, prefabricated elements that form the enclosure and subdivision of buildings. Every service penetration for piping, every shrinkage crack, and, in fact, any leaks at all seriously limit the ultimate performance of noise-isolating construction.

Sound-isolating construction

Since it is often impractical to use extremely heavy construction for walls and partitions in buildings, two separated elements, instead of a single heavy unit, are sometimes employed. Two layers of glass, for example, separated by a space of about six inches (15 centimetres), can be as effective as a heavy masonry wall. A double window of this sort is a good isolator provided it is completely weatherstripped. Ordinary sealed double glass, with an air space of 0.5 inch (1.25 centimetres) or less, is not really a double construction from the acoustical point of view, because the air space is too small to permit independent motion of the two pieces of glass. Double plaster walls, using separate framing for each face, are often used for party walls between apartments. This construction can be very effective, but its acoustical effect is seriously impaired by back-to-back convenience outlets or penetrations for heating pipes or plumbing.

Background noise. The amount of acoustical privacy between two spaces is related to the background noise level as well as to the sound isolation provided by the construction. Architects have often been surprised to find that a partition type that worked well in a city office building was far from satisfactory in a quieter location. In such cases, a small amount of white noise may be added to the ambient noise level.

Background noise should be bland, unnoticeable, and continuous. Steady traffic noise of the city or the characteristic sounds of many types of air diffusers are the kinds of background noise that people expect in a typical modern building. When there are no partitions, the background noise becomes even more important in determining privacy. A large, elegant, but sparsely filled dining room is often so quiet that patrons must speak in low tones to preserve privacy, while with a moderate level of conversation throughout an ordinary restaurant, good privacy can be had at comfortable voice levels.

Structure-borne noise and vibrations. Sound is also transmitted in a building when the structure sustains a direct impact or when a vibrating object is in rigid contact with it. The structure then radiates sound energy into the air. For this reason a neighbour's doorbell, if fastened directly to a common wall between apartments, can often be heard as loudly in one apartment as the other. The sound of movements from the floor above is a familiar structure-borne sound, as is the noise from a rigidly attached appliance, or from the great variety of mechanical devices that fill modern buildings. As is the case with airborne sound, the heavier the construction the less sound of impact and vibration is transmitted.

Resilient isolation

Combined with massive construction is the need for resilient isolation. The noise of a footfall overhead is considerably reduced when the floor is covered by heavy

carpeting. Footsteps can be made almost inaudible if, during construction, an additional layer of concrete or a wood floor is "floated" on special resilient pads on top of the basic structural floor. A doorbell on a resilient mounting is inaudible in a neighbour's apartment. A garbage grinder, attached to the sink through a rubber collar, is relatively quiet. Water pipes and drains, isolated from walls and floors with resilient materials, perform their functions with little noise. Fans and compressors, mounted so that they are free to vibrate on springs or rubber mountings, create little noise, although in the absence of such mountings, the whole structure acts as a sounding board for these devices.

Not only must moving equipment be properly mounted, but it must also be connected to necessary electrical and plumbing services through flexible connectors. Any rigid ties to the structure of the building nullify the value of resilient mountings.

CRITERIA FOR GOOD HEARING CONDITIONS

Good hearing conditions for an audience listening to speech or music, whether in an auditorium, courtroom, church, outdoor theatre, or living room, depend on four basic conditions. The space must be quiet, the sounds to be heard must have adequate loudness, there must be a good distribution of sound, and the sounds must be properly blended, with adequate separation for good articulation of music or speech.

Absence of noise. Perhaps the single most important factor in providing good hearing conditions is the absence of any interfering noise. Background noise that can be useful in providing privacy and freedom from distraction in the office or apartment has no place in the room in which there is serious listening to speech or music. Not only must the air-handling system be inaudible, but there must be no noise transmitted from other spaces in the building or from outside.

Adequate loudness and distribution. Obviously, the sounds to be heard should be loud enough and should be uniformly distributed. People sitting in the front of a room should not be subjected to greater volumes of sound than those in the rear. Hot spots, where the sound is too loud, and dead spots, where the sound is too faint, can be as unsatisfactory as seats in which everything is heard twice due to delayed echoes. Adequate loudness and good distribution of sound are determined largely by the shape and surface finish of the room.

Figure 1A illustrates how sound is distributed when an audience is seated on level ground outdoors. In a free field, a sound wave decreases in intensity by a factor equal to the square of the distance from the source. As the sound wave grazes the heads and clothing of the sound-absorptive audience, however, there are additional losses. People seated far back in such an audience receive less sound energy not only because they are far away from the sound source but because some of it is absorbed by the people in front of them. In such a seating area, loudness and distribution requirements are poorly met.

The ancient Greeks and Romans solved the problem by placing audiences on steep hillsides, which resulted not only in much better sight lines but in significantly smaller losses by absorption of the intervening audience (Figure 1B). Good sight lines almost always mean good hearing lines. An alternative to the steep hillside is to raise the source of sound high above the audience, but this is not usually convenient. A more satisfactory indoor arrangement is to let the ceiling of the room act as a sound mirror, bringing reflected sound down on top of the audience to reinforce the direct sound (Figure 2). In

Sound mirror

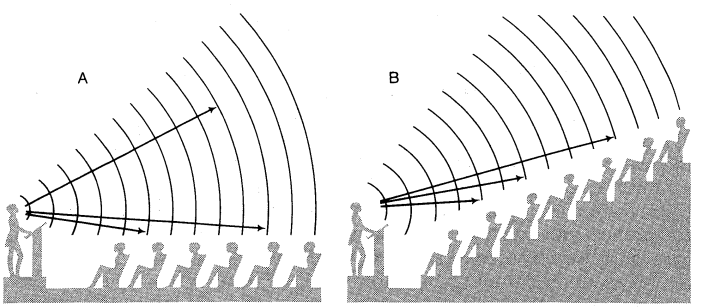


Figure 1: (A) Distribution of sound when audience is seated on ground level outdoors. (B) Distribution of sound in an outdoor theatre of the classical Greek design.

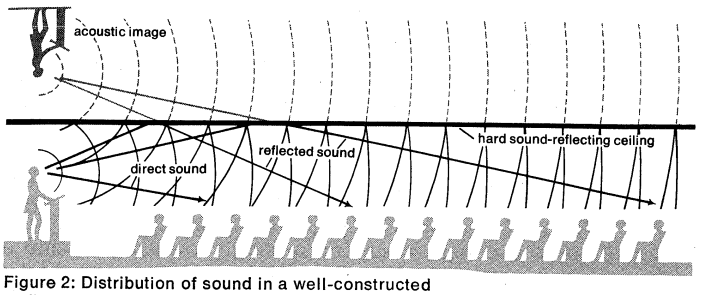


Figure 2: Distribution of sound in a well-constructed auditorium.

many auditoriums more sound energy is received from the ceiling and other reflectors in the room than directly from the source of sound itself. The ceiling is the most important surface in the room in assuring adequate loudness and good sound distribution. A ceiling in any room that must have good hearing conditions for an audience must always be hard and sound-reflecting, not sound-absorbing. This principle is often overlooked; in fact many auditoriums and other buildings for audience use have been deliberately but mistakenly equipped with sound-absorbing ceilings. Such use of sound-absorbing material inevitably produces bad hearing conditions because the room has been deprived of its most useful sound reflector. Sound-absorbing material is useful in rear wall surfaces to control echo and reverberation.

An equally important function of the sound-reflective ceiling is the mixing of sounds that emanate from various parts of a large performing group; *e.g.*, a congregation in a church or an orchestra or a chorus on a stage. Whenever people perform together, it is essential that all members of the group hear each other. An orchestra cannot perform satisfactorily in a conventional stage-house surrounded by velour draperies, because the various instrumentalists cannot hear each other. Some type of overhead reflector must be used.

If a listener is to receive sound energy by reflection from the ceiling, he must be seated directly under the ceiling; seats in a deep underbalcony area are usually the poorest in the house. The undersurface of such a balcony should always slope upward toward the front, with a sufficient opening to permit the rearmost seats a line of sight to part of the ceiling.

The reflections from ceiling and wall surfaces useful to the listener in improving loudness, clarity, and articulation must arrive at the listener's ear not more than about 30 milliseconds (thousandths of a second) after the initial sound reaches him. This means that the sound will have travelled over a path not more than about 30 feet (nine metres) greater than the distance between source and listener. Reflections delayed longer than this tend to detract from clarity and add confusion. If a reflection is delayed longer than about 70 milliseconds, it begins to be audible as a separate sound or echo.

To provide the members of the audience in the front half of the orchestra, or main floor, of an auditorium with these early reflections, the ceiling near the performing area cannot be more than 30 or 35 feet (nine or 11 metres) above the floor. Ceiling heights of 60 or 70 feet (20 metres) subject much of the main-floor audience to poor sound quality and lack of clarity. In the balconies and toward the rear of the main floor, reflections with a small time delay are received, even if the ceiling is rather high.

BLENDING AND SEPARATION: REVERBERATION TIME

Equal in importance to a quiet listening area and uniformly distributed, sufficiently loud sound is the blend-

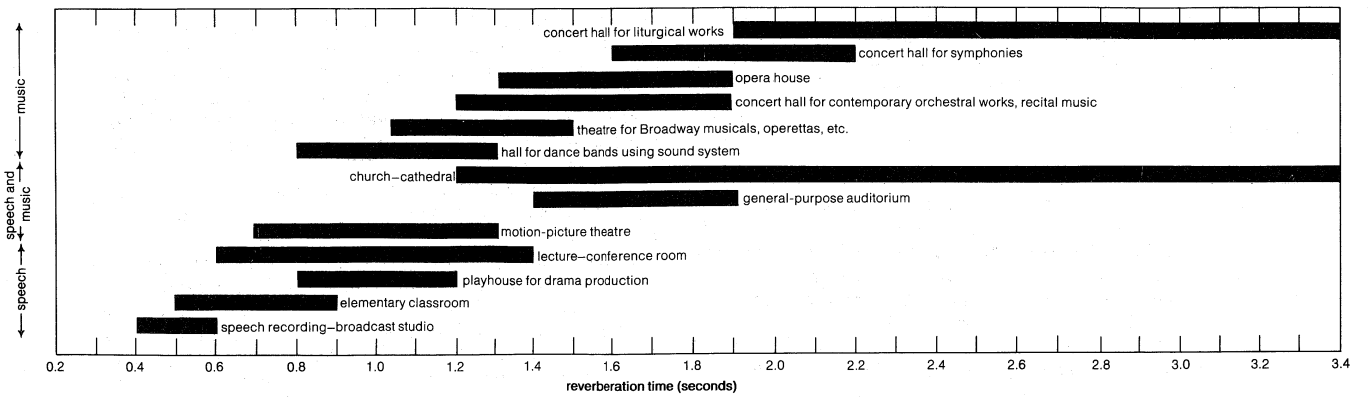


Figure 3: Optimum reverberation time for various types of facilities.
By courtesy of Bolt Beranek and Newman Inc.

ing of sounds to ensure full quality without excessive muddling. This blending process is called reverberation. The reverberation time in a room is arbitrarily defined as the time required for the sound to decrease to $1/1{,}000{,}000$ of its original intensity (or by 60 decibels) after the source is cut off.

Reverberation time is determined by the volume of the room and by the sound-absorbing properties of the finishes, furnishings, and people in the room. The greater the volume, the greater the reverberation time. The greater the area of sound-absorbing finishes, however, the less the reverberation time. In many theatres and auditoriums the audience is the principal absorber of sound. For this reason fabric-upholstered seats that absorb sound in the same way as the audience does are employed to minimize the variation in the reverberation time with a varying audience size. With hard seats, special fabric coverings are sometimes placed over the seats when the theatre is used empty for rehearsals.

In general, a relatively long reverberation time is needed for a sufficient blending of musical sounds but a shorter one is needed for the proper understanding of speech. Over the years, experienced listeners have listened to the sound quality in a large number of auditoriums and have correlated their likes and dislikes with actually measured reverberation time (see Figure 3). Even though a space has a reverberation time lying within the optimum range for a particular function, it does not necessarily follow that it will be an excellent space unless all other conditions for good hearing are satisfied.

Balancing early and reverberant energy

The requirement that reflections of sound reach the ears of the listener not more than about 30 milliseconds after the original signal determines the height of ceiling reflectors. A matter not well understood until recently has been the importance of balancing very carefully the amount of early sound (original signal plus early reflection) with the energy that reaches the ear in the reverberant field (after many reflections around the room). If there is too much early sound, the effect is too articulate and "dry". On the other hand, if there is too much reverberant energy, the effect is fuzzy, or "wet."

The optimum ratio between the early and the reverberant energy varies with the uses of the space. The overhead reflecting surface may take the form of a solid, reflecting canopy, or perhaps the reflecting surface may be made of many panels. The important point is that some early energy be reflected to the ears of the listeners and be balanced with the reverberant sound.

Older halls, noted for good sound quality, often lack such reflecting surfaces. Many have heavily ornamented wall and ceiling surfaces that scatter a great deal of energy back to the listeners, some of it early enough to be useful in improving articulation and clarity.

SOUND AMPLIFICATION

In many situations, adequate loudness and good distribution of sound can only be obtained by reinforcement of the natural transmission from source to listener by means of a sound-amplifying system. In large sports arenas, in airport terminal buildings, and other noisy locations, it is almost always necessary to provide reinforcement. Even in auditoriums where strong-voiced speakers can be heard clearly, the weaker voices must be amplified, and also recorded material or motion picture sound must often be amplified.

Loudspeaker systems

There are two principal types of sound amplification systems: central and distributed. In general, the most realistic amplification is obtained when a directional loudspeaker is placed not more than about 25 feet (eight metres) above the sound source. To the audience, the amplified sound seems to come from the same direction as the original source. Loudspeakers at each side of a platform give much poorer results, with a complete lack of directional realism. High-quality loudspeaker systems, however, are usually quite large, and so are not always easy to place above centre of the stage.

When it is not necessary to associate the amplified sound with the original source, as in an airport terminal, a distributed system of loudspeakers can be used. These must all be operated at a low level so that each listener hears only the loudspeaker near him. This type of system operates much like downlighting. The room is covered with small "pools" of sound. In some auditoriums in which it is impossible to locate a central loudspeaker above the live source of sound, a distributed system of loudspeakers can be used, either in the ceiling if it is low enough, or in the backs of seats. The realism of such a system can be greatly improved if time delay is introduced to allow the amplified sound to reach the listeners slightly later than the original sound.

A line source or column loudspeaker is often used to give greater directionality than can be obtained with a direct-radiator cone speaker. These units must be placed as carefully as any other type of loudspeaker and do not necessarily improve all situations.

Loudspeaker systems are often used unnecessarily. A well-designed hall or theatre seating 300 or 400 people almost never requires one. In auditoriums seating more than 500 a sound-reinforcing system is usually provided, though not always needed, and in larger halls it is frequently desirable to provide some reinforcement of natural sounds. In any case, the sound system must be carefully integrated with the design of the room and with the arrangement of the audience.

BIBLIOGRAPHY. WALLACE C. SABINE, *Collected Papers on Acoustics* (1922, reprinted 1964), a classic work on the development of the reverberation formula, still in use today, together with a fascinating discussion of solutions to acoustical problems, and a debunking of acoustical myths; C.M. HARRIS (ed.), *Handbook of Noise Control* (1957), a collection of 40 articles (each with bibliography) on the nature and control of noise in a variety of situations; L.L. BERANEK (ed.), *Noise Reduction* (1960; rev. ed., *Noise and Vibration Control*, 1971), a detailed text and reference work dealing with sound waves and their measurement, principles, criteria, and methods for noise control, and containing an extensive bibliography; *Music, Acoustics and Architecture* (1962), a critical analysis of factors that determine good room acoustics, including detailed drawings, photographs, and discussion of 54 halls in Europe and North and South America; R.B. NEWMAN and W.J. CAVANAUGH, "Acoustics," in J.H. CALLENDER (ed.), *Time-*

Saver Standards, 4th ed. (1966), a useful reference covering the basic principles of sound control, data on acoustical properties of common building materials, charts useful in design, and drawings of many details to achieve good acoustics in buildings.

(R.B.N.)

Acre

A state (*estado*) of the northern region of the Republic of Brazil, Acre covers the southwesternmost part of Brazil's Hiléia (Hylea), the forest zone of the Amazon River basin. Bounded north by Amazonas state (*estado*), it has western and southern frontiers with Peru and southeastern with Bolivia. With an area of 58,915 square miles (152,589 square kilometres), it had a population of about 220,000 in 1970. The capital is Rio Branco on the Rio Acre in the eastern part of the state. The state's name is derived from that of the Rio Acre, which seems to be of Indian origin. Covered by tropical rainforest, Acre produces the highest quality rubber in Brazil.

History. Portuguese explorers, making their way through the immense Amazon forest from the Atlantic estuary of the Amazon River, did not reach Acre before the middle decades of the 18th century, when there were no settled inhabitants but only roving bands of Indians. Under the Brazilian Empire, more expeditions began to penetrate the territory in the 1850s and 1860s; and although the whole area was ceded by Brazil to Bolivia in 1867 (by the Treaty of Ayacucho), the rubber boom of the following decades attracted more and more immigrants from northeastern Brazil. On July 4, 1899, during a local revolution, an independent Republic of Acre was proclaimed by Luís Gálvez Rodríguez, a Spanish adventurer, but this regime was short-lived. After further vicissitudes, in which a Brazilian immigrant, José Plácido de Castro, played the most active role, negotiations sponsored by Brazil's foreign minister culminated in 1903 in the Treaty of Petropólis, whereby Acre was reincorporated with Brazil. The frontier with Peru was agreed upon in 1909. Organized at first as a territory of Brazil, Acre achieved statehood in June 1962.

The Bolivian interregnum

Physical geography. With an average altitude of 600 feet (183 metres) above sea level, the land slopes gently down toward the Amazonian Plain from the higher ground in the west and in the south. The soils are sediments laid down within the past 65 million years (Cenozoic Era). An average annual temperature of 77° F (25° C) and an annual rainfall of 79–98 inches (2,000–2,500 millimetres) makes the climate warm and humid. The forest is traversed by numerous headstreams of two of the Amazon's major tributaries, the Juruá and the Purus. The principal trees of the forest are *Hevea brasiliensis* and *Bertholletia excelsa*, yielding rubber and Brazil nuts, respectively. The indigenous fauna include peccaries (piglike hoofed animals), red deer, capybaras (a rodent with no tail and partly webbed feet), agoutis (short-haired, short-eared, rabbitlike rodents), and tapirs (a large hoofed quadruped).

Population. The local Indian tribes are few and small, most of Acre's population consisting of immigrants, or descendants of immigrants, from northeastern Brazil. About 80 percent of the people live on the rivers and tracks that give access to the raw materials of the forests; the remainder live in the towns. The largest city, Rio Branco, had a population of 35,000 in 1970; other major towns are Cruzeiro do Sul, Sena Madureira, Tarauacá, and Feijó. The common language is Portuguese, the predominant religion Roman Catholicism. Amoebic dysentery, malaria, and leprosy are the most distressing endemic diseases.

Administration. According to the law of 1962, the system of internal government is determined by a state legislative assembly. There are 16 seats in the assembly occupied by nine state and seven federal representatives; elections are held every four years. There is also a state governor, elected for a five-year term.

Economy. Acre's chief natural resource is rubber, which constitutes 80 percent of the state's exports. The indigenous rubber tree yields the best rubber of all Amazonia, and Acre is Brazil's leading producer of that commodity, its annual output of 10,000 to 15,000 tons amounting to about one-third of the national total. Each tree can yield a yearly average of 4.4 pounds (2 kilograms), the highest yield obtainable from natural forest trees, and a single worker can extract 3,300 pounds (1,500 kilograms) per annum.

Rubber

Brazil nuts are the next most important natural product of the forest. About 4,000 to 8,000 tons are exported every year.

Agriculture, which is only of the subsistence type, is concerned with short-cycle crops, chiefly manioc, corn, and beans. Zebus (humped oxen) are being raised on the open-range system, but this activity is still at an early stage; pigs and chickens are bred quite widely. Though hunting and fishing contribute an important element to the people's diet, the state depends upon imports for most of its consumption. It exports practically all its marketable produce.

Transportation. Development has been held back by the high cost of production and by difficulties of transport. Rivers are the main channels of communication. Rio Branco is connected with Manaus, capital of Amazonas, by the Rio Acre-Purus-Amazon linkage and with Brasília—the federal capital—by a land highway; there are also air services connecting Rio Branco with both places. Belem, in Pará state, and Manaus are Acre's chief external markets.

(J.C.M.C.)

Acting

Whether on the stage, in films, or in television, acting is generally agreed to be a matter less of mimicry, exhibitionism, or imitation than of the ability to react to imaginary stimuli. Its essential elements remain the twin requisites enunciated in the 18th century by French actor François-Joseph Talma: "unusual sensitivity and extraordinary intelligence." The intelligence he refers to comes not from book learning but from the ability to understand the workings of the human personality.

The essential problems in acting—those of whether the actor actually "feels" or merely imitates, of whether he should speak naturally or rhetorically, and of what actually constitutes being natural—are as old as theatre itself. They are concerned not merely with "realistic" acting, which arose in the theatre in the 19th century, but with the nature of the acting process itself.

Such problems acquired increased urgency in recent times, in which an immense investment of time, effort, and money is required for a dramatic production in the commercial theatre, motion pictures, or television; often the viability of such a production depends on the actor's ability to revivify states of high emotion on command, with complete plausibility and without being overwhelmed by other aspects of production. In a film, for example, because of technical problems, the actor sometimes must re-enact a scene a dozen times—each time striving for perfection—and then proceed to an unrelated scene, from some other part of the story, demanding an entirely different attitude from him. In such circumstances the training of an actor is not a luxury but a necessity.

Yet, in an otherwise excellent contemporary handbook of the theatre, the editor apologized for the absence of a section on the actor, which subject, he said, had been assigned to one who, though practically and theoretically conversant with acting, finally acknowledged that he could offer no recipes to the professional nor could he find any logical foundation with which to begin. As early as the middle of the 18th century, the German critic and dramatist Gotthold Ephraim Lessing had already drawn attention to this difficulty. "We have actors but no art of acting." It remains as true today as when stated by George Henry Lewes in his excellent *On Actors and the Art of Acting* (1875):

> I have heard those for whose opinions in other directions my respect is great, utter judgments on this subject which proved that they had not even a suspicion of what the art of acting really is.

Difficulty of definition

Efforts to define the nature of an art or craft usually are based upon the masterpieces of that field. Without that necessary reference point, vague speculations and generalizations—without proof of validity—are likely. In the visual, musical, and literary arts, this foundation exists; the work of the great masters of the past and the present serves not only to elucidate the art but also to create standards to emulate. It is difficult to imagine what the present state of comprehension of music would be if only the music of today were available, and the achievements of Monteverdi, Bach, Beethoven, and Mozart had to be known only by hearsay. Yet, this is precisely the situation that exists in acting. The actor, in the words of the 19th-century American actor Lawrence Barrett, "is forever carving a statue of snow." That is why the understanding of acting has not equalled the appreciation of it and why the actor's creative process has defied comprehension.

Despite these problems, an attempt is made in this article to analyze the art of acting and discuss the techniques that have been found useful in helping actors to develop their art. No attempt is made here to present a "history of acting," because, as has been suggested, there is virtually no record of it apart from the literature on the history of theatre. There is much of interest on acting in other articles on the history of theatre, such as THEATRE, WESTERN and DANCE AND THEATRE, EAST ASIAN. The subject of acting also receives attention in the articles THEATRE, ART OF and DIRECTING, and, in the medium of films, MOTION PICTURES, ART OF.

THEORIES OF TRADITIONS

The actor as creative artist and interpreter

Throughout the history of theatre, continual confusion has been expressed over the question of whether the actor is a creative artist or simply an interpreter. Since the actor's performance is usually based on the play, and the dramatist is conceded to be a creative artist, it is sometimes concluded that the actor must be only an interpretive artist. Some modern exponents of the actor's creativity have indirectly accepted this view and have turned, therefore, to nonverbal theatre. But others deny that this recourse to primitivism is necessary in order to make acting a creative art. When composers like Schubert or Schumann create musical settings for the poems of Heine or Goethe, their music does not lose its essentially creative nature. Verdi used Shakespeare's Othello and Falstaff for his great operas, but his music is no less creative for that. When an artist uses the work of another artist in the same medium, that may properly be called noncreative imitation: the original artist has already solved the basic problems of execution, and his pattern is simply followed by the imitator. Such a work can be considered merely an exercise in skill (or in execution). An artist in one medium who uses an art work of another medium as subject matter, however, must solve all the problems posed by his own medium—a creative achievement. It is therefore quite proper to speak of a character as if he were the actor's creation—of John Gielgud's "Hamlet," for example, or John Barrymore's or Johnston Forbes-Robertson's. Because a medium offers the potential for creativity, of course, not all of its practitioners are necessarily creative: there are imitative artists in every medium. But acting can only be understood after it is first recognized as a creative medium demanding a creative act. In "The Art of Acting" the American drama teacher Brander Matthews remarked,

> The actor needs to have under control not only his gestures and his tones, but all other means of stimulating sensibility and these should be ready for use at all times, wholly independent of the words of the text.

In the same work he quoted with approval the words of the great 19th-century Italian tragedian Ernesto Rossi that a "great actor is independent of the poet, because the supreme essence of feeling does not reside in prose or in verse, but in the accent with which it is delivered." And even Denis Diderot, the French philosopher of the 18th century whose famous *The Paradox of Acting* (written 1773–78; published 1830) is dealt with below and who was himself a dramatist, stated:

> even with the clearest, the most precise, the most forceful of writers, words are no more, and never can be more, than symbols, indicating a thought, a feeling or an idea; symbols which need action, gesture, intonation, and a whole context of circumstances, to give them full significance.

If the art of acting is regarded as merely interpretive, the external elements of the actor's skill tend to be emphasized, but, when acting is recognized as a creative art, it leads inevitably to a search for the deeper resources that stimulate the actor's imagination and sensitivity. This search presents difficult problems. The actor must learn to train and to control the most sensitive material available to any craftsman: the living organism of a human being in all of its manifestations—mental, physical, and emotional. The actor is at once the piano and the pianist.

Acting should not be confused with pantomime, which is a form of external movements and gestures that represents an object or an event but not its symbolic significance. Similarly, the actor is not to be mistaken for an imitator. Many of the best imitators are unable to act in their own person or to create a character that is an extension of themselves rather than an imitation of someone else. Again acting should not be confused with make-believe, which functions only as a form of self-expression. It is not exhibitionism: a faculty for showing off or entertaining at parties is not the same as the talent demanded of the actor—the ability to put oneself into another character, to create a nonexistent event and perform it to its logical fulfillment, to repeat this performance not only when he is in a favourable mood but also at specified times and places, regardless of his own feelings on each occasion—in other words, the ability to create and to respond to imaginary objects and circumstances.

Genuine and feigned emotion. The most famous instance of supposed acting in ancient Greece was that of the actor Polus performing in the *Electra* of Sophocles, at Athens. The plot requires Electra to carry an urn supposed to contain the ashes of Orestes and lament and bewail the fate she believed had overtaken him. Accordingly, Polus, clad in the mourning garb of Electra, took from the tomb the ashes and urn of his own son (who had recently died), embraced them as if they were those of Orestes, and rendered not the appearance or imitation of sorrow but genuine grief and unfeigned lamentation. Rather than acting, this was in fact real grief being expressed.

From antiquity, rival traditions of acting can be discerned—one stressing the externals of voice, speech, and gesture and the other looking to the actual emotional processes of the actor. Aristotle defined acting as "the right management of the voice to express the various emotions," and this primacy of the voice as the actor's outstanding medium has been widely accepted. "Dramatic ability," he said further, "is a natural gift, and can hardly be taught. The principles of good diction can be so taught." Aristotle did not fall into the common mistake of thinking that acting is only good diction; rather, he simply recognized that diction, unlike acting, can be taught; he was well aware of something more than diction in acting but he knew no way of training it. He saw good acting resulting either from a great natural quickness of parts or an enthusiasm allied to madness. "By the first of these, we mold ourselves with facility to the imitation of every form; by the other, transported out of ourselves, we become what we imagine."

If the voice and diction were the essential elements in acting, it would have much in common with oratory, but in actuality the actor is not an orator. The orator speaks in his own character and his intention is to impress and persuade the audience; he may need some of the qualifications of the actor, but many actors are poor orators. The emphasis in oratory is on the skillful use of the voice, and many great actors do not possess impressive vocal instruments. When the great tragedian Edmund Kean appeared for the first time in London in the early 19th century, the critic William Hazlitt, who welcomed him enthusiastically, wished him well and hoped he would soon recover from his cold. He later realized that it was Kean's

natural timbre. Other leading actors, such as Albert Bassermann and Giovanni Grasso, have suffered from asthma and still conveyed a powerful temperament and maintained their vocal expressiveness.

The dichotomy noted in ancient Greece persisted through ancient Roman theatre and into modern times. On the one hand, there was a recognition of the need for the actor to be affected by the sensations he wishes to arouse in others; on the other hand, a need was also seen for a precise system of expression—the peculiar look, tone, and gesture appropriate to every emotion of the mind.

Beginning of modern acting

Modern acting began with the commedia dell'arte. These Italian comedies companies, the earliest mention of which is in 1545, rapidly achieved extraordinary popularity. Until then, acting was an amateur occupation. The actor was limited simply to illustrating the text by means of a narrow scheme of gesture and rhetorical speech. But in the commedia dell'arte the actor used only an outline, a plot; he improvised the play, giving free rein to the actor's art, developing his own characters or masks that he repeated in each play. Each character became an extension of the actor's own personality but elastic enough to respond to innumerable dramatic situations; thus, these actors began to develop the distinctive stage character of the theatre, whereas previously the emphasis had been on the literary aspect of theatre (see also COMMEDIA DELL' ARTE). Since this demanded high skill, the actors joined into companies. The actor became professional. By doing so, he stimulated the development of modern drama. The essential requisite for the drama is its performance. The dramatist's creation finds its fulfillment not in the writer's study but on the stage. This fulfillment can best be achieved through the contribution of the professional actor.

Nonetheless, actors continued to learn by doing. Their schools were professional companies; their classroom, the stage; their teachers, the audience and their fellow players. Schools of dramatic art, isolated from theatres or companies, are a relative innovation and are still unknown in several European countries.

Diderot's "Paradox of Acting." The most significant statement on acting is Diderot's *Paradox of Acting*. Because of its polemic brilliance, it remains the most widely known essay on acting. It has found little acceptance within the profession, though its famous paradox—that in order to move the audience the actor must himself remain unmoved—is still highly regarded.

Regardless of the untenable solutions it proposes, Diderot's essay contains an excellent description of the actor's problem. What bothered Diderot was the unsolved problem of how the actor, if he were full, really full, of feeling, could play the same part twice running with the same spirit and success. Full of fire at the first performance, he would be worn out and cold as marble at the third. Diderot confirmed his view by noting "the unequal acting of players who play from the heart. Their playing is alternately strong and feeble, fiery and cold, dull and sublime." This was the case with the actress who in her day was the outstanding example of emotional acting. "She comes on the stage without knowing what she is going to say; half the time she does not know what she is saying; but she has one sublime moment." Diderot knew that actors do feel and experience; he knew the nature of inspiration. But he also knew that some actors refused to recognize the need for craft, for training.

Diderot asked how the actor, if he is himself while he is playing, is to stop being himself: how he is to catch just the point at which he is to stay his hand? Diderot demanded unity in a performance; he demanded respect for the author's concept, and he understood the difficulty of repeating a performance. He demanded a definite course to the passion—a beginning, a middle, and an end. In Diderot's day, however, the problem of developing a technique for creating inspiration in the actor remained unsolved. The theoretic formulation had already been stated as well as it ever could be by the actor Talma:

> I call sensibility that faculty of exaltation which agitates an actor, takes possession of his senses, shakes even his very soul, and enables him to enter into the worst tragic situations, and the most terrible of the passions as if they were his own. The intelligence which accompanies sensibility judges the impressions which the latter has made us feel: it selects, arranges them, and subjects them to calculation.

The difficulty of solving the problem is illustrated by the work of the 19th-century French teacher François Delsarte, whose influence was widespread not only in France but also in the United States. Delsarte became dissatisfied with routine acting techniques. He observed their mechanical and stultifying character and realized that under the stress of natural instinct or emotion, the body assumes appropriate attitudes and gestures quite different from those described by his teachers. But when he attempted to formulate laws of speech and gesture, on the basis of years of diligent observation and study, he created a series of elaborate pictoral descriptions that were just as mechanical as those he originally criticized. Knowledge of affective behaviour had not advanced far enough to serve as an aid in solving the problem of the actor: there was still too little understanding of human behaviour, of the relation between the conscious and unconscious, and of the role of the senses.

Delsarte's laws of speech and gesture

Stanislavsky's contribution. It is in this context that the enormous contribution in the early 20th century of the great Russian actor and theorist Konstantin Stanislavsky can be appreciated. Stanislavsky was not an aesthetician but was primarily concerned with the problem of developing a workable technique. He applied himself to the very problems that Diderot and others had believed insoluble: the recapture and repetition of moments of spontaneity or inspiration, which could not be controlled and repeated at will even by many of the greatest actors. Stanislavsky dedicated himself to the central problem of how to stimulate the actor's creativity. Even early in his career, while watching performances by great actors, he had felt that all of them had something in common, something he encountered only in greatly talented actors. In his later work, as director of the Moscow Art Theatre, he often experienced those flashes of intuition or inspiration that stimulate the imagination and turn something that one understands with the mind into an emotional reality and experience. Stanislavsky described such a moment occurring at a low point in the rehearsals for Anton Chekhov's drama *Three Sisters*, when "the actors stopped in the middle of the play, ceased to act, seeing no sense in their work." Suddenly something incomprehensible happened: an accidental sound, of someone nervously scratching his fingernails on the bench on which he sat, reminded Stanislavsky of a scratching mouse, setting off an entire sequence of previously unconscious memories that put the work at hand into a new spiritual context.

Later, in examining many parts he had played, especially that of Dr. Stockmann in Henrik Ibsen's *An Enemy of the People*, a role he felt was better suited to him than any other in his repertoire, he became aware of how much his characterizations had been based unconsciously on his memories. With the passing of time, however, the memories and the feelings aroused by them were lost, and he began to repeat mechanically the fixed appurtenances of the role—the movements of the muscles, the mimetics of the face, eyes, arms, and body, and the physical signs of absent emotion. This led him to the perception that creativeness on the stage demands a condition which he called "the creative mood." To the genius on the stage, this condition almost always comes of itself, and less talented people receive it less often. Although everyone on the stage receives the creative mood sometimes, none seemed able to control it with their own will.

The role of the creative mood

Stanislavsky's description of the problem thus far had reached the point at which all previous examinations had stopped. By going further and inquiring into technical means for controlling the creative mood, Stanislavsky laid the foundation for the modern approach to the actor's problem. Stanislavsky had no intention of creating inspiration by artificial means; rather, he wanted to learn how to create favourable conditions for the appearance of inspiration by means of the will. He emphasized that

this problem could be evaded: other artists may create whenever they are of a mind or feel inspired, but "the artist of the stage must be the master of his own inspiration and must know how to call it forth when it is announced on the poster of the theatre." If he is unable to find a conscious path to unconscious creativeness, the actor is forced to rely on the superficial aspects of scenic craft and theatrical cliches.

Stanislavsky believed that the problem could be solved through advanced psychology, especially the concept of "affective memory" described by the French psychologist Théodule Ribot in the 1890s. Although there has been confusion and misunderstanding about it, and its very existence has been questioned, the concept of affective memory is of prime importance for the understanding of how spontaneous and emotional experiences occur and can be repeated on the stage.

Affective memory is a reliving of past experiences—likes and dislikes—when an analogous situation recurs. Something that has brought pain is anticipated with dislike the second time. This dislike, which is felt immediately, rather than remembered, is like a residue of previous appraisals. Affective memory may be linked directly to the memory of a traumatic experience, as the same situation or a similar one recurs, or to an experience that bears little apparent relation to the original, if the memory has been repressed. Of course, an experience need not necessarily be traumatic to leave an affective memory. The concept of affective memory has found a place in several schools of psychology, including the Freudian and the Pavlovian, though different explanations have been offered.

The concept embraces both sense memory and emotional memory. The latter term was used by Stanislavsky in his later work to refer to the experiences of an intense and explosive nature that are so necessary for the most dramatic moments in the theatre. The concept of affective memory is essential to an understanding of how the actor functions and the faculties that have to be trained to develop his talent. It is his unusually sensitive affective memory that enables the actor to respond to events that must be imagined on the stage and to repeat performances. This point was stressed by Stanislavsky's great pupil Yevgeny Vakhtangov, who emphasized that literal emotion—emotion that derives from the presence of an object that actually stimulates it—cannot be controlled and cannot be relied upon to provide the level of response that is required in every performance.

The use of affective memory is not limited only to acting. Wordsworth defined poetry as originating from "emotion recollected in tranquility." Marcel Proust, in a long passage in *Swann's Way*, brilliantly described the working of affective memory and illustrated precisely the way in which it can be recalled. Instances of its presence can be multiplied from all the arts—literary, visual, or musical. But, though in the other arts it can function unconsciously, the actor must learn to use it consciously to satisfy the unique conditions under which he must create.

Stanislavsky's "Method"

The "Method" is the name by which the totality of Stanislavsky's ideas have become most widely known. The Method represents a development of his procedures based not only on his writings but also on his actual achievement in his major productions. It includes the work of Vakhtangov, who demonstrated that Stanislavsky's ideas apply to the essential problems of the actor in any style and not only the realistic style most often associated with them. The Method became widely known in mid-20th century largely through the work in films of actors such as Marlon Brando, Rod Steiger, and Geraldine Page, who had studied at the Actors Studio in New York City. These actors made a powerful impression and showed a remarkable ability to bridge the gap between stage, screen, and television to an extent that aroused excitement and interest in the rest of the world. So strong was the fusion of performer and role that many of the traits of the character were confused with those of the actor, which led to serious misunderstanding. But at mid-20th century an American style of acting was being born.

Later developments. The post-Stanislavsky period has been influenced chiefly by the ideas of Antonin Artaud, Bertolt Brecht, and Jerzy Grotowski (who pointed out that the other two were more concerned with aesthetics than with methods). Artaud, a French avant-gardist director and actor, exerted an enormous posthumous influence on contemporary theatre through his writings. In them he proclaimed the "theatre of cruelty," which is based on the extreme development of gesture and sensory responses by the actors so that they can communicate with the audience at a more profound psychological level than is possible through words. Artaud's ideas achieved international attention through the productions of Peter Brook and the Royal Shakespeare Company, especially *The Persecution and Assassination of Jean-Paul Marat as Performed by the Inmates of Charenton under the Direction of the Marquis de Sade*, which called for emotional states verging on hysteria from most of the cast during each performance.

The ideas of Artaud, Brecht, and Grotowski

Contrary to the opinion of many, however, Artaud thought of the theatre not as a psychological but as a plastic and physical domain and of the actor as an "athlete of the heart." For every feeling, every mental action, and every leap of human emotion, there is a corresponding breath that is appropriate to it. It has been pointed out by Grotowski, however, that if Artaud's principles are analyzed in a practical way, "they lead to stereotypes: a particular type of movement to exteriorize a particular type of emotion. In the end, this leads to cliches."

Through his plays and the remarkable productions of the Berliner Ensemble in East Berlin in the 1950s, which represent the most important contribution to theatre of the post-Stanislavsky period, Bertolt Brecht generated ideas about acting that have received wide prominence and have usually been counterposed to those of Stanislavsky. Whereas in Stanislavsky-inspired productions the actors often seem to be exaggerating their individuality, Brecht's characters struck many observers as existing primarily as representatives of a class—in some cases self-effacing to the point of dehumanization. Brecht himself, however, denied that his ideas were opposed to Stanislavsky's. He drew attention to the relatively quiet style of acting that sometimes struck visitors to the Berliner Ensemble. Brecht called his approach epic realism. He stressed that the stage of a realistic theatre must be peopled by live, three-dimensional, self-contradictory people, with all their passions, unconsidered utterances, and actions. The actor has to be able to create such people. Brecht mentioned some of the procedures of Stanislavsky he felt indebted to—the creation of the given circumstances that motivate the beginning of an event, the emphasis on creating the activity of the day that helps to define the actor's behaviour, and the individualizing of the characters that make up a mass. And he warns that "we shall get empty, superficial, formalistic, mechanical acting if in our technical training we forget for a moment that it is the actor's duty to portray living people."

Brecht's epic realism

A contemporary Polish director, Jerzy Grotowski, made the most thorough effort to rediscover the elements of the actor's art. Although he credited Stanislavsky with having posed the most important questions, he was not satisfied either with Stanislavsky, who let natural impulses dominate, or with Brecht, who was too much concerned, Grotowski felt, with the construction of the role. To Grotowski, the actor is a man who works in public with his body, offering it publicly. The work with the actor's instrument consists of physical, plastic, and vocal training to guide him toward the right kind of concentration, to commit himself totally, and to achieve a state of "trance." The actors concentrate on the search for "signs," which express through sound and movement those impulses that waver on the borderline between dream and reality. By means of such signs, the actor's own psychoanalytic language of sounds and gestures is constructed, in the same way as a great poet creates his own language.

The actors of Grotowski's troupe are superbly trained physically and vocally, and they commit themselves to their task with total energy. They have been accused,

however, of conveying too little human emotion. Grotowski's criticism that Artaud's work leads to cliches has also been made of his own work. There is little evidence in it of the level of acting observed in the work of great actors.

THE ACTOR'S QUALIFICATIONS AND TRAINING

In view of the diversity of approaches to the actor's problems, it would seem difficult to arrive at any useful generalizations that are valid for all of them. Even among theatre groups that approach the production of a play from a fixed style or a fixed scale of expression, as in Japanese kabuki and classic Oriental theatre generally, the same basic concerns are apparent. The following is an attempt to set down an approach that has proven successful in a variety of professional procedures.

Looks and personality

The qualifications of the actor are generally thought to be a good physique, a retentive memory, an alert brain, a clear, resonant voice with good articulation, and controlled breathing. While looks and the even more important element of personality are undoubtedly factors, their characteristics are difficult to determine; they are usually recognized after the actor has become successful rather than before. Many actors do not possess them offstage but seem to ignite them as soon as they begin to perform. The modern mediums of the cinema and TV have intensified the importance of these external elements, continually searching for actors resembling those who have already achieved success, without helping to clarify these qualities. Yet the central element of the actor's talent, as differentiated from his means, is a special sensibility ("fire," "enthusiasm," "spirit," in the words of 18th-century theoreticians), an ability to respond to imaginary stimuli and situations, which makes it possible for him to enter into the experience and emotions of the character he is to represent. These elements have always been recognized as distinguishing the great actor but were assumed to be beyond the reach of the ordinary actor; they were regarded as elements "born in him" and not susceptible to training. This is precisely the area of the modern training of the actor.

The first stage in the training of the actor's control of his physical, mental, and psychical resources is the ability to relax. Because this ability seems to have little to do with the final achievements in acting, it is often disregarded, but it is basic to any expenditure of will and energy on his part. In a state of physical or mental tension, or both, the actor cannot think, the commands he gives himself are not transmitted, sensation is stifled, and expression is inhibited. The process of relaxation serves to clear the actor of the unnecessary pressures that he has accumulated before the moment of acting begins, to free him of blocks or interferences that may inhibit sensory responses. Physical and mental energies are comparatively easy to train, but sensory control is much more difficult. Relaxation is not a static state or effort. Often in the initial stages of training the actor is subject to strong eruptions of unconscious impulses. He must learn to continue the relaxation, to force his will to maintain his effort on the action of the nerves and the muscles. Relaxation is connected with the exercise of the actor's will.

The role of concentration

The converse of relaxation is concentration. Everything the actor does demands concentration. His training proceeds by work with imaginary objects: working with real objects often leads to pantomimic or to physical imitation, but the actor may begin with them in order to learn how to respond with his entire organism and to apply such responses to his work with imaginary objects—the real medium of the stage—as he would to real ones. This capacity to respond to stimuli that come not from outward reality but from the promptings of one's own imagination may be seen to some extent in every human being; something akin to it is found in psychology in the study of conditioned reflexes, of automatic and spontaneous reactions, and of behaviour patterns. In heightening the sensory awareness and stimulating the senses to respond more strongly in life, the actor acquires the ability to recreate any object, sensation, or experience in the imagination. Here the concept of affective memory, already discussed, plays a fundamental role.

The French actor Talma has already been quoted as stressing the actor's need for an "excess of sensibility, always more lively, more rapid, and more powerful." But he has also properly pointed out that

> his intelligence must always be on the watch, and acting in concert with his sensibility, regulate its movements and effects: for he cannot, like the painter and the poet, efface what he does. To form a great actor the union of sensibility and intelligence is required.

In strengthening his concentration, the actor uses not only will but also a process of self-awareness by which he observes what is happening and trains his instrument to respond to his commands. The very process of concentration and of commitment and involvement must include awareness. The more the actor learns to master concentration, the more aware he becomes.

The actor's sense of truth

An additional factor is the development of the actor's sense of truth—a faculty particularly stressed by Stanislavsky. The growth of self-awareness is useless if it is not accompanied by a correct evaluation of what is true and what is false. If the actor must rely on outside judgment and remain dependent on it, he may become insecure and lose his spontaneity and responsiveness. His mastery of inner relaxation and concentration helps him achieve a combination of spontaneity, commitment, and awareness. Thus, the actor's involvement and his awareness, rather than being in opposition to each other, are in accord.

The actor's sense of truth is also involved in another major area of the actor's training—his work with actions (the way he behaves physically on the stage), sometimes called the "business" of the actor. Some idea must supply an incentive or intention to pull together what could otherwise be a series of disconnected and unrelated physical deeds. Some purpose, some aim must motivate the actor's will and energy. Any performance thus may be seen as a series of actions—as the score of the play—which must be carried out not simply physically but logically and truthfully. They must accomplish their purpose anew each night at every performance rather than merely repeating the external movements.

To develop spontaneity, to train himself to behave logically and truthfully, and to listen and respond to his partner, the actor practices improvisation—dramatizing contrived situations without a script. Improvisation is of enormous importance in the process of training and also of performance. It teaches the actor to speak rather than to read his lines, and it breaks his unconscious adherence to conventional theatrical patterns of behaviour. It forces him to use his senses and often to discover not only the logic but also the significance of a scene. It compels the actor to work creatively and prevents him from reverting to skillful but mechanical repetition.

By means of exercises that may be remote from the actual roles he plays—such as the "song and dance" (in which a song is rendered in various ways, each conveying strongly a particular feeling about it) or, conversely, the spoken "inner monologue" (in which the actor recalls an important personal experience in minute detail) or others—the actor not only intensifies his capacity for experience, but also frees his blocked, or inhibited, impulses. He is enabled to deal with his own subjection to automatic habitual forms of behaviour and mannerisms and to acquire new means of expression, corresponding to the true nature and strength of his impulse. He is thus enabled to develop a much larger scale of expression than the customary casual means of behaviour in life permit.

Voice and body training

The basic means of the actor, which have traditionally served as the primary area of his training, are voice and body gesture. The methods used to train these tools of the actor derive from other fields, such as from the training of the singer's voice and of some forms of dance and pantomime. These contain many useful exercises for the strengthening of the respective muscles of the voice and body. But the singer's voice must retain its beauty and its musical quality regardless of the circumstances involved, and the dancer's body must always wind up in a formally acceptable attitude; and their respective training is so de-

signed. Their needs are quite different from the needs of the actor. His voice must be flexible and expressive of all situations and experiences. It must be able to deliver a "poor" voice or a vulgar, rough, angry, or harsh voice. It must vary as much as the events to be created. His attitudes must be those of the character—of a human who may be ill at ease, slovenly, awkward, debilitated, or natural—giving no indication that it is being accomplished by a skilled craftsman. The technical accomplishment in the singer and in the dancer may represent a large part of what is appreciated in their performances; but in the actor, the very fact of the accomplishment must remain hidden. Technical accomplishment should go unnoticed by the audience.

For this reason, many actors have expressed dissatisfaction with the accepted methods for training the actor's voice, speech, and body. In this respect, the work of Grotowski and of producing units throughout the world that have been influenced by him have demonstrated new possibilities. The skill of the actors in such groups has been generally admired, and their work could serve as a basis for training the actor. Relaxation, concentration, work on objects, sense memory, emotional memory, action and improvisation, and the training of voice and body constitute the actor's work on himself. Work with objects and sense memory are as fundamental to his art as are daily finger exercises to a pianist.

THE ACTOR'S APPROACH TO HIS ROLE

Stanislavsky suggested that the actor, in approaching his work on a scene, ask himself four questions: (1) who he is (character), (2) where he is (place), (3) what he is doing there (action and intention), and (4) what happened before he came there (given circumstances). The answers to these questions provide the actor with the necessary background for his performance, helping him to create the scene. In approaching the play in its entirety, the actor must subject his role to more intense analysis: he must search for the spine, or the kernel, of the play as well as its division into separate sections or units of actions. He must discern the beats of the play (*i.e.*, the smallest units of dramatic action into which each role can be divided) as well as the rhythms of the play as a whole, and he must determine what adjustments must be made in his performance for each of the other characters. For some plays an additional element is necessary: the overall mood, or pervading texture, that surrounds the play or out of which the play stems. The attempt to determine it, however, may lead to an excess of verbal and mental gymnastics that are of little actual value, unless the actors have been trained in the proper procedures. The actors must act out the elements involved in the analysis in order to receive any concrete benefit from it; otherwise it may remain superficial or merely intellectual.

Collaboration with the director

Another area deserving attention is the rehearsal process. This is primarily the time in which the director's conception of the play must be harmonized with those of the actors; it is of immense importance that the actor approach the rehearsal in a creative frame of mind, ready to enlarge both his own and his colleagues' interpretations. Without a logical sequence of rehearsals, the actor's creativity cannot be properly stimulated. Without an understanding of the psychology of the rehearsal procedure, much of the work of the actor and the director may be defeated in production. There are, for example, significant possibilities in the reading rehearsal, in which the actors, usually seated in a circle, read aloud from the script and discuss its meanings as they proceed through it. There is enormous value in improvisation, when it is understood and used correctly. The relation between the individual actor and the ensemble is welded during the rehearsals, and they are the proper time to encourage the actor to begin to develop his movements to block the scene, to memorize his lines.

Styles of performance. In an effort to bring new life to plays of the past and present and to advance the imaginative possibilities of theatre, there has been a rediscovery of style in the 20th century. Style is the attribute of any complete achievement; it is not, as is often assumed by laymen, merely the manners and customs of a particular period. Such manners may be more strikingly elegant compared with those of the present, but they remain only manners. The Elizabethan form of theatre had conflicting styles within it, judging from a description of them in *Hamlet*, and so did the Greek and the French classical theatre. Even in the Japanese Kabuki and Nō theatre there have been conflicts of styles like those in Western theatre.

Style is not, as is sometimes assumed, the opposite of realism. Neither is it necessarily characterized by an expansiveness or broadness in acting. Style is the angle from which reality is observed. It is an attribute of all creative activity—not just of period or classic plays. The search for the specific content and reality of a play leads to style. The search for style in itself or in the traditions of the past often leads to empty forms.

Just as style should not be identified with a particular period, neither should it be associated with specific playwrights. Such terms as a Shakespearean or a Chekhovian style actually refer to the theatrical conventions traditionally associated with those dramas—a rhetorical and "larger than life" manner in the first and a static "mood" in the latter. These elements are little related to style; otherwise great Shakespearian and Chekhovian productions could be re-created generation after generation in precisely the same way. The fact is that those dramas must be continually re-created from the new views of each emerging generation. Even the performances of Jean Racine's dramas re-created in the correct manner of French classical theatre do not attain great success in France. It is possible that more might be achieved by an interpretation of Racine along Chekhovian lines: the use of methods associated with the productions of Chekhov combined with a fresh understanding of Racine's character development might revise the opinion that his plays are excessively wordy at the expense of reality. This criticism was also levelled against Chekhov before Stanislavsky's Moscow Art Theatre demonstrated the inner logic of his plays.

Highly regarded theatrical performances

The most highly regarded performances in theatrical history were achieved by creating believable characters behaving logically and truthfully. The actors often used means that were criticized as being vulgar—either too naturalistic or too realistic—because they broke the formal patterns of poetic speech. These actors demonstrated little reliance on the formal elements that traditionally have been supposed to characterize style. French actors, who have been considered exemplars of this traditional notion of style, have not excelled in Shakespeare; this perhaps is an indication of the defects inherent in an approach based on the external concept of style.

The term style is often used incorrectly in reference to the theatrical conditions that simulate the original concept, structure, and dynamics of a play. The rediscovery in the 20th century of the Shakespearean stage, for example, led to a new quickness and fluidity, a nearly cinematic technique, in presenting Shakespeare's plays, but these techniques should not be interpreted as the original and therefore correct style of production. Shakespeare continues to be presented in a vast range of styles.

Plays that entail interaction between actor and audience also present no unique problems to the actor. The audience is treated with the same conviction and reality as if it were another presence in the drama, as if it were simply another character.

Acting in different media. The fundamentals of the actor's art remain the same no matter how bizarre the dramatic context: the actors may be abstractions, for example, as in Stanislavsky's 1908 production of Maurice Maeterlinck's allegorical fantasy *The Bluebird;* they may play a band of actors producing a play, which they then proceed to perform in a vivid theatrical fashion, as in Vakhtangov's production of *The Dybbuk*, a play about an evil spirit, by S. Ansky; they may imagine themselves into a satiric extravaganza, as in experiments by the Group Theatre (a major U.S. troupe of the 1930s, working on the basis of Stanislavsky's ideas) based on the satirical pictures of George Grosz, (1931); or they assume the

distorted attitudes appropriate to an expressionist world, as in the classic horror film *The Cabinet of Dr. Caligari* (1919).

Acting in motion pictures and television

The growth of motion pictures, especially the rise of the "talkies," beginning in 1927, greatly affected acting, as theatre talent was diverted from the stage. The requirements of acting in motion pictures, television, theatre, and opera are basically the same, although some of the techniques are different. It is possible to put strips of film together and create a performance that never was actually given. The performance is created by the director rather than by the actor. There have been performers in motion pictures who were thus completely products of the camera and contributed little from an acting point of view, depending rather on their physical charms and personality. Others, however, were authentic actors, who developed a style perfectly suited to the medium; Charlie Chaplin, for example, ranks as one of the greatest actors of all time in any medium.

Despite the technical demands that are unique to each medium, the properly trained actor moves easily from one medium to another without any diminution of his talent. Those who have been trained in the rhetorical and theatrical gesture approach, as many British and French actors have been, sometimes find difficulty in making the transition to films. The theatre tends to diminish action and voice, requiring a heightened intensity to project across the footlights. The camera, however, exaggerates action and emotion. Moreover, some actors find it difficult to perform scenes out of sequence, as is usually done in films, and for other actors the closeup can be intimidating. But the fact is that actors training for films usually use the same exercises as theatrical actors—working with imaginary objects and partners, performing appropriate physical and psychological tasks, and others. Moreover, most of the foremost actors in mid-20th century, such as Sir Laurence Olivier, Jean-Louis Barrault, and Katharine Hepburn, have been outstanding in both film and theatre.

The contemporary theatre is characterized by many plays that demand more dynamic and more imaginative physical actions of the actors than previously and that utilize a diversity of audio-visual effects and multimedia devices. Under the need to fulfill these demands, acting could easily revert to its old-fashioned externalized forms. In addition, the development of repertory theatres in the United States, England, and elsewhere, with their eclectic repertoires and their combinations of contemporary and classic plays, often leads to a search for meretricious "style" rather than for genuine content. These new pitfalls may be avoided, however, in much the same way as those that faced the actor in previous epochs, by understanding of the true fundamentals of the art of acting.

BIBLIOGRAPHY. The literature on actors and acting is overwhelming. Most of it is of little value. Significant are the statements of actors or theatre practitioners that must form the basis of our understanding. The best anthology in any language with useful notes and bibliography is TOBY COLE and H.K. CHINOY (eds.), *Actors on Acting: The Theories, Techniques, and Practices of the Great Actors of All Times As Told in Their Own Words*, rev. ed. (1970). The most valuable statements are those of Luigi Riccoboni, F.J. Talma, and William Gillette. Some of these are reprinted complete in *Papers on Acting*, ed. by BRANDER MATTHEWS (1958), which also includes the polemic between Henry Irving and C. Coquelin. A useful introduction to the literature is EDWIN DUERR, *The Length and Depth of Acting* (1962), flawed by the common inability to relate theory to practice. The famous essay by Diderot, in translation, *The Paradox of Acting*, is printed with WILLIAM ARCHER'S reply in *Masks or Faces* (1957). From the critical point of view the essay by PIERRE REMOND DE SAINTE-ALBINE, *Le Comédien* (1747; Eng. trans., *The Actor*, 1750 and 1755, reprinted 1969 with additional anecdotes of Garrick and other English actors), is still valuable. Important contributions are G.H. LEWES, *On Actors and the Art of Acting*, (1957); G.B. SHAW, *Our Theatres in the Nineties*, 3 vol. (1932); STARK YOUNG, *The Flower in Drama* (1923) and *Glamour* (1925). Basic to modern understanding are the works of STANISLAVSKY: *An Actor Prepares* (1936, reprinted 1956), *Building a Character* (1949), *Creating a Role* (1961). The English versions are well translated and edited by ELIZABETH REYNOLDS HAPGOOD. Complete versions with the useful notes of the Russian editions are available in German translation and are essential for serious study. A handy though somewhat superficial description based on secondary sources of the background surrounding his endeavour is CHRISTINE EDWARDS, *The Stanislavsky Heritage* (1965). Useful also are ROBERT LEWIS, *Method or Madness?* (1958); the essay by LEE STRASBERG on "Acting and the Training of the Actor," in JOHN GASSNER, *Producing the Play* (1941; rev. ed., 1953); ROBERT HETHMON (ed.), *Strasberg at the Actors Studio* (1965); RICHARD BOLESLAVSKI, *Acting: The First Six Lessons* (1949, reprinted 1962); and TOBY COLE, *Acting: A Handbook of the Stanislavski Method* (1955). For a proper appreciation of Stanislavsky's approach, the work of his pupil E.B. VAKHTANGOV is essential. A brilliant description is in NIKOLAI GORCHAKOV, *The Vakhtangov School of Stage Art* (1961; orig. pub. in Russian, 1957). The last period of Stanislavsky's work is unfortunately not available in English but is described in the German translation of WASSILI TOPORKOW, *K.S. Stanislawski bei der Probe* (1952). Recent contributions on acting are JERZY GROTOWSKI, *Toward a Poor Theatre* (1968); ANTONIN ARTAUD, *Le Théâtre et son double* (1938; Eng. trans., *The Theater and Its Double*, 1958); JOHN WILLETT (trans.), *Brecht on Theatre* (1964); MICHEL SAINT-DENIS, *Theatre: The Rediscovery of Style* (1960); VIOLA SPOLIN, *Improvisation for the Theater* (1963); and MICHAEL CHEKHOV, *To the Actor* (1953). The relation of music to theatrical declamation is touched on by ROMAIN ROLLAND in his essay on "Lully's Recitative and Racine's Declamation" in *Some Musicians of Former Days* (1915). Another important viewpoint relating the actor to his environment is F.S. BOAS, "Edmund Kean in His Heroic Parts," in *From Richardson to Pinero* (1936). This approach has recently been attempted on a larger scale in J.W. DONAHUE, *Dramatic Character in the English Romantic Age* (1970). No history of acting can be written without a knowledge of what acting consists of and the creative processes involved. German students have in many specialized studies and dissertations tried to formulate methods for studying the actor's work by examining critical descriptions, stage directions, and iconographical material. A bibliography is available in HANS KNUDSEN, *Methodik der Theaterwissenschaft* (1971).

(Le.S.)

Actinide Elements and Their Compounds

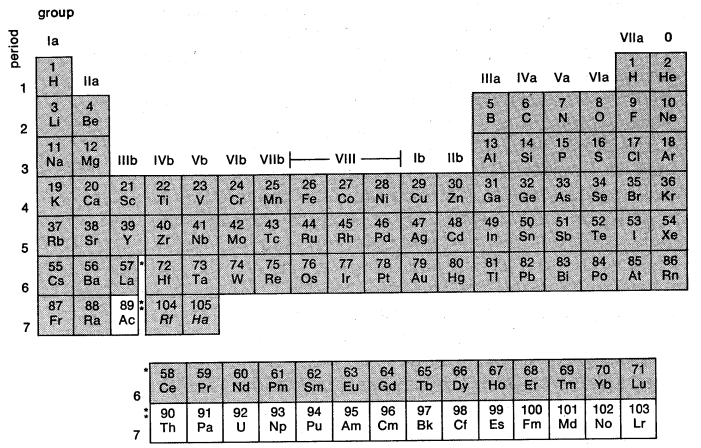
The actinide elements constitute a group of 15 consecutive elements in the periodic table from actinium, symbol Ac, with atomic number 89, to lawrencium, symbol Lr, and atomic number 103. As a group they are significant largely because of their radioactivity. Although several members of the group, including uranium (the most familiar), occur naturally, most are man-made. Both uranium and plutonium have been used in atomic bombs for their explosive power and currently are being employed in atomic-generating plants for the production of electrical power.

PROPERTIES OF THE ACTINIDES AS A GROUP

General comparisons. The properties of all of the actinide elements are quite similar because of the similarity of the arrangements of electrons around the nuclei of their atoms. In any atom the total number of electrically negative electrons equals the number of positively charged protons in the nucleus, thus keeping the atom as a whole electrically neutral. The periodic table is a catalogue of the elements so arranged that the atoms of any one element have one more proton that those of the element immediately before it in the succession. This means that the atom of each successive element must have one more electron to balance the one additional proton and keep the atom neutral (see ATOMIC STRUCTURE). The increasing number of electrons are arranged around the nucleus in concentric shells and subshells. These shells are identified by ascending numbers counting outward from the nucleus, 1, 2, 3, etc. The subshells, called orbitals, are identified by shell number and a letter (*s*, or *p*, or *d*, etc.). Each electron has a specific energy that is defined in terms of four energy quantum numbers. In a general discussion, however, it is enough to identify an electron as 1*s*, or 3*p*, or 5*d*, etc. There is a limit to the number of electrons that any shell may contain, and no atom may have more than eight electrons in its outermost shell. Such an octet arrangement is extremely stable, and the next element in the series must begin a new shell outside the octet. This means that shells that can contain a

maximum of more than eight electrons become filled later as inside shells. In the heaviest atoms, three such incomplete inner shells exist. The chemical properties, and many of the physical properties, of the elements depend almost wholly on the activity of the electrons in the outermost shell. Thus, similar configurations of the outermost shell in different elements give those elements similar properties. Electrons from inner shells may also participate in chemical activity, but their behaviour is of considerably lesser importance. All the electrons that may be active in chemical bonds are called valence electrons.

Atomic structure. The actinide elements follow one another in the seventh series of the periodic table (see the Figure). Each has 86 electrons arranged as in the atoms



Actinide elements in the periodic table of the elements.

of the noble gas radon (which precedes actinium by several spaces in the table), with three more electrons that may be positioned in the 6*d* and 7*s* orbitals (the seventh shell is outermost), and with additional electrons packing into inside orbitals. Specifically, the series is formed by the insertion of one more electron for each successive new element into an underlying 5*f* orbital. The valence electrons, however, are found mainly in the 6*d* and 7*s* orbitals. Thus, the chief difference among the atoms of the elements of the series is the single additional electron deep within the electron cloud; but the fact is that, because of its position in the 5th shell, this distinguishing electron actually affects the chemical properties of the actinides only in a relatively minor way; 5*f* electrons are not involved in the formation of chemical bonds with other atoms.

As is usual with the elements of any group, there are a number of exceptions to these generalities, particularly in the lower members of the series; but for most of these elements, the concept of a series of nearly identical actinide elements is a useful guide for predicting their chemical and physical properties.

Like all elements, each actinide has its own particular atomic number, equal to the number of protons in the nucleus and, consequently, to the number of electrons; at the same time the atoms of an element are capable of existing in a number of forms (isotopes), each of which has a different number of neutrons in its nucleus and hence a different atomic mass. Although isotopes of a given element behave alike chemically, they may have different stabilities in relation to radioactive decay, which is a property of the nucleus. No element beyond bismuth in the periodic classification—*i.e.,* no element that has an atomic number greater than 83—has any stable isotopes; radioactive isotopes of every element in the table can be produced in the laboratory. The actinides are unusual in forming a series of 15 elements having no stable isotopes; every actinide isotope undergoes radioactive decay, and, as a result, only a few of the lighter, more stable members of the series (such as thorium and uranium) are found in nature. The half-life, or the precise time required for one-half of any amount of a particular isotope to disappear due to radioactive decay, is a measure of the stability of that isotope. The naturally occurring isotopes in the actinide series have long half-lives, of the order of billions of years.

Usefulness. The practical significance of the actinides arises from the fissionability, or potential for splitting, of certain of their isotopes. When an atomic nucleus breaks apart, or undergoes fission, a far more disruptive process than ordinary radioactive decay, enormous amounts of energy are liberated. This energy can be allowed to generate an atomic explosion, or it can be controlled and used as a fuel to generate heat for the production of electrical power. Nuclear processes for power production give off no smoke, smog, or noxious gases, as conventional coal- or gas-fueled plants do. Nuclear power plants, however, do yield waste heat that may be considered as thermal pollution, and they also yield useless and dangerous radioactive wastes which, although they are pollutants, may be less undesirable than those from fossil-fuel generators. For this and other reasons, such as economy of operation, there is a potential for an enormous electrical energy production inherent in atomic energy generating technologies, and, since the actinide elements are the only known fissionable materials, the practical impact of their availability is great. The isotope of uranium with the atomic number 92 and mass 235, written as uranium-235 or, in chemical symbols, as ^{235}U, is present to the extent of only 0.7 percent in ordinary uranium, but it is a necessary fissionable material in the operation of an atomic reactor using natural uranium. Other fissionable isotopes of great importance are uranium-233, plutonium-239, and plutonium-241.

Nuclear fission

Fissionable plutonium isotopes are formed as by-products of fission in reactors using uranium; when neutrons are added to uranium-238, which is not itself fissionable, it is converted to the fissionable isotope plutonium-239. Thorium, also, is potentially of great economic value, because one of its isotopes, thorium-232, can be converted into the fissionable isotope uranium-233 in an atomic breeder reactor (*i.e.,* one that produces more fissionable material than it consumes), thus increasing by many times available supplies of fissionable materials. Since thorium is about three times more plentiful than uranium, the potential use of this element to produce atomic energy is significant.

The heavier actinides, those beyond plutonium in the periodic table, are of interest principally to research scientists, though they have some practical uses as sources of thermoelectric heat and neutrons. They are employed to some extent in cancer therapy.

Actinium, thorium, protactinium, and uranium are the only actinide elements found in nature to any significant extent. The remaining actinide elements are all man-made by bombarding naturally occurring actinides with neutrons or with heavy ions (charged particles) in particle accelerators (such as cyclotrons). The actinides beyond uranium do not occur in nature (except, in some cases, in trace amounts), since the stability of their isotopes decreases with increase in atomic number and whatever quantities may be produced decay too fast to accumulate. The half-life of uranium-238, the most stable uranium isotope, is 4.5×10^9 years. Plutonium-239 has a half-life of 24,400 years and is produced in reactors in ton amounts; but nobelium and lawrencium, elements 102 and 103, with half-lives of seconds, are produced a few atoms at a time. The first of these synthetic actinide elements to be discovered (1940) was neptunium, symbol Np, atomic number 93, which was prepared by bombardment of uranium metal with neutrons (see TRANSURANIUM ELEMENTS).

Chemical properties. The chemistry of any element can be understood best in terms of atomic structure and its effect on the formation of chemical bonds. Three general types of chemical bonds can be formed: ionic, in which valence electrons are shifted from the less to the more electronegative atom, producing oppositely charged ions, which then associate by virtue of their mutual attraction; covalent, in which a pair of electrons is shared by each atom; and metallic, in which the arrangement of

atoms allows the valence electrons to move about in the structure. These bonded atoms may be represented in several ways for purposes of chemical discussion; as molecular formula, structural formula, general formula, etc. In each case, an atom is signified by the symbol of the element, followed by a subscript that indicates the actual number of those atoms in the bonded structure. For example, water molecules are combinations of two hydrogen atoms and one oxygen atom; the molecular formula for water is H_2O. Hydrogen peroxide, with two oxygen atoms and two hydrogen atoms, has the formula H_2O_2. In a general formula, representative letters for groups of elements are used instead of the symbols for actual elements, and the letter *n* is used instead of a number when the numbers of atoms in the molecule are defined in terms of ratios. For example, any one alkali metal combines with any one halogen atom; therefore, the general formula for alkali halides is MX, M representing lithium, sodium, potassium, etc., and X representing fluorine, chlorine, bromine, etc.

There are two series of elements (actinide and lanthanide) in the periodic table (see the Figure) in which added electrons (with increasing atomic number) go into internal *f* orbitals, where they are partially buried and consequently not chemically active. These two series occur in Group IIIb of the table; because the outer, or valence, electrons of these elements are much the same, the chemical properties of the elements in the two series tend to resemble one another closely. A great deal is known about the lanthanides, all but one of which occur in nature as stable isotopes and, therefore, predictions about the chemistry of the actinides, some of which can be prepared only in minute quantities, can be made with some success by comparing their electron structures with those of the lanthanides. In the lanthanide series of elements, with atomic numbers from 57 to 71, and found in the periodic table directly above the actinides, each added electron goes into the *f* orbital of the fourth shell; these orbitals are designated as 4*f*. In the actinide elements, the added electrons also go into an *f* orbital in a similar manner, but in the fifth shell, instead. Electrons with larger quantum numbers generally are farther from the nucleus than those with smaller quantum numbers, and are therefore usually less strongly held by it. As expected, then, electrons in the 5*f* orbitals, being farther from the nucleus, are much less tightly bound than are those in 4*f* orbitals and, in fact, sometimes are active enough to take part in chemical reactions. The result is that the actinide elements, in which the 5*f* orbitals are being filled, have more variable valences (number of electrons available for chemical bonds) than do the lanthanides, in which the 4*f* orbitals are being filled.

Similar lanthanide and actinide series

The similarities between many lanthanide and actinide compounds are striking and offer a useful comparison. Under certain conditions, for example, actinium, americium, curium, and berkelium metals have the same crystal structure, as do many of the lanthanides. It is not yet known whether californium or einsteinium, the only other actinide elements with sufficiently stable isotopes for chemical work, have the same structure. Several of the lighter actinide elements from thorium through plutonium have different and unusual metallic structures, presumably due to the mixing of 5*f* and 6*d* orbitals in their atoms, some electrons entering unfilled 6*d* orbitals rather than the expected 5*f* orbitals.

The actinides generally show multiple oxidation states. When an element combines with others to form a compound, the number of electrons of its atom that are, or can be, involved in the bonds of that particular compound is called the oxidation state of the element. An element may have more than one possible oxidation state, and these are designated by the number of electrons involved, and by a plus or minus sign to indicate the role they play. Compounds of americium and californium with an oxidation state of +2 are known; there are reasons for expecting the existence of this state in some of the elements heavier than californium. For example, spectroscopic evidence for einsteinium(II) in the presence of the fluoride ion has been obtained. (Roman numerals following the name or symbol of an element are used to designate the oxidation state of the element.) Divalent actinides (that is, actinides in the +2 oxidation state) form compounds with nearly the same properties as those of the divalent lanthanides and, accordingly, iodides, bromides, and chlorides of divalent americium and californium have been found to be stable. If X symbolizes the nonactinide and M symbolizes the actinide element, then the general formula for these compounds would be one atom of actinide and two atoms of nonactinide: MX_2.

The +2 oxidation state

Great similarities in chemical behaviour are found in the actinides of oxidation state +3, from actinium to einsteinium; furthermore, they are much like the lanthanides of the same oxidation state. The crystal types and many physical properties of these trivalent actinides are dependent more on the size of the +3 ion (an atom that has given up three electrons and has become an ion with three positive charges, symbolized as Ac^{+3}, etc.) of the particular element that is involved. For instance, the solubility of the trifluorides formed by actinides with a +3 state (thorium and protactinium have no such state) is exceedingly low. The crystal-structure type for the actinide trifluorides is the same as that of lanthanum trifluoride, and, since the radius of the ion is a regular function of the atomic number, the circumstance allows extrapolation from the lanthanum compound to the actinide compound and interpolation between known compounds in the series to determine missing values. The hydroxides, phosphates, oxalates, and alkali double sulfates of the actinides are also insoluble, with many of each having identical crystal structures, or being isostructural. The chlorides, bromides, and iodides (*i.e.*, the halides) of the actinides are, for the most part, isostructural for any one halide, and the structure type can be predicted from a knowledge of the ionic radius. The solubility of these halides in water is generally great. The +3 oxides of actinides are also isostructural, with the general formula M_2O_3, in which M is any of the actinide elements; they form cubic (or hexagonal) crystals, and the sizes of the molecules are thus easily predictable. Generally, then, the chemistry of the actinides in the +3 oxidation state is similar, with the differences mainly due to ionic size. As a consequence of these similarities, separations of the elements and of their components are frequently difficult, necessitating the use of methods in which very slight physical differences of the atoms or ions serve to separate the chemically almost identical materials. Two methods are ion-exchange reactions, in which differences in ions and bonding are used to effect separation and solvent extraction, in which specific solvents are used to dissolve and withdraw from the mixture the desired element or its compound.

The +3 oxidation state

Actinides in the +4 oxidation state also are much alike (and also resemble the +4 lanthanides). The +4 actinides, thorium, protactinium, uranium, neptunium, plutonium, berkelium, and, to a lesser extent, americium and curium, are sufficiently stable to undergo chemical reactions in water solutions. Crystallized compounds in the +4 state exist for thorium, protactinium, uranium, neptunium, plutonium, americium, curium, berkelium, and californium. The oxides and many complex fluorides are known for all of these elements. The dioxides are all isostructural, as are the tetrafluorides. Most of these actinide compounds can be prepared in a dry state by igniting the metal itself, or one of its other compounds, in an atmosphere of oxygen or of fluorine. Some tetrachlorides, bromides, and iodides are known for thorium, uranium, and neptunium. The ease with which they can be formed decreases with increasing atomic number. Berkelium(IV) appears to be sufficiently stable to allow the preparation of the tetrachloride. Hydroxides of a number of these elements in the +4 state also are known; they are of very low solubility, as are the fluorides, oxalates, and phosphates. Again, many physical properties of the tetrafluorides are influenced more by ionic size than by atomic number, and isostructurality of these actinide and lanthanide compounds is the rule rather than the exception.

The +4 oxidation state

The similarities exhibited by the lanthanide and actinide compounds in the +3 and +4 oxidation states, as well as in some cases by the free elements, can be very useful. A great many individual differences, however, do arise. These are due, in part, to mixing of the orbitals (some electrons moving into *d* rather than *f* orbitals) and, in part, to the relative degrees of binding of the *f* electrons.

Special properties. *Chemical.* The metals thorium, protactinium, uranium, neptunium, and plutonium are for the most part different from one another. Uranium, neptunium, and plutonium have extremely dense metallic forms. Neptunium, for example, with a density of 20.45 grams per cubic centimetre when crystallized into the orthorhombic crystal form, is one of the densest metals known to man. A possible explanation for the fact that these metals show a number of different crystal forms is that the electrons in the 5*f* orbitals mix with those in the 6*d* orbitals and consequently form a number of hybrid electronic states of nearly the same energy. Beginning with americium, however, the electron energy levels seem to be sufficiently separated so that mixing does not occur.

The other main differences shown by the actinides is that some possess the +5, +6, and +7 oxidation states (no lanthanide element exceeds the +4 state). It appears that the 5*f* electrons of the actinides, being far enough from the positively charged nucleus, permit increasingly easier removal and consequent formation of higher oxidation states. The element protactinium shows the +5 state; uranium, neptunium, and americium exhibit the +5 and +6 states; only neptunium and plutonium have the +7 state.

There are two types of chemical reactions for the +5 and +6 states. If M symbolizes any actinide, and O, as usual, symbolizes oxygen, then the ions found both in aqueous (water) solution and in solids prepared from solution are represented by the general formulas MO_2^+ (meaning a molecule consisting of one atom of M with two of oxygen, the whole having a single positive charge) and MO_2^{2+}. In nonaqueous solution, and in solids prepared from them, compounds of M that do not contain oxygen are known. With the halogens (X being a general designation for a halogen—fluorine, chlorine, bromine, or iodine), compounds are known that can be represented as MX_5 (meaning a molecule consisting of one atom of an actinide with five atoms of a halide) and MX_6, as well as complexes of the type having the molecular formulas MX_6^-, MX_7^{2-}, and MX_8^{3-} for the +5 states, and MX_7^- and MX_8^{2-} for the +6 states. Neptunium(VII) and plutonium (VII) have been prepared in basic solution, and certain oxygenated ions (of the type represented by MO_5^{3-}) as well as a few solid compounds, have been identified with the same oxidation state. Complex oxides with alkali metals in which these two elements have the +7 state also have been prepared.

Physiological. All the actinide elements are heavy metals and, as such, are toxic, just as lead is toxic—relatively large amounts ingested over a long period cause serious illness. But, with the exception of the long-lived thorium and uranium isotopes, the real danger with the actinide elements lies in the radioactive properties of these elements. They are emitters of tissue-destroying and cancer-producing rays (alpha, beta, or gamma; see RADIOACTIVITY). Furthermore, the chemistry of many of these elements is such that, once ingested, they tend to remain in the body almost indefinitely. Several, such as plutonium and americium, migrate to the bone marrow, where their radiation interferes with the production of red blood cells. As a consequence, workers using these elements are required to take elaborate precautions to prevent ingestion. Less than one-millionth of a gram can be fatal.

Radioactive decay series

Nuclear. All of the actinide elements are unstable toward radioactive decay; the reason that actinium, thorium, protactinium, and uranium are found in nature at all is because some of their isotopes are unusually stable and others are being formed constantly by decay of the long-lived isotopes. It is convenient to divide the naturally occurring isotopes into families based on the relationships of their atomic masses to each other. The mass numbers of all isotopes of the so-called thorium series, for instance, turn out to be multiples of four, and the series is known as the 4n series. In the uranium series, masses have been shown to be such that they are represented by 4n + 2; and in the actinium series, by 4n + 3. Not found in nature to any significant extent, but which can be artificially produced by man, is the neptunium, or 4n + 1 series, named for its longest-lived member, neptunium-237.

In the various radioactive decay processes, several "rays" are emitted, the term ray being a holdover from the time when all of these emissions were thought to be rays. They are: (1) electrons, called beta particles to indicate their origin in radioactive decay, and designated as negative beta, or β^-, particles; (2) helium nuclei, called alpha particles and designated as α particles, or as helium with a plus two charge, He^{+2}; (3) gamma rays, which are electromagnetic waves of very high frequency, designated as γ rays; (4) positrons, which are positively charged electrons and are designated as positive beta, or β^+, particles. Finally, an orbital electron in a radioactive atom may be captured by the nucleus and taken into it. This radioactive event is called K-capture. Except for the emission of gamma rays, each of these processes leads to an isotope of a different element; that is, to a substance with a different atomic number. The emission of an alpha particle leads to a change in atomic weight as well, because this emission has significant mass. The changes occur in sequence, each decay process leading to still another unstable element, until a region of stable isotopes of the elements lead and bismuth is reached.

The most important nuclear reactions, however, involve the capture of neutrons by an actinide nucleus, followed by splitting, or fission, of that nucleus into two unequal parts, with the liberation of enormous quantities of energy plus two or more extra neutrons. Atomic reactors and atomic bombs depend upon the chain reaction set up by this process: the resulting neutrons react further, inducing more fission reactions, which produce more neutrons, which lead to still more fission reactions, with the result that, without control, a great deal of energy is liberated very quickly.

One gram of burning coal yields fewer than 10,000 calories of heat. The fission of one gram of uranium-235 produces 2×10^{10} calories, or about 2,000,000 times as much energy. With adequate control the energy can be released in useful form.

The most important actinide by far, because of its fissionability, is uranium, which has several isotopes. Natural uranium consists mostly of uranium-238, a non-fissionable isotope; uranium-235 occurs to the extent of only seven-tenths of one percent; and there are even smaller amounts of uranium-234.

Fissionable isotopes

The fissionable isotope uranium-235 is the valuable substance that causes the chain reactions with neutrons. Methods that separate isotopes by virtue of their different masses are used to eliminate the unwanted isotopes in the naturally occurring mixture of uranium ores and thus enrich natural uranium with respect to uranium-235. Such enriched uranium was used in the first atomic bomb exploded over Japan.

Two other fissionable actinide isotopes are important and undoubtedly will be more so in the future. The first is the isotope plutonium-239 prepared by neutron bombardment of uranium-238. First, uranium-239 is formed by neutron capture; this isotope then undergoes negative beta decay, giving neptunium-239, which in turn reacts by negative beta decay to give plutonium-239. Plutonium-239 thus is formed as a by-product in uranium piles in which uranium-238 is exposed to neutrons; it can be used in place of uranium-235 in atomic weapons or in reactions. The second fissionable isotope is potentially even more important, because it is produced from the element thorium, of which there are enormous reserves on Earth. When the isotope thorium-232 is bombarded by neutrons, one of which it captures, it is converted to thorium-233. This substance decays by

beta emission to protactinium-233, which again emits a beta particle to give uranium-233, also a fissionable isotope of uranium. Since uranium is a relatively scarce element, development of atomic power on a large scale is expected to deplete rapidly the uranium that can be produced economically. The use of thorium thus could extend the supplies of fissionable material by about threefold, and the use of plutonium as well could more than double it. Breeder power reactors are designed so that very few neutrons are lost through the surface or from absorption by impurities, and more fissionable material (either uranium-233 or plutonium-239) is produced than is consumed. Should such breeder reactors be made practicable, never-ending supplies of fissionable isotopes would be available.

Thermoelectric power generators

Although no other fissionable isotopes of the actinides occur in significant amount, there are practical uses of the nonfissionable isotopes due to the amounts of heat they produce by nuclear decay. For special power sources that require great dependability, thermoelectric generators using these isotopes can be considered. Alpha-emitting isotopes with radioactive half-lives in the range of several months to 100 years or more are suitable candidates. (Beta- or gamma-emitting isotopes require too much shielding to be usable.) Curium-244, which has a half-life of 17.6 years, produces 2.9 watts of heat per gram. Another isotope, plutonium-238, produces 0.57 watts per gram and has the much longer half-life of 86 years. Such power sources have been planted on the Moon to provide electrical energy for transmitting messages back to Earth. Plutonium-238 also is being considered for use as a power source for artificial hearts; it already has been implanted in the human body to power a heart pacemaker, which is not expected to need replacement for at least ten years.

COMPOUNDS OF THE INDIVIDUAL ACTINIDE ELEMENTS

Actinium. Actinium, atomic number 89, exhibits only a single oxidation state of plus three, and the compounds formed are the hydroxide, oxalate, fluoride, phosphate, sulfide, oxide, and the double sulfate salt with any one of the alkali metals, all of slight solubility. The other halides and oxyhalides also have been prepared. Because even the stablest isotope of actinium, actinium-227, decays to elements that are strongly radioactive (thorium-227 and francium-223), comparatively few chemical studies of actinium have been made.

Thorium. Thorium, atomic number 90, has two *6d* and two *7s* electrons outside the radon electron core; it readily loses these four electrons when it is dissolved or when it combines with other atoms. Thus, +4 is its only oxidation state. Although compounds have been prepared in which thorium atoms and iodine atoms are present in the ratios of 1:2 and 1:3, these are not ionic compounds but semiconducting substances consisting of thorium(IV), iodine, and electrons (see SEMICONDUCTORS AND INSULATORS, THEORY OF). Similar compounds of thorium with nitrogen, sulfur, selenium, carbon, and other elements are also known.

The chemistry of thorium(IV) is that of the thorium ion, Th^{4+}. Insoluble oxides, peroxides, hydroxides, fluorides, phosphates, and oxalates are known. All of the halides have been prepared by conventional techniques, and, with the exception of thorium fluoride, ThF_4, all are soluble in dilute acids. Many double fluoride salts of the alkali and alkaline-earth metals with thorium have been prepared, and a very complicated series of such complex compounds with many different crystalline forms is known.

Thorium metal can be alloyed with relatively few other elements because its atomic diameter, 3.6 Å (1 Å, or angstrom unit, equals 10^{-10} metre), is too large for it to fit into their crystalline structures. It does form intermetallic compounds with aluminum, beryllium, cobalt, copper, iron, magnesium, manganese, mercury, nickel, zinc, and certain other elements.

The compound of thorium produced in largest quantity is the nitrate, which has the formula $Th(NO_3)_4 \cdot xH_2O$, in which x may have various values. The production of mantles for incandescent gas lighting formerly consumed large amounts of this compound; when cotton impregnated with thorium nitrate is burned, the nitrate is converted into thorium dioxide, which emits a brilliant white light upon heating by a gas flame. The addition of a small proportion of thorium sulfate to the nitrate improves the mechanical strength of the dioxide; small proportions of cerium salts increase the brightness of the light produced.

Thorium dioxide, often called thoria, is the thorium compound commonly employed as a nuclear energy source in breeder reactors (see above *Usefulness;* see also NUCLEAR REACTOR). Dispersion of a small amount of thoria in nickel renders that metal unusually resistant to corrosion even at temperatures as high as 2,400° F.

A few organometallic compounds of thorium, such as the cyclopentadienyl derivative, are known, as are many coordination compounds (*q.v.*) with negatively charged organic ions, including those formed from ethylenediaminetetraacetic acid, 2,4-pentanedione, and the arsenic-containing dyestuff thorin. Although the thorium atom itself does not impart visible colour to its compounds, many of its organic coordination compounds are highly coloured, a fact that is utilized in analytical procedures for the detection and determination of thorium.

Protactinium. Protactinium, atomic number 91, has five electrons added to the radon electron structure. It could therefore have +3, +4, or +5 oxidation states. Experimentally, it is found that protactinium(V) is the stable species, although protactinium(IV) occurs under mild reducing conditions. Protactinium(III) is exceedingly difficult to obtain, and there is, in fact, reason to suppose it cannot exist at all.

Because of the high electrical charge of pentavalent protactinium, the ion Pa^{5+} does not exist as an individual particle; stable solutions of the element in this oxidation state, however, can be prepared in the presence of strong acids containing ions such as chloride, fluoride, sulfate, or nitrate, which are called ligands when they bond to the metal atom in a specific way (ligands donate electron pairs in forming covalent bonds; see COORDINATION COMPOUNDS). In the absence of such ions, which convert it into comparatively simple species, protactinium forms very insoluble, intractable macromolecules or polymers, in which repeating structural units are linked together in long chains. Complex protactinium fluorides have been tentatively identified in aqueous solutions containing fluoride ions, and compounds containing these complex ions have been isolated. All of the halides of protactinium(V) have been synthesized, as have the well-known pentoxide (molecular formula Pa_2O_5) and a stable oxyfluoride (Pa_2OF_8). Numerous complex compounds of the protactinium halides with the halides of the alkali and alkaline-earth metals also have been made.

The conversion of compounds of protactinium(V) to those of protactinium(IV) is relatively easy; the processes are examples of chemical reduction (see OXIDATION-REDUCTION REACTIONS). Tetravalent protactinium behaves as a normal +4 actinide. Its compounds, which are similar to those known for thorium, are formed by the reactions of hydrogen gas with solid compounds of protactinium(V) at 400° C (750° F). The dioxide (with a molecular formula of PaO_2), all of the tetrahalides, and a number of complex compounds have been prepared and identified.

Alloys of protactinium

Several alloys of protactinium, particularly those with uranium, have been prepared and studied. Little work has been done on organometallic compounds of protactinium, but there is every reason to believe that they will prove accessible.

Uranium. Uranium, atomic number 92, has six electrons added to the radon electron core. As might be expected, therefore, it has +3, +4, +5, and +6 oxidation states. The +3 state is easily oxidized to the +4 state. In aqueous solution, this uranium(IV) is further oxidized by oxygen in the air to uranium(VI), uranium(V) being unstable.

Uranium forms the compounds that would be expected for an element having the +3, +4, +5, and +6 oxidation states. It forms series of halides corresponding to

the general formulas (X representing any halide) UX_3 (meaning a molecule consisting of one atom of uranium and three atoms of a halogen), UX_4, UX_5, and UX_6, as well as complexes of these with alkali metal halides. The compounds of the +3 state are stable only as dry compounds; in water solution they are oxidized to the +4 state. The uranium(IV) compounds behave as predicted; *i.e.*, the hydroxide and fluoride are insoluble, and extensive hydrolysis of them occurs in dilute acid solutions. Many compounds of uranium(V) that do not contain oxygen are known, and are stable in the absence of water or oxygen. All halides of uranium(V), except the iodide, are known, as are many complexes with alkali metals. The same two classes of ions are found with uranium(VI). In water, the uranyl ion, UO_2^{2+}, occurs, and compounds containing this ion, such as sodium uranyl acetate (used in analytical chemistry), are known. In basic solution, complex uranates containing up to seven uranium atoms per molecule are formed. In the absence of oxygen or water the fluoride of uranium(VI), the hexafluoride, is stable and is used in the commercially important gaseous diffusion process for the enrichment of uranium-235 in ordinary uranium. Uranium hexachloride also has been prepared, but the hexabromide and hexaiodide are unknown.

The hydride of uranium is a well-defined compound used as a generator of pure hydrogen; when heated to 432° C, hydrogen is given off at a pressure of one atmosphere. The uranium–oxygen system is complicated; uranium oxides having compositions represented by the formulas UO_2, U_3O_8, and UO_3 are known, in addition to others that have less simple formulas. Oxides of many other metals react with uranium or with its oxides, forming compounds containing oxygen, uranium, and the other metal; for example, calcium oxide, CaO, combines with uranium trioxide, UO_3, to produce calcium uranate(VI), $CaUO_4$. There are three nitrides of uranium, having the formulas UN, U_2N_3, and UN_2; the same atomic ratios are found in compounds of the uranium–sulfur system and also in the uranium–carbon system. Uranium carbides are particularly useful as fuels in nuclear reactors that operate at elevated temperatures. Uranium forms a large number of coloured complexes with organic compounds; certain of these are used for colorimetric analysis of small amounts of uranium. Many alloys of uranium have been prepared and characterized. Numerous organometallic compounds have been reported, such as the cyclopentadienyl derivative. Literally hundreds of uranium compounds have been prepared and studied, chiefly because of the use of the element as a nuclear fuel.

Neptunium. Neptunium, atomic number 93, has the radon electron core plus seven outer electrons. There is some uncertainty about the existence of *5f* electrons in elemental neptunium, but their presence in the various compounds of neptunium(III), neptunium(IV), neptunium(V), and neptunium(VI) has been demonstrated. Neptunium in aqueous solution exists as ions in the +3 and +4 states and as oxygenated ions in the +5 and +6 states. The +3 state is unstable in air, being oxidized to the +4 state. It is, however, much more stable than is the +3 state of uranium. The trend towards greater stability for the lower oxidation states with increasing atomic number is noticeable throughout the actinide series. Although the uranyl(VI) ion is the most stable of the uranium ions, the neptunium(V) ion is the stable ion of neptunium in aqueous solution. In addition to the oxygen-containing neptunyl(V) and neptunyl(VI) ions, NpO_2^+ and NpO_2^{2+}, nonoxygenated neptunium(V) and neptunium(VI) compounds are known in the solid state. In particular, neptunium(V) exists in complex ions having the formulas NpF_6^-, NpF_7^{2-}, and NpF_8^{3-}, and neptunium(VI) is present in the neutral molecules of the compound neptunium hexafluoride, NpF_6. All these chemical species have close analogues in the chemistry of uranium. Whereas uranium(VI) has no electrons left beyond the radon core, neptunium(VI) has one. This electron can be removed in basic solution to form neptunium(VII).

Compounds containing neptunium(III) are made most easily by dry reactions, since in solution it is rapidly air-oxidized to neptunium(IV). Neptunium(III) forms compounds typical of the trivalent state; all of the halides are known, as well as a complex salt of potassium with neptunium(III) fluoride. Neptunium(III) fluoride is insoluble in dilute acid. Neptunium in the +4 state is a typical tetravalent actinide, stable in the absence of oxidizing agents and forming the expected compounds. The tetrafluoride, chloride, and bromide of neptunium(IV) are known, but the decreasing stability of the higher oxidation states with increasing atomic number is shown by the lack of existence of the iodide. Oxygenated ions are known both in solution and in numerous solid compounds. As is true of uranium, nonoxygenated compounds of neptunium(V) and neptunium(VI) have been prepared. Salts of alkali metals and complex anions with fluorine exist. Neptunium(VII) forms a few moderately insoluble compounds in basic solution, but not much information about them is available. Complex oxides containing neptunium(VII) and alkali metal ions have been prepared.

A hydride of neptunium is known. The neptunium–oxygen system is complicated in the same way that the uranium–oxygen system is. Carbides, phosphides, nitrides, and silicides also are known. A number of organoneptunium compounds have been prepared. Alloy systems have been studied, and the existence of a number of intermetallic compounds has been reported.

Plutonium. Plutonium, atomic number 94, has the electron core of radon plus eight outer electrons. The amazing number of complex forms of the metal is thought to be due to a mixing of *5f* and *6d* orbitals, which have nearly the same energy. In its compounds, plutonium has been shown beyond any doubt to contain *5f* electrons. Well-characterized oxidation states of +3, +4, +5, +6, and +7 exist, with numerous compounds of these known. The decreasing stability of actinide elements of higher oxidation states with increasing atomic number is again evident; plutonium(IV) is the stable state in aqueous solution. Like uranium and neptunium, plutonium forms oxygenated plutonyl ions as well as fluoride complex ions. Plutonyl(V) is unstable in solution. Plutonium(VI) fluoride is known; it resembles the fluorides of neptunium(VI) and uranium(VI) in its physical properties, although it is less stable. Plutonium(VII) has been shown to exist in strongly basic solution; it is considerably less stable than is neptunium(VII).

Plutonium(III) forms compounds typical of a +3 actinide. Although it is air-oxidized to plutonium(IV), many mild reducing agents stabilize the +3 state, thereby allowing the preparation of many compounds of this oxidation state. Plutonium(IV) is the stable state in aqueous solution, and its compounds can conveniently be prepared from solution. Plutonium(V) has been prepared as an alkali metal–fluoride complex, although plutonium(V) fluoride itself is unknown. In aqueous solution only relatively mild oxidation is necessary to prepare the plutonyl(VI) ion. Only two compounds of plutonium(VII) have been prepared from solution (with barium and strontium). Complex oxides of the alkali metals and plutonium(VII) also have been prepared.

Two hydrides of plutonium are known. Plutonium(IV) forms the oxide of highest oxidation number, which again reflects the lowered stability of higher oxidation states with increasing atomic number. Other oxides are known, and a number of oxides of plutonium and other metals have been reported. Carbides, silicides, nitrides, phosphides, sulfides, and similar compounds also have been prepared.

A considerable number of organic plutonium compounds are known. There are many intermetallic compounds of plutonium with beryllium, aluminum, manganese, iron, cobalt, nickel, and other metals. The intermetallic compounds formed by plutonium differ from those of the other actinides because the structures of plutonium metal are different.

Americium. Americium, atomic number 95, has the radon electron core plus nine outer electrons, seven of which are in the *5f* shell. The stable valence state in solu-

tion is americium(III), in which there are six *5f* electrons, analogous to the lanthanide element, europium. By analogy with europium, americium might be expected to have a stable +2 oxidation state, and solid compounds with halogens, such as americium(II) iodide, AmI_2, and americium(II) bromide, $AmBr_2$, are known. In addition to solutions of americium(III), solutions and solid compounds of americium(IV), (V), and (VI) are known. For the +5 and +6 states, however, only two ions have been found. There is no evidence for the existence of americium(VII) in basic solutions; it appears to be much less stable than either neptunium(VII) or plutonium(VII), and no solid compounds of americium(VII) have been made.

Americium(III) forms typical +3 actinide compounds; insoluble hydroxides and fluorides can be prepared from its stable solutions. The only known halide of americium(IV) is the fluoride. An americium(IV) hydroxide also has been made. Several oxides have been prepared. A few salts of americium(V), such as the carbonates, fluorides, and oxalates have been made. For the +6 state, oxides and sodium americyl acetate have been reported. No compound of americium(V) or (VI) has been prepared that is not in the form of the oxygenated ion.

Americium forms a hydride. A few other compounds, such as sulfides, phosphides, and carbides, have been reported, but there has been much less study of americium chemistry than that of the preceding actinides, due in part to the rarity of americium.

Curium. Curium, atomic number 96, has the radon electron core plus ten outer electrons, seven of which are *5f* electrons. Thus, the *5f* orbital is half filled in this element, as the *4f* orbital is in gadolinium.

This electronic configuration is especially stable, and only curium(III) might be expected to exist, as does only gadolinium(III). In fact, however, curium(IV) can be prepared. This is considered to be further evidence that *5f* electrons are more loosely held than are *4f* electrons. No oxidation states higher than curium(IV) are known. Curium(IV) is a strong oxidizing agent.

Almost all of the known chemistry is that of a typical +3 actinide. Curium(III) is the stable state in solution. All of the halides have been prepared, as have hydroxides, fluorides, phosphates, and other compounds. The only two binary compounds of curium(IV) are the tetrafluoride and the dioxide.

A lower oxide is known; it melts at about 2,000° C and may be useful as a thermoelectric-power source. A number of double fluoride salts of alkali metals and curium(IV) have been prepared. The study of curium chemistry is complicated by the short half-lives of the available isotopes, 5 months for curium-242 and 17.6 years for curium-244, and by the rarity of the element.

Berkelium. Berkelium, atomic number 97, has the radon electron core plus 11 outer electrons, eight of which are *5f* orbital electrons, one more than that of the half-filled shell of curium. Berkelium might be expected to have a +3 state and, due to the stability of the half-filled shell, might also be expected to lose one more electron to form a +4 state. In fact, both +3 and +4 states are found. The upper state is a quite strong oxidizing agent, but it is sufficiently stable to occur both in solution and in the solid state. All of the berkelium(III) halides have been synthesized, but the only berkelium(IV) halide known is the tetrafluoride. Without doubt, a number of complex fluorides with alkali metals can be prepared. Both the sesquioxide and the dioxide have been identified. The chemical behaviour of berkelium is that of a typical actinide capable of existing in +3 and +4 oxidation states.

Californium. Californium, atomic number 98, has the radon core plus 12 electrons, ten of which are *5f* electrons. In solution, californium(III) is the only oxidation state found. Californium(IV) fluoride, however, has been prepared by the use of fluorine, and oxygen under high pressure has furnished the dioxide; these are the only californium(IV) compounds known. A few californium(III) compounds have been studied, such as the trichloride, the oxychloride, and the trioxide. Apparently californium(III) and californium(IV) behave as typical actinides in the +3 and +4 oxidation states. Compounds containing californium(II) that had been prepared by the early 1970s included the iodide, CfI_2, and the bromide, $CfBr_2$.

Einsteinium. Einsteinium, atomic number 99, has the radon electron core plus 13 electrons; 11 of these are in the *5f* orbital. Very little of its chemistry has been explored, but evidence points to the existence of einsteinium(III) as the stable state, both in solution and in solids. The trichloride and the oxychloride have been reported. The alpha particles emitted by the isotope einsteinium-253 available for these studies tear the crystals apart so that no X-ray diffraction lines are obtained at room temperature. It was found necessary to heat the crystals to about 400° C so that the disruption was healed continuously by crystallization processes before meaningful data could be obtained. The oxide of Es(III), Es_2O_3, has been prepared and identified by electron diffraction. Einsteinium(II) was shown by spectroscopic means to be present within a crystal of calcium fluoride. There is some evidence that einsteinium(II) can be prepared from aqueous solution as amalgams with alkali metals. There is no evidence for einsteinium(IV) or higher states, but, as of the early 1970s, little research in this direction had been carried out.

Fermium. Fermium, atomic number 100, has 14 outer electrons, of which 12 are in the *5f* orbital. The order in which fermium comes off an ion-exchange column is the best proof of the existence of fermium(III). The lanthanide and actinide elements are routinely separated by this procedure, and the order in which they are separated depends in part on their oxidation states. No compounds of the element have been made. There is evidence for the existence of fermium(II) based on alkali metal-amalgam reductions in solution. No other chemistry is known, and, due to the element's rarity, little more is expected to be known soon.

Mendelevium. Mendelevium, atomic number 101, possesses 13 electrons in the *5f* orbital, one less than the completed shell. The chemical knowledge of mendelevium is based almost entirely on its behaviour in ion exchange separations, in which it acts as a typical +3 actinide. Sodium amalgam reduction experiments suggest a quite stable +2 state, which can be prepared rather easily from the +3 state. There are no known compounds of mendelevium, and due to the short half-lives of the isotopes, none can be identified.

Nobelium. Nobelium, atomic number 102, has a completed *5f* orbital containing 14 electrons. Even less of its chemistry is known than that of mendelevium. Two oxidation states, +2 and +3, have been shown to exist. Nobelium(II) is believed to be the stable state in solution—in the absence of any but the strongest oxidizing agents. That this should be the only actinide element with an extremely stable divalent state is quite in line with the trend toward increasing stability of lower valence states with increasing atomic number. The lanthanide analogue of nobelium, ytterbium, also forms a stable +2 state.

Lawrencium. Lawrencium, atomic number 103, is the last of the actinide elements; it has a completed *5f* orbital containing 14 electrons. Its known isotopes are so unstable that the chemical behaviour of lawrencium is largely conjectural; its ion-exchange behaviour suggests that it is a typical +3 actinide. Different valence states may exist in solid compounds, but probably not in solution.

BIBLIOGRAPHY. L.B. ASPREY and R.A. PENNEMAN, "The Chemistry of the Actinides," *Chem. Engng. News*, 45:75–91 (1967), a review article covering all aspects of the actinides; J.J. KATZ and G.T. SEABORG, *The Chemistry of the Actinide Elements* (1957), a thorough but now outdated coverage of actinide chemistry; D. BROWN, *Halides of the Lanthanides and Actinides* (1968), a recent compilation on the numerous halides formed by the actinides; G.T. SEABORG and J.J. KATZ (eds.), *The Actinide Elements*, in the "National Nuclear Energy Series," Manhattan Project Technical Section, Division 4, vol. 14B (1954), individual papers by the original scientists on work done during World War II.

(L.B.As.)

Acton, Lord

An English moralist and historian who, though Catholic by tradition and inclination, was deeply opposed to the doctrine of papal infallibility as formulated by the Vatican Council of 1870, John Emerich Edward Dalberg Acton propounded an austerely liberal philosophy, in which was, nevertheless, present an awareness of the threats to liberty contained in state and society.

By courtesy of the National Portrait Gallery, London



Acton, oil painting by Franz von Lenbach (1836–1904). In the National Portrait Gallery, London.

A stern critic of nationalism, his liberalism was rooted in Christianity. "I fully admit that political Rights proceed directly from religious duties, and hold this to be the true basis of Liberalism." For him, conscience was the fount of freedom, and its claims were superior to those of the state. "The nation is responsible to Heaven for the acts of the State." If democracy could not restrain itself, liberty would be lost. The test of a country's freedom was the amount of security enjoyed by minorities. For Acton, in his judgment of politics as of history, morality was fundamental. He was the great modern philosopher of resistance to the evil state.

Civilized, cosmopolitan, rich, learned, and widely connected, he is remembered as much for his few historical writings as for his prescient concern with the problems of political morality.

Acton was born in Naples on January 10, 1834, the only son of Sir Ferdinand Richard Edward Acton (1801–37) by his marriage to Marie Louise Pelline von Dalberg, heiress to a very respectable German title.

In 1840 his widowed mother married Lord Leveson, the future Lord Granville and Liberal foreign secretary, an alliance that brought Acton early into the intimate circle of the great Whigs. Educated at Oscott College, Warwickshire, he went to Munich to study under the German Catholic Church historian Johann Joseph Ignaz von Döllinger, who grounded him in the new German methods of historical research.

Early training in historical research

Having spent much time in the United States and Europe, he returned to England, settled at the family seat in Aldenham, Shropshire, and was elected to the House of Commons for Carlow, Shropshire, in 1859. In the same year he became editor, following John Henry Newman, of the Roman Catholic monthly the *Rambler*, but this experience was unfortunate. He himself believed that a rigorously scientific history would ultimately strengthen the church and that Christian doctrine would be harmonized with the findings of history. When Pius IX indicated that it was wrong, though not heretical, to reject the opinion of the Roman congregations, Acton did not dispute the Pope's authority but held that it condemned the intellectual independence of his journal (by then called the *Home and Foreign Review*) and so laid down the editorship in 1864. Though he was no upholder of the papacy's temporal power and was ready to denounce its unfortunate effects on the course of history, he was firmly attached to Catholicism: unlike Döllinger, he remained faithful to the Catholic Church after 1870. He regarded "communion with Rome as dearer than life."

His parliamentary career had ended in 1865—he was an almost silent member—but his friendship with William Gladstone, the Liberal leader and prime minister, was of great consequence for the minds of both men. The poet and educationist Matthew Arnold used to say that "Gladstone influences all around him but Acton; it is Acton who influences Gladstone." He claimed, in fact, to be the author of the Home Rule policy for Ireland.

Acton was raised to the peerage on Gladstone's recommendation in 1869, and in 1892 Gladstone repaid his services as adviser by having him made a lord-in-waiting to Queen Victoria.

In 1865 he married Marie von Arco-Valley, daughter of a Bavarian count, by whom he was to have one son and three daughters.

After 1870 Acton wrote little, his only notable publications being a masterly essay in the *Quarterly Review* (January 1878) on "Democracy in Europe"; two lectures delivered at Bridgnorth in 1877 on *The History of Freedom in Antiquity* and *The History of Freedom in Christianity* (both published in 1907)—these last the only tangible portions put together by him of his long-projected "History of Liberty"—and an essay on modern German historians in the first number of the *English Historical Review*, which he helped to found (1886). In 1895, on the death of Sir John Seeley, the prime minister Lord Rosebery had him appointed to the regius professorship of modern history at Cambridge. His inaugural *Lecture on the Study of History* (published in 1895) made a great impression in the university, and his influence on historical study was felt. He delivered two valuable courses of lectures, on the French Revolution and on modern history, but it was in private that the influence of his teaching was most marked.

His influence as a lecturer in history

In 1899 and 1900 he devoted much of his energy to coordinating the project of *The Cambridge Modern History*, a monument of objective, detailed, collaborative scholarship.

His efforts to secure, direct, and coordinate contributors for the project exhausted him. Suffering from the effects of a paralytic stroke that he had suffered in 1901, Acton withdrew to a Bavarian property belonging to his wife's family at Tegernsee. He died there on June 19, 1902.

His life, as stated in his own words, "is the story of a man who started in life believing himself a sincere Catholic and a sincere Liberal; who therefore renounced everything in Catholicism which was not compatible with Liberty and everything in Politics not compatible with Catholicism."

BIBLIOGRAPHY. W.A. SHAW, *A Bibliography of the Historical Works of . . . Lord Acton* (1903); GERTRUDE HIMMELFARB, *Lord Acton: A Study in Conscience and Politics* (1952), an excellent intellectual biography; GEORGE FASNACHT, *Acton's Political Philosophy* (1953), and LIONEL KOCHAN, *Acton on History* (1954), two competent, specific studies; DAVID MATHEW, *Acton: The Formative Years* (1946); HUGH MACDOUGALL, *The Acton-Newman Relations* (1962).

(A.W.J.)

Adam, Robert

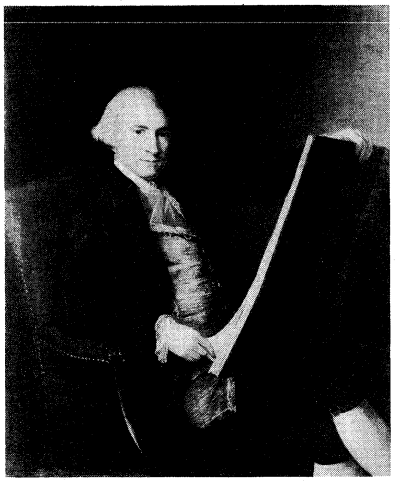
One of the greatest British architects of the 18th century and the originator of the delicate neoclassical Adam style of decoration, Robert Adam was born at Kirkcaldy, Fifeshire, Scotland, on July 3, 1728, the second son of William Adam, the foremost Scottish architect of his time. William, who as Master Mason to the Ordnance in North Britain supervised the design of military buildings, also designed numerous country houses in a conservative Palladian style—the modified classic Roman style that was originally developed by the 16th-century architect Andrea Palladio.

Early life

The Adam children grew up in the cultured atmosphere

of a propertied and well-connected 18th-century family. Shortly after Robert's birth, the family moved to Edinburgh, where, at the age of six, he entered the Edinburgh High School. In 1743 he enrolled at Edinburgh College (now University of Edinburgh), but in 1745 he abandoned his studies and the following year entered his father's offices as an apprentice and assistant.

By courtesy of the National Portrait Gallery, London



Adam, oil painting by an unknown artist. In the National Portrait Gallery, London.

William Adam died in 1748, and his post as Master Mason to the Board of Ordnance passed to his eldest son, John, who took Robert into partnership; and in the succeeding few years both benefitted from the lucrative contracts that resulted from the appointment. Besides building Ft. George in the Moray Firth near Inverness, the Adam brothers were also engaged on another important assignment—the completion of the interior of the Earl of Hopetoun's house. In their interiors, the brothers introduced into Scotland a new, lighter, almost Rococo, style of decoration. The other important private commission of these years was Dumfries House, Ayrshire, for Lord Dumfries.

European influences

In 1754 Robert, who by then considered himself to be worth £5,000, was invited to accompany the Honourable Charles Hope, the Earl's younger brother, to Italy. He thus had the opportunity to realize the dream he had been saving for since his father's death, and, just as important, he had the social advantages of travelling with the brother of an earl. He was as much concerned with meeting young noblemen abroad as with acquiring more architectural knowledge from a study of the monuments of Roman antiquity. The letters he wrote to his family during his years abroad give a picture of Robert as a madly ambitious young man, an arrogant social climber, and yet still a dedicated artist.

He met Charles Hope in Brussels, and together they proceeded to Paris, where Robert fitted himself out in the latest fashions and set out to "lay in a stock of good acquaintance that may be of use to me hereafter." After fewer than three weeks in Paris, they set off for Italy by way of the south of France, visiting en route the Roman sites of Nîmes, Arles, the Pont du Gard, and Montpellier. They reached Genoa early in January 1755 and proceeded to Florence via Leghorn. Arriving at the end of the month, they were immediately caught up in the social whirl for which Robert had hoped.

While he was in Florence Robert made the acquaintance of a man who was to have an important professional influence upon him. This man was the talented young French architect and draftsman Charles-Louis Clérisseau, who agreed to accompany him on his tour as instructor and draftsman. Clérisseau had been a student at the French Academy in Rome but had left in 1754 after a dispute with its director. As a result of his friendship with Clérisseau, Robert came in contact with avant-garde architectural theory in Rome. "I hope," he wrote, "to have my ideas greatly enlarged and my taste formed upon the solid foundation of genuine antiquity." Clérisseau agreed to

> serve [him] as an antiquarian . . . teach [him] perspective and drawing . . . [and] give [him] copies of all [Clérisseau's] studies of the antique, bas-reliefs and other ornaments . . . .

Adam left Florence in February 1755 and travelled to Rome, where he had to decide whether to devote himself to elegant society or to architecture

> If I am known in Rome to be an architect, if I am seen drawing or with a pencil in my hand, I cannot enter into genteel company who will not admit an artist or, if they do admit him, will very probably rub affronts on him in order to prevent his appearing at their card-playing, balls and concerts.

He had to decide: "Shall I lose Hope and my introduction to the great, or shall I lose Clérisseau and my taste for the grand?"

He quarrelled with Hope, and the two separated. Taking rooms for himself and Clérisseau, Robert settled down to serious study, visiting, sketching, and measuring the monuments of antiquity. Among the important figures he met in Rome were the art collector Cardinal Albani and the engraver Giambattista Piranesi, who dedicated an engraving and a book to him.

In May 1757 Adam and Clérisseau left Rome and travelled to Dalmatia via Venice to visit the ruins of Diocletian's palace at Spalatro (Split). Adam felt he

> could not help considering my knowledge of Architecture as imperfect, unless I should be able to add the observation of a private edifice of the Ancients to my study of their public works.

They spent five weeks at Spalatro, preparing the drawings that were eventually to be published in 1764 as the *Ruins of the Palace of the Emperor Diocletian at Spalatro in Dalmatia.*

Having nearly exhausted his money and anxious to return to England, Robert had to forgo the pleasures of further expeditions to Greece and Egypt. He returned to London in January 1758, his head full of details of Roman antiquities. The current Palladian style was losing its appeal, and the public was ready for a new architectural style. Adam lost no time in making his reputation, and by the mid 1760s he had, with the help of his younger brother James, who joined him in London in 1763, created and fully developed the Adam style. That style, which he and James later claimed had "brought about, in this country . . . a kind of revolution in the whole system of this useful and elegant art," was marked by a then new lightness and freedom in the use of the classical elements of architecture. In the Royal Society of Arts building (1772–74), for instance, Adam combined an Ionic pilaster—a rectangular column with scroll-like ornaments projecting from a wall—with a Doric entablature —the upper section of a wall designed in the oldest, simplest Greek style, a liberty a Palladian would never have dared take. This new style consisted basically of a fresh combination of many architectural elements. The various influences included the Palladianism of Lord Burlington and William Kent, both architects; the movement and vigour of the architecture of Sir John Vanbrugh; contemporary French influence, discernible particularly in details, planning, and furniture design; Roman archaeological influences; and the influence of Italian Renaissance decoration, particularly the fanciful ornamentation of the 16th century. Adam's genius was thus a result not of isolated influences but of the synthesis of various lines of development. The Adam style was essentially a decorative style, and it is as a designer of interiors that Adam is chiefly remembered. He gave meticulous attention to every part of each room, from the carpets to the most unobtrusive decoration.

The Adam style

First work in London

Robert's first important work in London was the Admiralty Screen (1758). Through the influence of Lord Bute, a friend of the King, he was appointed Architect of the King's Works in November 1761 along with Sir William Chambers, his principal architectural rival. By the early 1760s he had many domestic commissions; almost without exception these consisted of the completion

or redecoration of earlier houses. It was ironic that despite his fame and ability Adam was rarely called upon to build completely new houses, nor was he to realize his grandiose ideas for public buildings until the very end of his life.

The first Adam interiors at Hatchlands, Surrey (1758–61), and Shardeloes, Buckinghamshire (1759–61), were still near-Palladian in style. But by 1761 his mature style was beginning to develop, and important commissions from this time include Harewood, Yorkshire; Croome Court, Worcestershire; Kedleston Hall, Derbyshire; Bowood, Wiltshire; and Osterley Park, Middlesex. In 1762 he was employed to redesign the interior of Syon House, Middlesex.

Robert produced an important plan that proposed filling an old centre court with a vast domed pantheon-like hall; it was never executed, however. The entrance hall of Syon, based on a basilica—a rectangular building divided into three areas by two rows of columns—with its half-domed ends, is one of the most significant Neoclassical interiors in England.

Other houses from this early phase include the first completely new house, Mersham-le-Hatch, Kent (1762–72); Lansdowne House, Berkeley Square, London (1762–68); Luton Hoo, Bedfordshire (1766–75); Newby Hall, Yorkshire (*c.*1767–85); and Kenwood, London (1767–68).

The south front of Kedleston Hall provides an example of Adam's exterior treatment. His theme of a triumphal arch as the exterior expressions of the great domed interior hall is the first use of this particular Roman form in domestic architecture. The double portico (an open space created by a roof held up by columns) at Osterley Park, derived from the Portico of Octavia, Rome, is a similar Neoclassical motif.

In 1768 Robert and James Adam leased a site on the Thames for a speculative development to be known as the Adelphi (it was almost totally destroyed in 1936). They invested a large sum on embanking the site and building several terraces of houses (1768–72) in which the Adam interior style of slim pilasters supporting a shallow frieze and cornice—the middle and uppermost sections of an entablature—was brought out of doors. It was, however, a financial disaster. In 1773 they again speculated unsuccessfully in a group of stuccoed terraces in Portland Place, London.

The Adams built three major London houses in the early 1770s, which were superb examples of their mature style—No. 20, St. James's Square for Sir Watkin Williams-Wynn (1772–74); No. 20, Portman Square for the Countess of Home (1775–77; now the Courtauld Institute of Art); and Derby House in Grosvenor Square for the Earl of Derby (1773–74; demolished 1862).

In 1773 they published the first volume of *The Works in Architecture of Robert and James Adam.* A second volume followed in 1779, a third was published posthumously in 1822. In the preface to the first volume they explain their idea of "movement," an essential aspect of the Adam style:

> *Movement* is meant to express, the rise and fall, the advance and recess, with other diversity of form, in the different parts of a building, so as to add greatly to the picturesque of the composition.

By 1780 Robert Adam's popularity was beginning to decline, and Horace Walpole, after visiting the architect Henry Holland's new Carlton House, wrote, "How sick one shall be, after this chaste palace, of Mr. Adam's gingerbread and sippets of embroidery."

Neo-Gothic work

Robert Adam designed and built a number of romantic Neo-Gothic castles, mostly dating from the 1780s, in Scotland. The most important of the castles is Culzean, Ayrshire, for the earls of Cassilis (1777–90). Another important work in the Gothic style was the interior at Alnwick Castle, Northumberland (*c.* 1770–80; destroyed in the 19th century).

Toward the end of his life, Robert built the Register House, Edinburgh (1772–92), in which he at last realized the conception of a monumental domed hall within a square, envisaged at Syon some years earlier; and in 1789 he designed the University of Edinburgh, whose entrance front is perhaps his most successful exterior. At Fitzroy Square, London (1790), and Charlotte Square, Edinburgh (1791), he experimented for the last time with the introduction of movement into street architecture.

As a designer of furniture, Adam played a leading role. The furniture style he evolved, popularized by the cabinetmaker George Hepplewhite, was always meant to harmonize with the rest of the interiors. In this field, too, he was prolific, turning his hand to everything from organ cases and sedan chairs to saltcellars and door fittings. It is one of the outstanding features of an Adam interior that everything, even the smallest detail, was part of the unified scheme created by the architect.

Robert Adam died on March 3, 1792, and was buried in Westminster Abbey. The bulk of the nearly 9,000 drawings he left were purchased by the architect Sir John Soane in 1833 and are now in the Soane Museum, London.

MAJOR WORKS

PUBLIC BUILDINGS: Screen wall to the Admiralty, London (1760; mutilated 1827, since restored); Gunton Church, Norfolk (1769); Pulteney Bridge, Bath, Somerset (1769–74); Chandos House, Portland Place, London (1770–71); Royal Society of Arts, London (1772–74); Register House, Edinburgh (1772–92); Mistley Church, Essex (1776; demolished except for twin towers, 1870); Portland Place, London (1776–*c.* 1780; partly rebuilt or demolished).

COUNTRY HOUSES: Shardeloes, Buckinghamshire, portico, interior decorations, and stables (1759–61); Harewood House, Yorkshire, modified John Carr's designs and designed the interiors (1759–71; house much altered, 1843–50); Osterley Park, Middlesex, remodelling of interior (1761–80); Syon House, Middlesex, remodelling of interior (1762–69); Mersham-le-Hatch, Kent (1762–72); Kedleston Hall, Derbyshire (*c.* 1765–70); Luton Hoo, Bedfordshire (1766–75; altered *c.* 1816; reconstructed after a fire in 1843); Nostell Priory, Yorkshire, redecoration (1766–85); Kenwood House, Hampstead, London, remodelling (1767–68); Newby Hall, Yorkshire, added porch and wings and remodelled library (*c.* 1767–85); Saltram House, Devon, decoration (1768 and 1779); Culzean Castle, Ayrshire (1777–90).

TOWN HOUSES: Lansdowne House, Berkeley Square, London (1762–68); Adelphi Buildings, Strand, London (1768–72, demolished); 20, St. James's Square, London (1772–74); Derby House, Grosvenor Square, London (1773–74, demolished); Home House, 20, Portman Square, London (1775–77).

WRITINGS: *Ruins of the Palace of the Emperor Diocletian at Spalatro in Dalmatia* (1764); *The Works in Architecture of Robert and James Adam,* 3 vol. (1773–1822).

BIBLIOGRAPHY. A.T. BOLTON, *The Architecture of Robert and James Adam,* 2 vol. (1922), the standard work on the subject, with comprehensive lists of works and many illustrations; JOHN FLEMING, *Robert Adam and His Circle* (1962), an excellent biographical work based on original letters; EILEEN HARRIS, *The Furniture of Robert Adam* (1963), the best and most complete account of Adam's activity as a furniture designer; DAMIE STILLMAN, *The Decorative Work of Robert Adam* (1966), an account of Adam's work in the field of interior decoration, well-illustrated with original drawings; DOREEN YARWOOD, *Robert Adam* (1970), a popular and often incorrect recent biography concentrating on Adam the man rather than the artist.

(S.Mi.)

Adams, Henry

An American historian, novelist, and man of letters, Henry Adams is the author of one of the outstanding autobiographies in Western literature and a towering figure in 19th-century American intellectual history. Belonging to a patrician family of statesmen and scholars, he felt out of place amid the modern egalitarian and technological forces that were transforming Western society at the end of the 19th century.

Henry Brooks Adams, born in Boston on February 16, 1838, was the product of that city's Brahmin class, a cultured elite that traced its lineage to Puritan New England. He was the great-grandson of John Adams and the grandson of John Quincy Adams. The Adams family tradition of leadership was carried on by his father, Charles Francis Adams (1807–86), a diplomat, historian, and congressman. His younger brother, Brooks (1848–



Henry Adams.
By courtesy of Harvard University Archives

1927), was also a historian; his older brother, Charles Francis, Jr. (1835–1915), was an author and railroad executive. Through his mother, Abigail Brown Brooks, Adams was related to one of the most distinguished and wealthiest families in Boston. Tradition ingrained a deep sense of morality in Adams. He never escaped his heritage and often spoke of himself as a child of the 17th and 18th centuries who was forced to come to terms with the new world of the 20th century.

Adams was graduated from Harvard in 1858 and, in typical patrician fashion, embarked upon a grand tour of Europe in search of amusement and a vocation. Anticipating a career as an attorney, he spent the winter of 1859 attending lectures in civil law at the University of Berlin. With the outbreak of the U.S. Civil War in 1861, Pres. Abraham Lincoln appointed Adams' father minister to England. Henry, age 23, accompanied him to London, acting as his private secretary until 1868.

Career as a correspondent

Returning to the United States, Adams travelled to Washington, D.C., as a newspaper correspondent for *The Nation* and other leading journals. He plunged into the capital's social and political life, anxious to begin the reconstruction of a nation shattered by war. He called for civil service reform and retention of the gold standard. Adams wrote numerous essays exposing political corruption and warning against the growing power of economic monopolies, particularly railroads. These articles were published in *Chapters of Erie and Other Essays* (1871). The mediocrity of the nation's "statesmen" constantly irritated him. Adams liked to repeat Pres. Ulysses S. Grant's remark that Venice would be a fine city if it were drained.

Adams continued his reformist activities as editor of the *North American Review* (1870–76). Moreover, he participated in the Liberal Republican movement. This group of insurgents, repelled by partisanship and the scandals of the Grant administration, bolted the Republican Party in 1872 and nominated the Democrat Horace Greeley for president. Their crusade soon foundered. Adams grew disillusioned with a world he characterized as devoid of principle. He was disgusted with demagogic politicians and a society in which all became "servant[s] of the powerhouse." Americans, he wrote, "had no time for thought; they saw, and could see, nothing beyond their day's work; their attitude to the universe outside them was that of the deep-sea fish." His anonymously published novel *Democracy, an American Novel* (1880) reflected his loss of faith. The heroine, Madeleine Lee, like Adams himself, becomes an intimate of Washington's political circles. As confidante of a Midwestern senator, Madeleine is introduced to the democratic process. She meets the President and other figures who are equally vacuous. After her contact with the power brokers, Madeleine concluded: "Democracy has shaken my nerves to pieces."

In 1870 Charles W. Eliot, president of Harvard College, appointed Adams professor of medieval history. He was the first American to employ the seminar method in teaching history. In 1877 he resigned to edit the papers of Thomas Jefferson's treasury secretary, Albert Gallatin. Pursuing his interest in U.S. history, Adams completed two biographies, *The Life of Albert Gallatin* (1879) and *John Randolph* (1882). He continued to delve into the nation's early national period, hoping to understand the nature of an evolving American democracy. This study culminated in his nine-volume *History of the United States of America* during the administrations of Jefferson and Madison (1889–91), a scholarly work that received immediate acclaim. In this work he explored the dilemma of governing an egalitarian society in a political world in which the predominant tendency was to aggrandize power. In 1884 Adams wrote another novel, *Esther*. Published under a pseudonym, *Esther* dealt with the relationship between religion and modern science, a theme that engaged Adams throughout his life.

His nine-volume history of the United States

Adams was stunned when, in 1885, his wife of 13 years, Marian Hooper, committed suicide. Distraught, he arranged for the sculpture of a mysterious, cloaked woman to be placed upon her grave. The union had produced no children, and Adams never remarried. After his wife's death, Adams began a period of restless wandering. He travelled the globe from the South Sea islands to the Middle East. Gradually the circuit narrowed to winters in Washington and summers in Paris.

Though Adams referred to his existence during this period as that of a "cave-dweller," his life was quite the opposite. From the 1870s until his last years, intellectuals gravitated to his home to discuss art, science, politics, and literature. Among them were the British diplomat Sir Cecil Arthur Spring-Rice, the architect Henry Hobson Richardson, and Sen. Henry Cabot Lodge. His closest friends were the geologist Clarence King and the painter John La Farge. Adams, King, and La Farge were inseparable. Their letters remain a rich source of information on everything from gossip to the most current trends of thought.

While in France, Adams pushed further into the recesses of history in search of "a fixed point . . . from which he might measure motion down to his own time." That point became medieval Christendom in the 13th century. In *Mont-Saint-Michel and Chartres* (printed privately, 1904; published, 1913) he described the medieval world view as reflected in its cathedrals. These buildings, he believed, expressed "an emotion, the deepest man ever felt—the struggle of his own littleness to grasp the infinite." Adams' attraction to the Middle Ages lay in the era's ideological unity; a coherence expressed in Catholicism and symbolized by the Virgin Mary.

The autobiography

The Education of Henry Adams (printed privately, 1907) was a companion volume to *Chartres*. The *Education* remains Adams' best known work and one of the most distinguished of all autobiographies. In contrast to *Chartres*, the *Education* centred upon the 20th-century universe of multiplicity, particularly the exploding world of science and technology. In opposition to the medieval Virgin, Adams saw a new godhead—the dynamo—symbol of modern history's anarchic energies. The *Education* recorded his failure to understand the centrifugal forces of contemporary life. The book traced Adams' confrontations with reality as he moved from the custom-bound world of his birth into the modern, existential universe in which certainties had vanished. The *Education* was published posthumously in 1918.

Neither history nor education provided an answer for Henry Adams. Individuals, he believed, could not face reality; to endure, one adopts illusions. His attempt to draw lines of continuity from the 13th to the 20th centuries ended in futility. "All he could prove," Adams concluded, "was change."

In 1908 Adams edited the letters and diary of his friend John Hay, secretary of state from 1898 to 1905. His last book, *The Life of George Cabot Lodge*, was published in 1911. In two speculative essays, "Rule of Phase Applied to History" (1909) and *Letter to American Teachers*

of History (1910), Adams calculated the demise of the world. Basing his theory on a scientific law, the dissipation of energy, he described civilization as having retrogressed through four stages: the religious, mechanical, electrical, and ethereal. The cataclysm, he prophesied, would occur in 1921. How literally Adams intended his prediction remains a point of dispute.

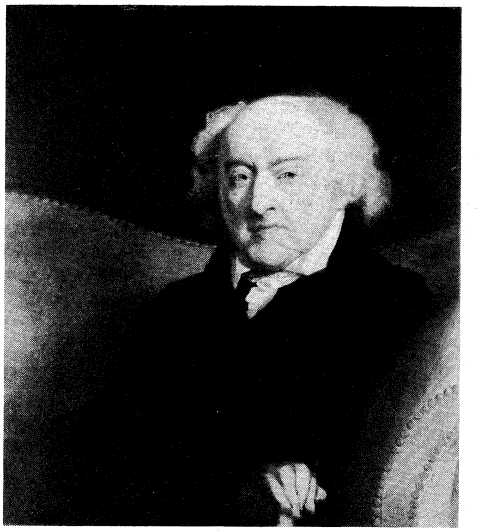
In 1912, at the age of 74, Adams suffered a stroke. His haunting fear of senility became real for a short time. For three months he lay partially paralyzed, his mind hovering between reason and delirium. He recovered sufficiently, however, to travel to Europe once again. On March 27, 1918, Adams died in his sleep in his Washington home and, according to his wish, was buried next to his wife in an unmarked grave. In 1919 he was posthumously awarded a Pulitzer Prize for the *Education*.

Assessment

Adams is noted for an ironic literary style coupled with a detached, often bitter, tone. These characteristics have led some critics to view him as an irascible misfit. They contend that his fascination with the Middle Ages and his continuous emphasis upon failure were masks behind which he hid a misanthropic alienation from the world. More sympathetic commentators see Adams as a romantic figure who sought meaning in the chaos and violence of the 20th century. As Adams described it, he was in pursuit of ". . . a world that sensitive and timid natures could regard without a shudder."

BIBLIOGRAPHY. HENRY ADAMS, *The Degradation of the Democratic Dogma* (1919, reprinted 1969), a collection of Adams' theoretical essays with a lengthy introduction by his brother Brooks; "The Great Secession Winter 1860–1861," *Proceedings*, Massachusetts Historical Society, 43:656–687 (1909–10), an essay, written during the secession crisis of 1860–61 but not published for 50 years, which analyzed the political developments that led to the Civil War; *Essays in Anglo-Saxon Law* (1876), a collection of studies in early British history written by Adams and his seminar students; *Historical Essays* (1891), a reprint of articles Adams had previously published in various journals; and introductions to *Documents Relating to New England Federalism* (1877), which shed light on the politics of the early national period; WORTHINGTON CHAUNCEY FORD (ed.), *A Cycle of Adams Letters, 1861–1865*, 2 vol. (1920, reprinted in 1 vol. 1969), *Letters of Henry Adams, 1858–1891* (1930, reprinted 1969), and *Letters of Henry Adams, 1892–1918* (1938, reprinted 1969), all contain a rich selection of Adams' letters; ERNEST SAMUELS, *The Young Henry Adams* (1948), *Henry Adams: The Middle Years* (1958), and *Henry Adams: The Major Phase* (1964), the most comprehensive and distinguished biography of Adams; J.C. LEVENSON, *The Mind and Art of Henry Adams* (1957), an unexcelled interpretative work centring upon both the man and his writings; WILLIAM H. JORDY, *Henry Adams: Scientific Historian* (1952), focussed upon Adams' scientific approach and his quest for a historical synthesis; MELVIN LYON, *Symbol and Idea in Henry Adams* (1970), an examination of the problem of illusion and reality, which Adams often expressed through symbols, as seen in his six major works; VERN WAGNER, *The Suspension of Henry Adams: A Study of Manner and Matter* (1969), a discussion of Adams' writing style as an example of unique literary artistry; FREDERIC COPLE JAHER, *Doubters and Dissenters: Cataclysmic Thought in America, 1885–1918* (1964), a critical study of Adams, seeing him as a displaced Brahmin afloat in an industrialized world he was helpless to understand.

(C.McH.)



John Adams, oil painting by Gilbert Stuart, 1826. In the National Collection of Fine Arts, Washington, D.C.
By courtesy of the National Collection of Fine Arts, Smithsonian Institution, Washington, D.C.

Adams, John

The first vice president and the second president of the United States, John Adams played a major role in the entire Revolutionary movement—as an aggressive delegate from Massachusetts to the Continental Congress from 1774 to 1778, as a learned advisor and drafter of new state constitutions after 1776, and as a blunt but effective diplomat who helped the 13 American Continental Colonies to secure peace and independence from Britain in 1783. Elevated to high political office in later years, he remained a nonpolitical politician, who remained blind to the relevance of political parties or to any kind of opposition. In his opinion, only the proper organization of government through a system of checks and balances could prevent tyranny, which he identified with the "democracy" of the French Revolution. His belief in his own political rectitude and his love of books, if not his robust love of husbandry, were inherited by his eldest son, John Quincy Adams, who became the sixth president of the U. S.

Youth. John Adams was born on October 30, 1735 (new style), the eldest of three sons of Deacon John Adams, a farmer and selectman, and Susanna Boylston of Braintree, Massachusetts. There he grew up, relishing all the outdoor joys of boyhood, and was educated for college. After graduating in 1755 from Harvard, he studied law in Worcester from 1756 to 1758 in the office of James Putnam, a sophisticated and learned man, in whose household he lived while teaching in the local grammar school. Having returned to Braintree in 1758 and been admitted to the Boston bar, Adams in 1764 married Abigail Smith, a minister's daughter of neighbouring Weymouth who had important family connections that were to benefit his practice.

Intelligent, vivacious, and witty, Abigail composed brilliant letters that together with an uninhibited diary kept by Adams reveal the earthy culture of the New England countryside of the late 18th century and its avidity for manners and learning. In this remarkably self-revealing diary, Adams exhibits a notable talent for the delineation of character in others. A bundle of contradictions himself, he was sometimes outspoken to the point of rudeness, but at other times, scholarly and urbane; irascible, jealous, and suspicious, he was also often playful and tender.

Career as a lawyer in Braintree

As a young lawyer Adams was not only proud of his learning but apparently ambitious of seeing himself in print. In this he succeeded in 1763 by making contributions simultaneously to different newspapers, so that, as has been remarked, "he was carrying on a dialogue with himself in the two leading Boston papers." This was at the time when "sides" were beginning to be formed in Massachusetts politics, the radical wing of colonial opposition to British measures being led by James Otis, Jr., a popular leader at the bar. Adams had long admired Otis, particularly his eloquent defense of the rights of the colonists in the Writs of Assistance case (1761), in which he argued the unconstitutionality of any writ that entitled customhouse officers to search colonists' homes for smuggled goods without specific evidence. Adams remembered this issue in his old age as the one that signalled the opening of the contest with Britain. Doubtless his recollection of the incident was somewhat dramatized, but it inspired him with zeal for the cause of the patriots, and there can be no doubt about his own vigorous and learned participation in the crisis four years later over the Stamp Act, a British revenue law that required publications and documents in the American colonies to bear a stamp. Although he opposed the riots associated with

the act, he regarded it as illegal because the colonists had not consented to it. This argument resulted in his famous articles in the *Boston Gazette*, beginning in August 1765, which were reprinted in London in 1768 as *A Dissertation On the Canon and Feudal Law*. The *Dissertation* argued that the opposition of the New England colonies to the Stamp Act was a part of the protest of New World individualism against Old World collective authority. In 1768 Adams moved his family from Braintree to Boston, where he proudly wore the robes of a barrister recently introduced into the practice of the Superior Court by Chief Justice Thomas Hutchinson (1711–80), later governor of Massachusetts, a man for whom Adams, somewhat paradoxically, bore the utmost dislike. In 1770 Adams courageously participated in the defense of the British soldiers accused of murder in the so-called Boston Massacre. Attempting to maintain order and to enforce the Townshend Acts, which imposed import duties on glass, lead, tea, paper, and paint, the soldiers became involved in a riot that resulted in the deaths of five persons. Public sentiment was strongly against the accused. The trial, however, resulted in the acquittal of the commanding officer and most of the soldiers.

Adams' public career received a temporary setback in 1771. His law practice had recently been enlarged to the point where he had to hire two clerks, but, suffering a brief relapse in health and spirits, he moved back to Braintree and made a long horseback trip "to take the waters" at Stafford Springs. Apparently much improved, he returned to Boston in 1772 to earn a living for his growing family, a girl born in 1765 and his first son, John Quincy, in 1767. His private interests, however, were soon again being sacrificed to public ends, for he was elected to serve in the Massachusetts House of Representatives.

Continental Congress. In August 1774, Adams journeyed to Philadelphia with the Massachusetts delegation to the First Continental Congress, the first federal legislature of the 13 American colonies. Entries in his diary give a rare insight into the composition of that memorable body, which clearly yielded to the radicals in giving first place to grievances rather than to plans for reconciliation with Britain. Here, Adams served as an outspoken but shrewd member for three stormy years. Meantime, in early 1775 he published letters in the *Boston Gazette* under the pseudonym Novanglus, which were rebuttals to letters by the Loyalist writer Daniel Leonard, or "Massachusettensis." Novanglus, not unlike Thomas Jefferson in his "Summary View" of the previous year, emphasized an advanced stage in revolutionary thought: that the colonies having always been outside the realm had never properly been subject to imperial control by Parliament.

Only the personal tie with the sovereign now remained, and this Adams proceeded to sever. In the Second Continental Congress, in June 1775, he nominated George Washington of Virginia to be commander in chief of the almost nonexistent army, a move that he hoped would gain Virginia's support for the Revolutionary policies. A year later, he seconded the resolution of Richard Henry Lee of Virginia "that these United Colonies are, and of a right ought to be, free and independent States" and on June 11, 1776, was appointed to serve on a committee to draft a declaration of independence. Written primarily by Thomas Jefferson, the Declaration was defended on the floor of Congress primarily by Adams. Meanwhile, he had had altercations with some of the more conciliatory members of that body, notably with John Dickinson of Delaware, whom he had described in an imprudent letter, captured and gleefully published by the British, as a "piddling genius." His dislike of hesitant congressmen was accompanied by mixed feelings about a newly arrived journalist, Thomas Paine, the Englishman, whose "Common Sense" nevertheless hastened the cause of independence that Adams strongly favoured in 1776. At a time of general confusion, Adams' *Thoughts On Government* were circulated by several of his colleagues for use throughout the colonies in the making of new constitutions. The *Thoughts* embodied for the first time Adams' conviction that the principal functions of government must be separated in order to prevent tyranny.

Defense of the Declaration of Independence

Revolutionary politics. Between 1776 and 1778, Adams divided his time in promoting the Revolutionary cause in his hometown and state and in serving on numerous congressional committees, including one to create a navy and one to review foreign affairs. The Revolutionaries had begun to hope that they could improve their position by involving France with its old enemy, Britain. Late in 1777, Adams was named by Congress to serve with Benjamin Franklin as a joint commissioner to France, replacing Silas Deane, an original American commissioner whose loyalty to the Revolutionary cause had been called into question. Adams sailed on a continental frigate in February, together with his eldest son, 11-year-old John Quincy Adams. In France, he learned that Franklin had already successfully concluded a French alliance, and later in the year, when the joint American commission in France was dissolved, he returned home, leaving Franklin in sole charge. He arrived in America in the summer of 1779, in time to play a major role in the drafting of the most famous of all Revolutionary state constitutions—that of Massachusetts in 1780—along lines that he considered the proper divisions of functions: "that it may be a government of laws and not of men." He reacquainted himself with congressional politics and again found himself the questionable beneficiary of the old quarrel over Deane. Appointed one of the commissioners to negotiate commercial treaties and to make peace with Great Britain, it was clear that Adams' friends in Congress were noticeably not of the "pro-French" faction but advocates of the blunt, frank school of "shirt-sleeve" diplomacy.

Role in the French alliance

Once again he had to set out on a long and perilous journey overseas, again taking John Quincy with him and a younger son, Charles, and again leaving his wife to run the farm at Braintree. It was a major separation in the lives of the Adamses, not to be healed until six years later when Abigail and her 18-year-old daughter sailed to rejoin the family at Auteuil, near Paris. This was after Adams, together with John Jay, one of the peace commissioners appointed by Congress, and Franklin, had helped bring about a final peace with Britain—on favourable terms and in ways not entirely relished by the French allies—and had negotiated several badly needed congressional loans in Holland. In 1785 he was made the first United States minister to the Court of St. James and moved with his family to England.

Foreign service. In middle age John Adams thus entered into what may be called the second major phase of his life, a career in foreign service in which his son and grandson were also one day to win distinction. The transformation of the simple Braintree lawyer into an envoy capable of dealing with experienced European diplomats was, however, less astonishing than it seems. Although his long years of foreign residence were later to influence American public opinion to the effect that the "Duke of Braintree" had become an aristocrat, such vilification was always basically political. John Adams the envoy was essentially the same blunt New Englander he had always been, and what he principally learned abroad between 1778 and 1788 seems to have been to distrust French politics of the Old Régime (preceding the French Revolution of 1789) and the morals of Europeans, generally; neither idea was wholly new.

While minister to Great Britain, and a year before he returned to America, Adams published his *Defence of the Constitutions of Government of the United States of America* (1787), expressing himself strongly along what some people considered to be "aristocratic" lines—as in the suggestion that "the rich, the well-born and the able" would have to compose the upper chamber of a republic if it were to survive. The work was filled with learned and frequently unacknowledged extracts from an awesome range of writers on government and human nature. Yet despite all its pretentiousness, it merely emphasized ideas held by Adams since 1776 at least—that all branches of government should be separated and that the legislature itself should be two-chambered. It was a schol-

ar's view, based more or less upon what the French political philosopher Montesquieu had once understood to be the make-up of the British government and its system of checks and balances. The political reputation of the *Defence* became an important factor in explaining some of the political difficulties experienced by John Adams in later years. More importantly, the first volume of the *Defence* had arrived in America in time to influence the members of the Philadelphia Convention had they so desired. Offended by the aristocratic tone of the work, many electors, in 1789, failed to cast their votes for Adams. While George Washington received the vote of every elector and was thus elected president, Adams received only 34 out of 69, the second largest number, and was thus declared vice president.

Vice presidency

Vice presidency and presidency. The differences of opinion among the members of the new government over the policies to be pursued resulted in the formation of two well-defined political groups: the Federalists—the conservative faction led by Alexander Hamilton, which advocated a strong, centralized government and favoured industry, landowners, and merchants—and the Democratic-Republicans, led by Thomas Jefferson, who emphasized personal liberty and limited powers for the federal government. Adams had been a supporter of the new Federal plan for a republic from the very first, and although he was less conservative than many others in the party, he became one of the leading Federalists. As vice president, moreover, he was to think of himself, somewhat pompously it is true, as a lightning rod for the whole national experiment. He even fancied himself at one time as a "whipping boy" for some of Hamilton's measures, although his agrarian nature eventually rebelled against the consequence of those policies. The newspaper publication of his *Discourses on Davila* alienated Hamilton, who criticized the work for fear it might bring the federal experiment into disrepute, while Jefferson, on the other hand, thought it verged on promulgating a hereditary monarchy and aristocracy. This alleged predilection for a "mix'd monarchy" later influenced opposition to Adams as a Federalist.

It was no doubt symptomatic of a certain unpopularity of Adams that he could secure only about half of the electoral votes possible in the elections of 1789 and 1792. His eight years under Washington were politically insignificant—as indeed he termed the office of vice president itself—although as presiding officer in the Senate, "His Rotundity" (as his detractors called him) invariably voted in a Federalist direction when breaking ties. Not until 1796, when he was selected by a congressional caucus to succeed Washington, did opposition in his own party become apparent. The fact that he received only a few more electoral votes than Jefferson was due in part to the number of votes cast for another Federalist, Thomas Pinckney of South Carolina, the recently successful envoy to Spain and a favourite with party leaders. Thus, Adams began his term as president with deep suspicions, not only about Jefferson but about Hamilton and his followers. Nevertheless, he kept the cabinet of Washington more or less intact until 1800, despite the fact that several of its members relied upon the advice of Hamilton, Washington's brilliant secretary of the treasury, who was now trying to rehabilitate his fortunes in law practice in New York. The retention of most members of Washington's cabinet was a serious mistake that was partly due to Adams' continuing inability to grasp the significance of "parties." He proudly thought of himself like Washington, as a leader of the whole nation.

Election to the presidency

The Adams administration went from crisis to crisis, for the French Revolutionists were threatening to create a new imperium in Europe, and in America they were stirring up their numerous admirers. Adams had been appalled by the nature of this new revolutionary upheaval from the very beginning, so different, as he saw it, from what had happened in America. Together with his son John Quincy he had taken up the gauntlet seemingly thrown down by Jefferson in 1791, when the latter had endorsed the publication in America of Tom Paine's "Rights of Man," which defended the French Revolution. But if Adams was angered by his Francophile critics, Abigail was outraged, for criticism of her husband was to her little less than treason. Adams did not originate the Alien and Sedition Acts in 1798, which were designed to restrict the activities of friends of Revolutionary France who either hoped to come to America or were already there, but he did sign them into law, and doubtless with Abigail's hearty concurrence. Hating militarism, he decided in 1799 not to go to war with France but to reopen negotiations—after military preparations had been started by Hamilton with the support of the more warlike Federalists. This decision drew the sting from the extremists, many of whom now saw Adams himself as the true enemy. Adams seems to have been more friendly to moderate Federalists than he himself realized. More popular than his own party, he was defeated for re-election in 1800 not only by the strategy of the opposition, the Democratic-Republican Party led by Jefferson and Aaron Burr, but also by the reluctant endorsement of Hamilton who retaliated against Adams' conciliatory policy toward France, by which war was averted. In office to the bitter end, March 3, 1801, he signed in a number of "midnight judges," under an act reforming the national judiciary that he had recommended, having already appointed his secretary of state, John Marshall of Virginia, as chief justice. Thus, Adams attempted to keep Federalists in control of the national judiciary, at least, and he hurried away from the capital without bothering to greet Thomas Jefferson, the new president. He was to remember his critics with great bitterness.

Signing of the Alien and Sedition Acts

Retirement at Quincy. Adams retired to Quincy and lived on into a vigorous and combative old age, composing parts of an "Autobiography" on three different occasions and in varying moods; writing innumerable letters to set people straight about the American Revolution; heaping coals of fire on the head of his neighbour and former friend, Mercy Otis Warren, for the misinformation allegedly conveyed in her *History* of the American Revolution; rejoicing in the size of his manure pile; enjoying his tankard of hard cider each morning before breakfast; and glorying in the diplomatic success and political eminence of his son, John Quincy Adams. Even after Abigail's death in 1818, he strove to maintain an active life, taking long walks in the countryside, dictating to dutiful nieces, and enjoying the role of a senior citizen become a sage.

Gilbert Stuart's portrait of Adams in his 89th year shows him in characteristically bold mood but shrunken frame, obviously in need of teeth, his palsied fingers transparently thin. His last great joy was the elevation of his son to the presidency in 1825. He died on July 4, 1826, just 50 years after the formal Declaration of Independence, on the same day and only a few hours after the death of Thomas Jefferson, the one-time friend and collaborator who had become his political enemy only to become reconciled and friendly again late in life. He lies buried, next to Abigail, in the Stone Temple in downtown Quincy, about halfway between his humble place of birth and the Adams Mansion.

BIBLIOGRAPHY. PAGE SMITH, *John Adams*, 2 vol. (1962), is an exhaustive modern account, enriched with details from the microfilm of the recently opened Adams Family Papers. For broad interpretation, an older biography, GILBERT CHINARD, *Honest John Adams* (1933), is charming and still useful. C.D. BOWEN, *John Adams and the American Revolution* (1950), is a novelized and highly readable story of the early years. The Adams Family Papers have provided the raw material for the modern edition of the *Diary and Autobiography of John Adams*, ed. by L.H. BUTTERFIELD *et al.*, 4 vol. (1961), supplemented by *The Earliest Diary of John Adams* (1966). These bring up-to-date the 19th-century edition by CHARLES FRANCIS ADAMS of his grandfather's major writings, *The Works of John Adams, Second President of the United States: With a Life of the Author*, 10 vol. (1850–56). The opening portions of the *Life* in this edition were written by John Adams' son, John Quincy Adams, the 6th president. It should be noted that both the *Diary* and the *Autobiography* contain virtually nothing on John Adams' later career as vice president and president; his "second life" after retirement to Quincy in 1801 is even more neglected. The original

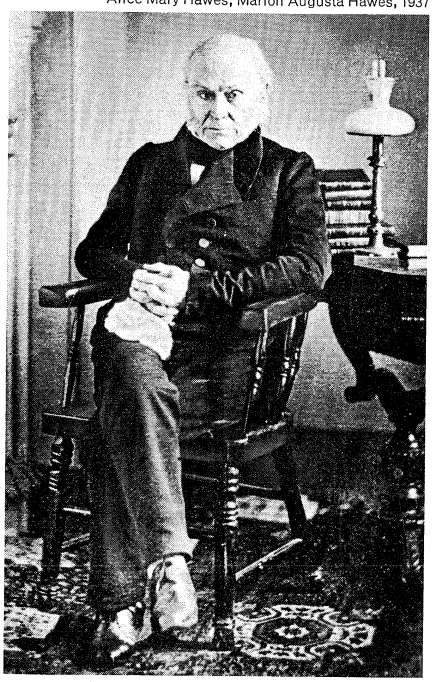
sketch of Adams by EDWARD CHANNING in older editions of the *Encyclopædia Britannica* is still worthwhile, simply because Channing wrote it.

(R.A.E.)

Adams, John Quincy

Adams, John Quincy, eldest son of Pres. John Adams and sixth president of the United States, had two notable careers: first as one of America's greatest diplomats and second as outstanding champion of the antislavery cause in the House of Representatives during the last 16 years of his long life. Between these two careers was the period of his presidency of the United States, from 1825 to 1829.

By courtesy of the Metropolitan Museum of Art, New York, gift of I. N. Phelps Stokes, Edward S. Hawes, Alice Mary Hawes, Marion Augusta Hawes, 1937



John Quincy Adams, daguerreotype by A.S. Southworth and J.J. Hawes.

Early life. Adams was born on July 11, 1767, in that part of Braintree that is now Quincy, Massachusetts. He entered the world at the same time his maternal great-grandfather, John Quincy, for many years a prominent member of the Massachusetts legislature, was leaving it; hence his name. He grew up as a child of the American Revolution. He watched the Battle of Bunker Hill from Penn's Hill and heard the cannons roar across the Back Bay. His patriot father, at that time a delegate to the Continental Congress, and his patriot mother, one of the intellectual women of those times, had a strong molding influence on his education after the war had deprived Braintree of its only schoolmaster. In 1778 and again in 1780 the boy accompanied his father to Europe. He studied at a private school in Paris in 1778–79 and at the University of Leiden in 1780. Thus, at an early age he acquired an excellent knowledge of the French language and a smattering of Dutch. In 1780, also, he began to keep regularly the diary that forms so conspicuous a record of the doings of himself and his contemporaries through the next 60 years of American history.

Early influences

Early diplomatic and political career. In 1781, at the age of 14, he accompanied Francis Dana, United States envoy to Russia, as his private secretary and interpreter of French. Dana, after lingering for more than a year in St. Petersburg, was not received by the Russian government; and in 1782 Adams, returning by way of Scandinavia, Hanover, and the Netherlands, joined his father in Paris. There he acted, in an informal way, as an additional secretary to the American commissioners in the negotiation of the treaty of peace that concluded the American Revolution. Instead of remaining in London with his father, who had been appointed United States minister to the Court of St. James, he chose to return to Massachusetts, where he graduated from Harvard College in 1787. He then read law at Newburyport under the tutelage of Theophilus Parsons, and in 1790 he was admitted to the bar in Boston. While struggling for a practice, he wrote a series of articles for the newspapers in which he controverted some of the doctrines in Thomas Paine's "Rights of Man." In another later series he ably supported the neutrality policy of George Washington's administration as it faced the war that broke out between France and England in 1793. These articles were brought to President Washington's attention and resulted in Adams' appointment as U.S. minister to the Netherlands in May 1794.

Diplomatic posts

The Hague was then the best diplomatic listening post in Europe for the war of the first coalition against Revolutionary France. Young Adams' official dispatches to the Secretary of State and his informal letters to his father, who was now the vice president, kept the government well informed of the diplomatic activities and wars of the distressed Continent and the danger of becoming involved in the European vortex. These letters were also read by President Washington: some of Adams' phrases, in fact, appeared in Washington's "Farewell Address" of 1796. During the absence of the regular minister at London, Thomas Pinckney, Adams transacted public business with the British Foreign Office relating to exchange of ratifications of the Jay Treaty of 1794 between the United States and Great Britain. In 1796 Washington, who came to regard young Adams as the ablest officer in the foreign service, appointed him minister to Portugal, but before his departure his father John Adams became president and changed the young diplomat's destination to Prussia.

John Quincy Adams was married in London in 1797, on the eve of his departure for Berlin, to Louisa Catherine Johnson (1775–1852), daughter of the United States consul Joshua Johnson, a Marylander by birth, and his wife, Katherine Nuth, an Englishwoman.

While in Berlin, Adams negotiated (1799) a treaty of amity and commerce with Prussia. Recalled from Berlin by President Adams after the election of Thomas Jefferson to the presidency in 1800, the younger Adams reached Boston in 1801 and the next year was elected to the Massachusetts Senate. In 1803 the Massachusetts legislature elected him as a member of the Senate of the United States.

Up to this time John Quincy Adams was regarded as belonging to the Federalist Party, but he found its general policy displeasing. He was frowned upon as the son of his father by the followers of Alexander Hamilton and by reactionary groups, and he soon found himself practically powerless as an unpopular member of an unpopular minority. Actually he was not then, and indeed never was, a strict party man; all through his life, ever aspiring to higher public service, he considered himself a "man of my whole country." Adams arrived in Washington too late to vote for ratification of the treaty for the purchase of Louisiana, opposed by the other Federalist senators, but he voted for the appropriations to carry it into effect and announced that he would have voted for the purchase treaty itself. Nevertheless, he joined his Federalist colleagues in voting against a bill to enable the President to place officials of his own appointment in control of the newly acquired territory; such a bill, Adams vainly protested, overstepped the constitutional powers of the presidency, violated the right of self-government, and imposed taxation without representation. In December 1807 he supported President Jefferson's suggestion of an embargo prohibiting all foreign commerce (an attempt to gain British recognition of American rights) and vigorously urged instant action, saying: "The President has recommended the measure on his high responsibility. I would not consider, I would not deliberate; I would act!" Within five hours the Senate had passed the embargo bill and sent it to the House of Representatives. Support of this measure, hated by the

Break with the Federalists

Federalists and unpopular in New England because it stifled the region's economy, cost Adams his seat in the Senate. His successor was chosen on June 3, 1808, several months before the usual time of electing a senator for the next term, and five days later Adams resigned. In the same year he attended the Republican congressional caucus, which nominated James Madison for the presidency, and thus allied himself with that party. From 1806 to 1809 Adams was Boylston professor of rhetoric and oratory at Harvard College.

In 1809 President Madison sent Adams to Russia to represent the United States at the court of the tsars. He arrived at St. Petersburg at the psychological moment when the Tsar had made up his mind to break with Napoleon. Adams therefore met with a favourable reception and a disposition to further the interests of American commerce in every possible way. From this vantage point he watched and reported Napoleon's invasion of Russia and the final disastrous retreat and dissolution of France's *grande armée*. On the outbreak of the war between the United States and England in 1812, he was still at St. Petersburg. That September the Russian government suggested that the Tsar was willing to act as mediator between the two belligerents. Madison precipitately accepted this proposition and sent Albert Gallatin and James Bayard to act as commissioners with Adams, but England would have nothing to do with it. In August 1814, however, these gentlemen, with Henry Clay and Jonathan Russell, began negotiations with English commissioners that resulted in the signing of the Treaty of Ghent on December 24 of that year. Adams then visited Paris, where he witnessed the return of Napoleon from Elba, and next went to London, where, with Clay and Gallatin, he negotiated (1815) a "Convention to Regulate Commerce and Navigation." Soon afterward he became U.S. minister to Great Britain, as his father had been before him, and as his son, Charles Francis Adams, was to be after him. After accomplishing little in London, he returned to the United States in the summer of 1817 to become secretary of state in the Cabinet of Pres. James Monroe. This appointment was primarily because of his diplomatic experience but also because of the President's desire to have a sectionally well-balanced Cabinet in what came to be known as the Era of Good Feeling.

Secretary of state

As secretary of state Adams played the leading part in the acquisition of Florida. Ever since the acquisition of Louisiana, successive administrations had sought to include at least a part of Florida in that purchase. In 1819, after long negotiations, Adams succeeded in getting the Spanish minister to agree to a treaty in which Spain would abandon all claims to territory east of the Mississippi, the United States would relinquish all claim to what is now Texas, and a boundary of the United States would be drawn (for the first time) from the Atlantic to the Pacific Ocean. This Transcontinental Treaty was perhaps the greatest victory ever won by a single man in the diplomatic history of the United States. Adams himself was responsible for the idea of extending the country's northern boundary westward from the Rocky Mountains to the Pacific—considered a stroke of diplomatic genius. To use his own word, it marked a triumphant "epocha" in U.S. continental expansion. Before the Spanish government ratified the Transcontinental Treaty in 1819, however, Mexico (including Texas) had thrown off allegiance to the mother country, and the United States had occupied Florida by force of arms. As secretary of state, Adams was also responsible for conclusion of the treaty of 1818 with Great Britain, laying down the northern boundary of the United States from the Lake of the Woods to the Rocky Mountains along the line of 49° north latitude. Years later, as a member of the House of Representatives he supported 49° north latitude as the boundary of Oregon from the Rocky Mountains to the Pacific Ocean: "I want that country for our Western pioneers." Polk's Oregon treaty of 1846 drew that boundary along the line of 49°. The Monroe Doctrine rightly bears the name of the president who in 1823 assumed the responsibility for its promulgation, but it was the work of John Quincy Adams more than of any other single man.

Presidency and feud with Jackson. As President Monroe's second term drew to a close in 1824, there was a lack of good feeling among his official advisers, three of whom—Secretary of State John Quincy Adams, Secretary of War John C. Calhoun, and Secretary of the Treasury William H. Crawford—aspired to succeed him in his high office. Henry Clay, speaker of the house, and Gen. Andrew Jackson were also candidates. Calhoun was nominated for the vice presidency. Of the other four, Jackson received 99 electoral votes for the presidency, Adams 84, Crawford 41, and Clay 37; because no one had a majority, the decision was made by the House of Representatives, which was confined in its choice to the three candidates who had received the largest number of votes. Clay, who had for years assumed a censorious attitude toward Jackson, cast his influence for Adams, whose election was thereby secured on the first ballot. A few days later Adams offered Clay the office of secretary of state, which was accepted. The charge of "bargain and corruption" followed, and the feud thus created between Adams and Jackson greatly influenced the history of the United States.

The election of 1824

Up to this point Adams' career had been almost uniformly successful, but his presidency (1825–29), during which the country prospered, was in most respects a political failure because of the virulent opposition of the Jacksonians. In 1828 Jackson was elected president over Adams. It was during Jackson's administration that irreconcilable differences developed between the followers of Adams and the followers of Jackson, the former becoming known as the National Republicans, who, with the Anti-Masons, were the precursors of the Whigs. In 1829 Adams retired to private life in the town of Quincy, but only for a brief period; in 1830, supported largely by members of the Anti-Mason movement (a political force formed initially in opposition to Freemasonry), he was elected a member of the national House of Representatives. When it was suggested to him that his acceptance of this position would degrade a former president, Adams replied that no person could be degraded by serving the people as a representative in Congress or, he added, as a selectman of his town. He served in the House of Representatives from 1831 until his death, in 1848. But he had not abandoned his hopes for a re-election to the presidency—whether as nominee of the Anti-Masonic Party (in which he was very active as long as that party had political possibilities) or of the National Republican Party or of a union of both or even of the later Whig Party—always in his own mind as a "man of the whole nation." Gradually, these hopes faded.

Second career in Congress. His long second career in Congress was at least as important as his earlier career as a diplomat. Throughout, he was conspicuous as an opponent of the expansion of slavery and was at heart an Abolitionist, though he never became one in the political sense of the word. In 1839 he presented to the House of Representatives a resolution for a constitutional amendment providing that every child born in the United States after July 4, 1842, should be born free; that, with the exception of Florida, no new state should be admitted into the Union with slavery; and that neither slavery nor the slave trade should exist in the District of Columbia after July 4, 1845. The "gag rules," a resolution passed by Southern members of Congress against all discussion of slavery in the House of Representatives, effectively blocked any discussion of Adams' proposed amendment. His prolonged fight for the repeal of the gag rules and for the right of petition to Congress for the mitigation or abolition of slavery was one of the most dramatic contests in the history of Congress. These petitions, from individuals and groups of individuals from all over the northern states, were increasingly sent to Adams, and he dutifully presented them. Adams contended that the gag rules were a direct violation of the First Amendment to the federal Constitution, and he refused to be silenced on the question, fighting indomitably for repeal, in spite of the bitter denunciation of his opponents. Each year the num-

Opposition to slavery

ber of antislavery petitions received and presented by him grew in great numbers. Perhaps the climax was in 1837 when Adams presented a petition from 22 slaves and, threatened by his opponents with censure, defended himself with remarkable keenness and ability. At each session the majority against him decreased until, in 1844, his motion to repeal the then standing 21st (gag) rule of the House was carried by a vote of 108 to 80, and his long battle was over.

Another spectacular contribution of Adams to the antislavery cause was his championing of the cause of the Africans of the slave ship "Amistad"—slaves who had mutinied and escaped from their Spanish owners off the coast of Cuba and brought the slave ship into United States waters near Long Island. Adams defended these Negroes as freemen before the Supreme Court in 1841 against efforts of the administration of Pres. Martin Van Buren to return them to their masters and to inevitable death. Adams won their freedom.

As a member of Congress—in fact, throughout his life—Adams was a constant supporter of improvement of the arts and sciences and the diffusion of knowledge; and he did much to conserve the bequest of James Smithson (an eccentric Englishman) to the United States and to create and endow the Smithsonian Institution with the money from Smithson's estate.

Adams' death

Perhaps the most dramatic event in Adams' life was its end. On February 21, 1848, in the act of protesting an honorary grant of swords by Congress to the generals who had won what Adams considered a "most unrighteous war" with Mexico, he suffered a cerebral stroke, fell unconscious to the floor of the House, and died two days later in the capitol building. His obsequies in Washington and in his native Massachusetts assumed the character of a countrywide pageant of mourning.

Personality. Few men in American public life have possessed more independence, more public spirit, and more ability than did Adams; but throughout his political career he was handicapped by a certain personal reserve and austerity and coolness of manner, which prevented him from appealing to the imaginations and affections of the people. He had few intimate friends, and not many men in American history have been regarded, during the period of their lifetime, with so much hostility or attacked with so much rancour by their political opponents.

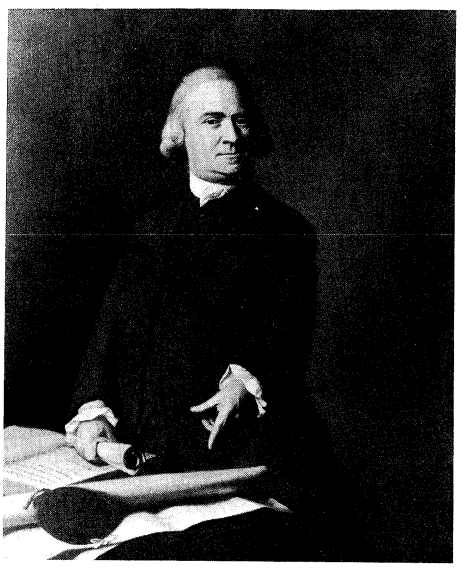
BIBLIOGRAPHY. C.F. ADAMS (ed.), *Memoirs of John Quincy Adams, Comprising Portions of His Diary from 1795 to 1848* (1874–77, reprinted 1969); J.T. MORSE, *John Quincy Adams* (1883; new ed., 1899); DEXTER PERKINS, "John Quincy Adams," in *American Secretaries of State and Their Diplomacy*, vol. 4 (1928, reprinted 1963); BENNETT CHAMP CLARK, *John Quincy Adams, "Old Man Eloquent"* (1932). Two basic volumes by SAMUEL FLAGG BEMIS, *John Quincy Adams and the Foundations of American Foreign Policy* (1949) and *John Quincy Adams and the Union* (1956), utilized the fabulous archives of the Adams family, now in the possession of the Massachusetts Historical Society and in the process of voluminous publication, and a wealth of other official and private collections.

(S.F.Be.)

Adams, Samuel

A radical champion of revolution and a skillful politician and polemicist, Samuel Adams did more than any other American to arouse opposition against English rule in the Colonies. He was born in Boston, Massachusetts, on Sept. 27, 1722, a second cousin of John Adams, second president of the United States. He graduated from Harvard College in 1740, briefly studied law, and failed in several business ventures. As a tax collector in Boston, he neglected to collect the public levies and to keep proper accounts, thus exposing himself to suit.

Although unsuccessful in conducting personal or public business, Adams took an active and influential part in local politics. By the time the English Parliament passed the Sugar Act (1764) taxing molasses for revenue, Adams was a powerful figure in the opposition to British authority in the Colonies. He denounced the act, being one of the first of the colonials to cry out against taxation without representation. He played an important part in instigating the Stamp Act riots in Boston that were directed against the new requirement to pay taxes on all legal and commercial documents, newspapers, and college diplomas.



Samuel Adams, oil portrait by John Singleton Copley, c. 1770-72. In the Museum of Fine Arts, Boston.
By courtesy of the Museum of Fine Arts, Boston, deposited by the City of Boston

His influence was soon second only to James Otis, the lawyer and politician who gained prominence by his resistance to the revenue acts. Elected to the lower house of the Massachusetts general court from Boston, Adams served in that body until 1774, after 1766 as its clerk. In 1769 Adams assumed the leadership of the Massachusetts radicals. There is some reason to believe that he had committed himself to American independence a year earlier. John Adams may have erred in ascribing this extreme stand to his cousin at so early a time, but certainly Samuel Adams was one of the first American leaders to deny Parliament's authority over the Colonies; and he was also one of the first—certainly by 1774—to establish independence as the proper goal.

Commitment to American independence

John Adams described his cousin as a plain, modest, and virtuous man. But in addition, Samuel Adams was a propagandist who was not overscrupulous in his attacks upon British officials and policies, and a passionate politician as well. In innumerable newspaper letters and essays over various signatures, he described British measures and the behaviour of royal governors, judges, and customs men in the darkest colours. He was a master of organization, arranging for the election of men who agreed with him, procuring committees that would act as he wished, and securing the passage of resolutions that he desired.

During the crisis over the Townshend duties (1767–70), the new import taxes on previously duty-free products proposed by cabinet minister Charles Townshend, Adams was unable to persuade the Massachusetts colonists to take extreme steps, partly because of the moderating influence of Otis. The British troops sent to Boston in 1768, however, offered a fine target for this propaganda, and Adams saw to it that they were portrayed in the colonial newspapers as brutal soldiery oppressing inoffensive citizens and assailing their wives and daughters. He was one of the leaders in the town meeting that demanded and secured the removal of the troops from Boston after some British soldiers fired into a mob and killed five Americans. When news came that the Townshend duties, except for that on tea, had been repealed, his following dwindled. Nevertheless, during the years 1770–73, when other colonial leaders were inactive, Adams persistently revived old issues and found new ones; he was responsible for the foundation (1772) of the committee of correspondence of Boston that kept in contact with similar bodies

Foundation of committee of correspondence

in other towns, in whose establishment he also had a hand. These committees later became effective instruments in the fight against the British.

The passage by Parliament of the Tea Act of 1773, which granted the East India Company a monopoly on tea sales in the Colonies, gave Adams ample opportunity to exercise his remarkable talents. Although he did not participate in the Boston Tea Party, he was undoubtedly one of its planners. He was again a leading figure in the opposition of Massachusetts to the execution of the Intolerable (Coercive) Acts passed by the British Parliament in retaliation for the dumping of tea in Boston Harbor; and as a member of the First Continental Congress, which spoke for the 13 Colonies, he insisted that the delegates take a vigorous stand against Britain. A member of the provincial congress of Massachusetts in 1774–75, he participated in making preparations for warfare should Britain resort to arms. When the British troops marched out of Boston to Concord, Adams and the president of the Continental Congress, John Hancock, were staying in a farmhouse near the line of march; and it has been said that the arrest of the two men was one of the purposes of the expedition. But the troops made no effort to find them, and British orders called only for destruction of military supplies gathered at Concord. When Gen. Thomas Gage issued an offer of pardon to the rebels some weeks later, however, he excepted Adams and Hancock.

Member of Continental Congress

As a member of the Continental Congress, in which he served as a delegate until 1781, Adams was less conspicuous than he was in town meetings and the Massachusetts legislature, for the congress contained many men as able as he. He and John Adams were among the first to call for a final separation from Britain, both signed the Declaration of Independence, and both exerted considerable influence in the congress.

Adams was a member of the convention that framed the Massachusetts constitution of 1780 and also sat in the convention of his state that ratified the Federal Constitution. He was at first an anti-Federalist who opposed the ratification of the Constitution for fear that it would vest too much power in the federal government but finally abandoned his opposition when the Federalists promised to support a number of future amendments, including a bill of rights. He was defeated in the first congressional election. Returning to political power as a follower of Hancock, he was lieutenant governor of Massachusetts from 1789 to 1793 and governor from 1794 to 1797. When parties developed on a national scale, he affiliated himself with the Democratic Republicans, the forerunner of the Democratic Party. After being defeated as a presidential elector favouring Thomas Jefferson in the election of 1796, he retired to private life. He died on Oct. 2, 1803.

BIBLIOGRAPHY. The best biography of Samuel Adams is JOHN C. MILLER, *Sam Adams: Pioneer in Propaganda* (1936). There are several others: WILLIAM V. WELLS, *The Life and Public Services of Samuel Adams*, 3 vol. (1865); JAMES K. HOSMER, *Samuel Adams* (1885); RALPH V. HARLOW, *Samuel Adams: Promoter of the American Revolution* (1923); and STEWART BEACH, *Samuel Adams: The Fateful Years, 1764–1776* (1965). Many of Samuel Adams' writings have been lost or destroyed, but a substantial collection of them, edited by HARRY A. CUSHING, *The Writings of Samuel Adams*, 4 vol., was published in 1904–08.

(J.R.Al.)

Addis Ababa

Addis Ababa, the capital city of Ethiopia, is located at the country's geographical centre. Situated at an altitude of 8,000 feet above sea level, it is the highest city in Africa. Its municipal boundaries enclose more than 90 square miles of forested hills and valleys, among which the city has grown in chaotic fashion. The fast-growing population of about 960,000 reflects the city's importance as a symbol of modernity in a largely poor and underdeveloped country.

As well as being the seat of government, Addis Ababa is also the centre of Ethiopia's economic activity. It is the main distribution centre for agricultural and consumer goods and houses much of the country's fledgling industry. Although isolated from the world before World War II, since the 1960s it has become the diplomatic centre for black Africa.

History. Only since the late 19th century has Addis Ababa been the capital of the Ethiopian state. Its immediate predecessor, Entoto, was situated on a high tableland and was found to be unsatisfactory because of extreme cold and an acute shortage of firewood. The empress Taitu, wife of Emperor Menelik II (reigned 1889–1913), persuaded the Emperor to build a house near the hot springs at the foot of the tableland and to grant land in the area to members of the nobility. The city was thus founded in 1887 and was named Addis Ababa ("new flower") by the Empress.

Founding of the city

In its first years the city was more like a military encampment than a town. The central focus was the Emperor's palace, which was surrounded by the dwellings of his troops and of his innumerable retainers. The rest of the town reflected a similar pattern; the residences of the nobility were widely scattered, and each was surrounded by dwellings of troops and retainers.

As the population increased, the capital increasingly experienced a shortage of firewood. In 1905 several varieties of the fast-growing eucalyptus tree were imported from Australia; the trees spread and provided a forest cover for the city. In 1917 the construction of the railway between Addis Ababa and the French Somaliland port of Djibouti connected the landlocked city to the Indian Ocean. At the time of the coming of the Italians in 1935, the population of Addis Ababa was estimated at more than 100,000. It had by that time some of the features of a modern town, including a few hospitals and schools, a power and water supply, a police force, and a rudimentary system of paved roads.

Addis Ababa was the capital of Italian East Africa from 1935 to 1941. The Italians attempted to turn it into a colonial town with strictly segregated residential areas. They introduced the concept of zoning and built a new native market, still known as the Mercato (Italian: "market"). Modern stone houses were built during this period, particularly in the areas of European residence, and many roads were paved. Other innovations included the large water reservoir at Gefarsa to the west and the hydroelectric station at Akaki to the south.

There were only limited changes in Addis Ababa between 1941 and 1960, but development has been impressive since then. Nearly half of the housing units in Addis Ababa in 1970 were built during the 1960s, as were a number of high-rise office and apartment buildings, luxurious residential villas, and low-cost housing projects. The road system and the water and power supply systems were also expanded in order to serve the growing city.

The contemporary city. Addis Ababa is located on the wide Shewan Plateau, which is crossed by several streams, including the Kebena, Kechene, Kurtume, and Akaki rivers. The lower slopes of the Entoto Hills form the northern limit of the city. The rugged site slopes generally southward, and altitudes vary from 8,500 to 7,200 feet within a distance of about six miles.

The only unusual feature in the otherwise ideal climate is the annual average rainfall of 49 inches, which comes in violent thundershowers mostly in April, July, and August. The temperature varies between a minimum monthly average of 59° F (15° C) in December and a maximum average of 66° F (19° C) in May, with a comfortable annual average temperature of 61° F (16° C).

The climate is favourable for a wide range of plant life. There are extensive eucalyptus groves in and around the city, as well as trees natural to the area such as certain species of acacia and fir. Vegetables, cereals, and livestock, raised in the surrounding region, help to feed the growing population. Wildlife consists mainly of the many hyenas and wild dogs that often venture far into the city at night, scavenging for refuse.

The inadequate sewage system presents significant sanitation problems. The city's streams are badly polluted, but serious air pollution is only a potential threat.

The municipal boundaries of Addis Ababa enclose 92 square miles (238 square kilometres), but in the early 1970s only about one-third of this area was built-up. Opportunities for further expansion are ample, except to the north, where the Entoto Hills are an effective barrier.

Principal zones of the city

Because of unplanned growth, zoning is diffuse, but there are general concentrations of function. The governmental and educational section begins at Asfa Wossen Street and ends at Meskel Square, running south through the centre of the city. Along this route are found three campuses of the university, several schools, the National Museum, six governmental ministries, several palaces, the houses of Parliament, and Africa Hall.

The industrial zone in the far south extends chiefly along the road leading to the town of Debre Zeit and reaches beyond the municipal boundaries. There is also a small concentration of factories to the south of Mexico Square in the southwest.

The commercial zone has two focuses—the Mercato, located in the west, and the Piazza (officially, de Gaulle Square), in the centre of town. The Mercato is reputed to be one of the largest indigenous markets in Africa and attracts buyers and sellers from a 50-mile radius on Saturdays. In the Piazza and adjoining areas are the more expensive European-style shopping centres. Smaller markets and shopping areas are scattered everywhere.

The purely residential areas of expensive homes and villas are situated along the road to the new airport in the southeast and to the south and west of the old airport in the southwest.

Unplanned growth has led to the haphazard development of the road network. Only a few roads are large enough to accommodate the growing volume of traffic. The major roads are paved, but these are mostly in the commercial areas. Access by car to many homes—where it exists at all—is down tortuous cobblestone streets or along muddy dirt roads. Only the main arteries have official street names, and there are no numbered addresses because of the chaotic, twisted street pattern. Public transport is provided by a fleet of about 150 buses that run along the main roads.

Addis Ababa is the hub of the country's transportation network. The five road outlets connect the city to provincial capitals; the only railway is that running to Djibouti. There are air connections to about 40 domestic airports, as well as scheduled international flights to Europe, to India, and to other African cities.

With more than 46,000 motor vehicles registered in Addis Ababa in the early 1970s, as well as many vehicles coming from out of town, traffic was congested. The problem was also complicated by heavy pedestrian and domestic-animal traffic in some parts of the city.

The city's population

The total city population in the early 1970s was about 960,000. The most populous administrative districts were Teklehaimanot and Mekakelegna, in the centre of town, and Bole, in the south. The densest population occurred in the Arada, Teklehaimanot, and Mekakelegna districts. The average population density was about 10,000 persons per square mile.

There is a slight excess of males over females in the relatively young population, of which 40 percent is less than 15 years of age. Ethnic composition is highly varied; the most numerous group is the Amharas, who constitute well over half of the population. The Gallas make up about one-fifth of the total inhabitants; the rest consist of Gurages, Hareris, Tigres, Walamos (Wollamos), Somalis, Dorses, and other groups. There is also a significant foreign community, consisting mainly of Arabs, Italians, Greeks, and Armenians. The largest religious group is composed of the adherents of the Ethiopian Orthodox Church. There is a large Muslim community, and there are also adherents of other Christian denominations.

About two-thirds of the population live in housing units that average about three persons to one room. Of these units, about one-fourth are owned by their occupiers, and the remainder are rented; a few are occupied either rent-free or in exchange for services rendered. Only about 7 percent of the housing units have piped water. About half, however, have access to piped water within a distance of 80 feet; others are near a river or well.

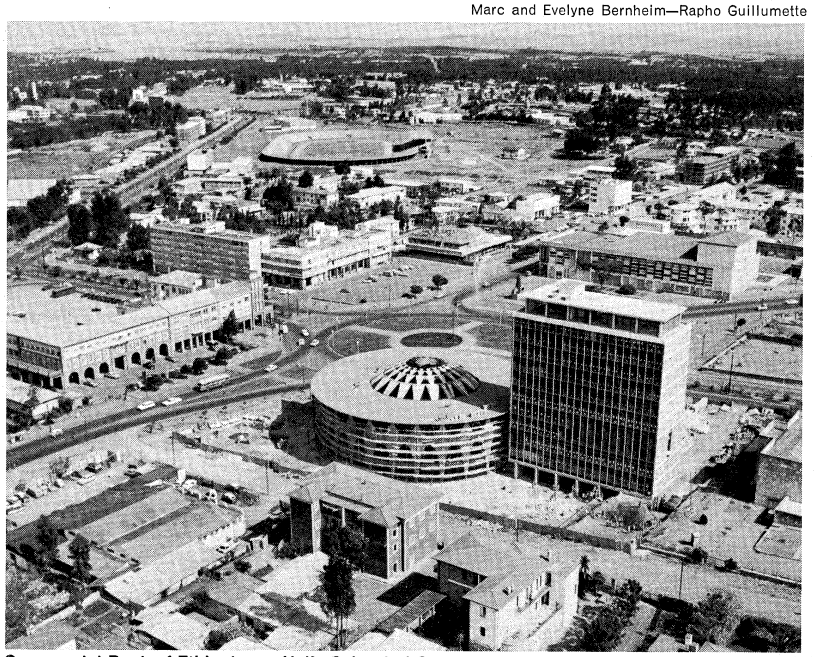
In 1967 about 40 percent of the housing units consisted of permanent structures made of durable materials such as stone, bricks, concrete, tiles, or mud and wood. The remainder were constructed of locally available materials and had roofs of thatch or of corrugated iron.

The economy. Addis Ababa contains about 60 percent of the industries and about 50 percent of the industrial workers of Ethiopia. The variety of consumer products manufactured for the local market includes textiles, shoes, food, beverages, wood products, plastics, and chemical products. Most of the country's service industries are also located in the city. Industry, however, employs no more than 3 percent of the total city population.

Larger industries are mostly government-owned; smaller industries and larger retail outlets are generally owned and operated by aliens. Arabs are gradually losing control of the small retail business to Ethiopians.

The bulk of the export and import trade of Ethiopia is channelled through Addis Ababa on its way to or from

Marc and Evelyne Bernheim—Rapho Guillumette



Commercial Bank of Ethiopia, on Haile Selassie I Square in Addis Ababa.

the ports of Djibouti, on the Gulf of Aden, or Aseb (Assab), on the Red Sea. The city is also the collection and distribution centre for much of the country's internal trade. Banking and insurance services are concentrated in Addis Ababa, and there is a small but growing stock exchange.

As the country's industrial, commercial, and administrative centre, Addis Ababa contributes a major proportion of the country's gross domestic product. Per capita income in the city is several times higher than that in the nation as a whole; it attracts a constant flow of unskilled labour from the countryside, which has resulted in an increasing amount of unemployment in the city.

Administration and social conditions. Addis Ababa is the centre of all governmental activity. The 14 government ministries are located in the city, as are the houses of Parliament, the imperial palaces, and nearly all of the autonomous governmental agencies. Several international organizations have their headquarters in Addis Ababa, the most important of which are the Organization of African Unity and the United Nations Economic Commission for Africa (located in Africa Hall).

City government

According to the city charter of 1954, all legislative power is vested in an elective 20-member municipal council that may enact local legislation, adopt budgets, and determine policies. Two council members are elected from each of the city's 10 districts for two-year terms. Executive power rests in the mayor, or *kantiba*, who is appointed directly by the emperor and is the ex officio president of the municipal council. The Ministry of Interior has general supervisory powers over the mayor and the council.

The city is divided into ten administrative districts, or *weredawoch*. These are Teklehaimanot, Arada, and Mekakelegna districts, in the centre; Gefarsa, Gulele, and Entoto districts, to the north; Lideta and Bole districts, to the south; Yekka district, to the east; and Keranio district, to the west.

The water supply was augmented in 1970 by the construction of a dam over the Legedadi River, about 15 miles east of the city. The Legedadi reservoir, with a capacity of about 1,400,000,000 cubic feet, and the reservoir on the Akaki River, with a capacity of over 200,000,000 cubic feet, are together expected to supply sufficient water to meet current and future demands. Electric power for the city and the industrial ribbon to the south comes chiefly from the two hydroelectric plants on the Awash River. The plants have a combined installed generating capacity of about 75,000 kilowatts.

In 1970 Addis Ababa had more than 50 medical clinics and 13 hospitals, with a total capacity of more than 2,500 beds. Excluding military medical personnel, there were about 170 doctors and about 300 trained nurses. Nearly all the medical services in the city are either supported or heavily subsidized by the state. The city is relatively healthy and is free from malaria. Other infectious and parasitic diseases such as amoebic dysentery and infectious hepatitis, however, have a high occurrence.

There are six police stations and a large police force. The rate of major crimes is relatively low for a city of comparable size, and no organized crime is known to exist. Pocket picking and robbery, however, are prevalent, and most of the larger households employ watchmen. The three fire stations are well equipped, and firemen also provide miscellaneous rescue services.

Educational services

In the early 1970s the government operated more than 30 elementary, 25 junior secondary, and more than 10 senior secondary schools. In addition, there were schools on all three levels operated by private individuals, by missions, or by the foreign communities. In-service vocational training was offered by private and government institutions, and many schools also offered evening classes. The total school population, numbering about 120,000 in 1970, included about 74 percent of the school-age children.

Haile Selassie I University was established in Addis Ababa in 1961. Eight of its ten faculties are located in the city on four campuses with a student body of more than 4,000 and a faculty numbering nearly 600. A junior college for the training of teachers was established in 1969.

Cultural life. The traditional custodians of culture in Ethiopia have been religious institutions, but the promotion of contemporary cultural life is mostly in the hands of secular institutions. The Department of Fine Arts of the Ministry of Education operates the National School of Music and the art school, while the Antiquities Administration runs the National Library and the National Museum. The university has the Museum of the Institute of Ethiopian Studies, which is primarily anthropological in emphasis. The government-supported patriotic association (Hager Fikir Mahber) and the Haile Selassie I Theatre have popular folklore troupes and theatre groups. Art, Amharic literature, and Ethiopian studies are supported by the Haile Selassie I Prize Trust, which awards annual prizes.

The mass media

A government-owned radio station, with medium-wave service for the Addis Ababa region and shortwave service for the rest of the nation, broadcasts news and information, primarily in Amharic and English. Music programs are primarily Ethiopian, but they include a substantial amount of Western popular music. There has been some experimentation with educational programs designed for use in schools. A powerful private radio station, owned by the Lutheran World Federation, also operates from Addis Ababa and beams basically religious broadcasts in several languages to African and Asian countries. There is a government television station catering to the Addis Ababa region as far south as Asella; it broadcasts daily for about four hours in Amharic and English.

The two major daily newspapers in Addis Ababa—the *Addis Zemen* ("New Era") in Amharic and *The Ethiopian Herald* in English—are published by the Ministry of Information. The ministry also publishes two Amharic weeklies—*Ye-zareitu Ityopia* ("Ethiopia of Today") and *Ye-sport fana* ("The Light of Sports")—one Amharic monthly, *Menen*, and one illustrated English quarterly, *The Ethiopian Mirror*. The most popular newspaper is the Amharic fortnightly *Polisenna Ermejjaw* ("Police and Progress"), which is published by the police force and relies heavily on crime reporting.

Formally designated recreational areas are limited, but open spaces suitable for recreational purposes abound within and around the built-up areas. A small zoo of the imperial lions is located in a park near the university. The more affluent members of the population have access to the lake region, which is a short drive to the south; resorts in that area provide facilities for boating, water-skiing, bathing, and bird-watching.

The most popular spectator sport is soccer, which is played at the Haile Selassie I Stadium. Basketball, volleyball, and other sports are also played, chiefly by school teams. Facilities are available at private clubs and hotels for golf, polo, tennis, and swimming.

BIBLIOGRAPHY. The most extensive work on the city for the period up to 1960 is EDOUARD BERLAN, *Addis-Abeba: la plus haute ville d'Afrique* (1963). No comprehensive work exists in English, but a number of useful articles have appeared in the *Ethiopia Observer;* see especially MARTIN JOHNSON, "Addis Ababa from the Air," and RICHARD PANKHURST, "The Foundation and Growth of Addis Ababa to 1935," both in vol. 6, no. 1 (1962). An account of the indigenous market place is given in MAURICE DE YOUNG, "An African Emporium: The Addis Ababa Markato," *Journal of Ethiopian Studies*, vol. 5, no. 2 (1967). Much practical information on the city is given in the municipality's *Pocket Guide of Addis Ababa* (1964); and a good map of the city is edition 3 of the Mapping and Geography Institute (MGI) of Ethiopia (1964).

(A.Dem.)

Addison, Joseph

Essayist, poet, dramatist, and statesman, Joseph Addison successfully combined two careers in his short life. As a writer he produced one of the great tragedies of the 18th century in *Cato* and brought to perfection the art of the periodical essay in his journal, *The Spectator*. As a civil servant he became an influential supporter of the Whigs (who sought to further the constitutional prin-

ciples established by the Revolution of 1688) in a number of government posts, finally becoming secretary of state under George I. He achieved early fame as a writer of Latin and English verse, but it was his poem on the battle of Blenheim that brought him to the attention of the Whig leaders and paved the way to government employment and literary fame. Dr. Johnson's praise of *The Spectator* as a model of prose style established Addison as one of the most universally admired and influential masters of prose in the language.

By courtesy of the National Portrait Gallery, London



Addison, oil painting by M. Dahl, 1719.
In the National Portrait Gallery, London.

Youth and early career. Addison was born in Milston, Wiltshire, on May 1, 1672, the eldest son of the Rev. Lancelot Addison, later archdeacon of Coventry and dean of Lichfield. After schooling in Amesbury and Salisbury and at Lichfield Grammar School, he was enrolled at the age of 14 in the Charterhouse in London. Here began his lifelong friendship with Richard Steele, who later became his literary collaborator. Both went on to Oxford, where Addison matriculated at Queen's College in May 1687. Through distinction in Latin verse he won election as Demy to Magdalen College on July 30, 1689, and took the degree of M.A. on February 14, 1693. At Magdalen he spent ten years as tutor in preparation for a career as a scholar and man of letters. In 1695 *A Poem to his Majesty* (William III), with a dedication to Lord Keeper Somers, the influential Whig statesman, brought favourable notice not only from Somers but also Charles Montague (later earl of Halifax), who saw in Addison a writer whose services were of potential use to the crown. A treasury grant offered opportunity for travel and preparation for government service. Much of his Latin verse had been published in the second volume of *Musarum Anglicanarum Analecta* (1699), which Addison himself edited. He had also attained distinction by contributing the preface to Virgil's *Georgics*, in John Dryden's great translation of 1697.

European influences on Addison

The European tour (1699–1704) enabled Addison not only to become acquainted with English diplomats abroad but also to meet contemporary European men of letters and to observe at first hand the background of the Latin poets. His letters from Paris record visits to such celebrities as the literary critic Boileau, who was the most famous poet of his day, and to Nicolas Malebranche, the French philosopher, as well as his impressions of French drama and opera. He spent the year 1701 in leisurely travel in Italy, viewing ancient ruins, mountains, and rivers, as well as statues, paintings, and medals, against the background of Horace and Virgil, Claudian and Silius Italicus, who were to figure in his *Remarks on several Parts of Italy* (1705). "A Letter from Italy" (1704), a poetic epistle to Lord Halifax, records his delight in surveying these "poetic fields." The prose *Remarks*, addressed to Lord Somers, reveals not only his interest in the arts but also a true Whig's hatred of despotic power and the concentration of wealth in the papal treasury, contrasted with the widespread poverty of the people. From Italy Addison crossed the Alps at Mont Cenis into Switzerland, where, in Geneva, he learned in March 1702 of the death of William III and the consequent loss of power of his two chief patrons, Somers and Halifax. Instead of returning to England, therefore, he remained in Switzerland through the summer and then made an extended tour to Vienna, Dresden, Hanover (for a visit to the electress Sophia, the mother of the future King George I of England), and Hamburg, finally reaching Holland. There he lingered nearly a year, with visits to such men of letters as Pierre Bayle and contacts with Englishmen in Amsterdam, including the publisher Jacob Tonson (who was then in Holland arranging for the first English publication of Bayle's great *Dictionary*), before returning to England in 1704.

Government service. In London Addison renewed his friendship with Somers and Halifax and other members of the Kit-Cat Club, founded by Tonson as an association of prominent Whig leaders and literary figures of the day—among them Steele, William Congreve, Sir John Vanbrugh, and Sir Samuel Garth. In August of this year London was electrified by the news of the Duke of Marlborough's sweeping victory over the French at Blenheim, and Addison was approached by government leaders to write a poem worthy of the great occasion. Even before the final draft was completed, Addison was appointed commissioner of appeals in excise, a sinecure left vacant by the death of John Locke. *The Campaign*, addressed to Marlborough, was published on December 14 (though dated 1705), the very day the Duke returned in triumph to London. By its rejection of conventional classical imagery (what Dr. Johnson called its "rational and manly contempt of fiction") and its effective portrayal of Marlborough's military genius, it was an immediate success, and three editions were sold before the year's end. It perfectly expressed the nation's great hour of victory.

Years as undersecretary. The Whig success in the election of May 1705, which saw the return of Somers and Halifax to the Privy Council, brought Addison increased financial security in an appointment as undersecretary to the secretary of state for southern affairs, a position of considerable administrative importance, with an income of over £500 a year. With an office at the Cockpit (the government offices in Whitehall), Addison found himself in a busy and lucrative post, involving correspondence with diplomats abroad and a variety of business at home. Early in 1706 he accompanied Halifax to the court of Hanover to discuss details of the Act of Settlement, providing for the Hanoverian succession upon the death of Queen Anne; here and in Holland contacts were renewed with Bayle, the German philosopher Leibniz, and other men of letters. Addison's retention in a new, more powerful Whig administration in the autumn of 1706 reflected his further rise in government service.

Collaboration with Steele

At this time he began to see much of Steele, helping him write *The Tender Husband* (1705), for which he also provided a prologue; when the play was published it carried a dedication to Addison in honour of "an inviolable friendship." In practical ways Addison also assisted Steele with substantial loans and the appointment as editor of the official London *Gazette*. Moreover, as undersecretary, Addison frequently supplied him with material for the paper. He lodged with Steele during the summer of 1707, until the latter's marriage in September, when Addison moved to larger quarters.

As a challenge to the popularity of Italian opera, Addison chose to write an opera in English centred on the legend of "fair Rosamond" and set in Woodstock Park, site of Marlborough's Blenheim Palace. For the music he engaged Thomas Clayton, composer of *Arsinoe, Queen of Cyprus*—an unfortunate choice as it turned out; when *Rosamond* was staged at Drury Lane (March 4, 1707), it ran for only three nights. The libretto itself was well constructed and the lyrics light and effective. Later in the century, with new music by Thomas Augustine Arne, it in fact proved very popular.

Duties in Ireland. In 1708 Addison was elected to Parliament for Lostwithiel in Cornwall and later in the same year was made secretary to the Earl of Wharton, the new lord lieutenant of Ireland. Addison's post was in effect that of secretary of state for Irish affairs, with a revenue of some £2,000 a year. In addition to routine duties as Irish secretary, he was responsible to the crown for much official policy in Ireland, where such delicate matters as the Test Act, the laws against Roman Catholics, and the settlement of the Protestant palatines required a tactful and firm hand. He was elected a member of the Irish Parliament taking his seat as representative for Cavan on May 15, 1709. He also interested himself in the preservation and arrangement of the Irish public records, purchasing the office of keeper of the records in Bermingham's Tower, which he developed into a responsible office with an annual salary of £400.

Addison served as Irish secretary for the two years during which Wharton was lord lieutenant, residing in Dublin Castle from April to September 1709 and from May to August 1710, spending the winter months in England. It was during his term as Irish secretary that his friend Steele began publishing *The Tatler*, which appeared three times a week under the pseudonym of Isaac Bickerstaff. The first number appeared on April 12, 1709, while Addison was still in England, but according to his first biographer (Thomas Tickell), it was not until the appearance of No. 6 (on April 23) that Addison recognized an observation on Virgil that he himself had made to Steele and knew him to be "Isaac Bickerstaff." It seems unlikely, however, that Steele should have kept the project a secret from Addison, since the two were close friends and had virtually collaborated on the *Gazette*. At any rate, while still in Ireland, Addison began contributing to the new periodical. Back in London in September 1709, he supplied most of the essays during the winter of 1709–10 before returning to Ireland in May.

Publication of *The Tatler*

Years of Tory dominance. The year 1710 was marked by the overturn of the Whigs from power and a substantial Tory victory at the polls. Although Addison easily retained his seat in the Commons—this time for Malmesbury in Wiltshire—his old and powerful patrons were again out of favour, and for the first time since his appointment as undersecretary in 1705, Addison found himself without employment. He was thus able to devote even more time to literary activity and to cultivation of personal friendships not only with Steele and other Kit-Cats but, for a short period, with Jonathan Swift—until Swift's shift of allegiance to the rising Tory leaders resulted in estrangement. Addison continued contributing to the final numbers of *The Tatler*, which Steele finally brought to a close on January 2, 1711.

Thanks to Addison's help *The Tatler* was an undoubted success, although it had met with criticism because of Steele's attacks on the new Tory government. By the end of 1710 Steele had enough material for a collected edition of *The Tatler;* thereupon, he and Addison decided to make a fresh start with a new periodical. *The Spectator*, appearing six days a week, from March 1, 1711, to December 6, 1712, offered a wide range of material to its readers, from discussion of the latest fashions to serious disquisitions on criticism and morality, including Addison's weekly papers on *Paradise Lost* and the series on the "pleasures of the imagination." The members of the Spectator Club, whose speculations the new paper professed to publish, were (besides Mr. Spectator himself) Sir Roger de Coverley, an eccentric and old-fashioned country squire; Captain Sentry, a retired army officer; Sir Andrew Freeport, a merchant "of great eminence in the City"; a member of the Inner Temple and frequenter of playhouses; Will Honeycomb, an elderly beau who fancied himself knowledgeable in the ways of women; and finally, a grave and learned clergyman. In bringing learning "out of closets and libraries, schools and colleges, to dwell in clubs and assemblies, at tea-tables, and in coffee-houses," *The Spectator* was eminently successful—as shown by the great variety of letters from readers and the steadily increasing number of advertisements printed. Over 3,000 copies were published daily, and the 555 numbers were then collected into seven volumes. Two years later (from June 18 to December 20, 1714) Addison published 80 additional numbers, with the help of two assistants, and these were later reprinted as volume eight.

Subject matter of *The Spectator*

Addison's other notable literary production, while in the political wilderness during the years of Tory dominance, was his tragedy *Cato*. Performed at Drury Lane on April 14, 1713, the play was a resounding success—largely, no doubt, because of the political overtones that both parties read into the play. To the Whigs Cato seemed the resolute defender of liberty against French tyranny; while the Tories were able to interpret the domineering Caesar as a kind of Roman Marlborough whose military victories were a threat to English liberties. The play enjoyed an unusual run of 20 performances in April and May 1713, eight editions were sold within the year, and it continued to be performed, read, and quoted throughout the century.

Success of *Cato*

Later years. With the death of Queen Anne on August 1, 1714, and the accession of George I, Addison's political fortunes rose. He was appointed secretary to the regents (who governed until the arrival of the new monarch from Hanover) and in April 1717 was made secretary of state. Ill health, however, prevented his taking a very active part in government affairs, and he resigned the office the following year. Meanwhile, he had married the dowager Countess of Warwick and spent the remaining years of his life in comparative affluence at Holland House in Kensington. A series of political essays, *The Free-Holder, or Political Essays*, was published from December 23, 1715, to June 29, 1716, and his comedy *The Drummer* was produced at Drury Lane on March 10, 1716. The dispute over Lord Sunderland's bill for restricting the peerage, in which Addison and Steele took opposing sides, unfortunately estranged the two friends during the last year of Addison's life. He died on June 17, 1719, and was buried in Westminster Abbey, near the grave of his old patron and friend Lord Halifax.

MAJOR WORKS

POEMS: "A Letter from Italy" in Dryden's fifth miscellany (1704); *The Campaign* (1705); Latin verse in vol. 2 of *Musarum Anglicanarum Analecta*, ed. by Addison (1699).

DRAMATIC WORKS: *Rosamond: An Opera* (1707); *Cato: A Tragedy* (1713); *The Drummer; or, the Haunted-House, a Comedy* (1716).

PERIODICAL ESSAYS: *The Tatler* (1709–11), about 42 numbers; *The Whig Examiner* (1710), all 5 numbers; *The Spectator* (1711–12), 251 numbers; *The Guardian* (1713), 53 numbers; *The Lover* (1714), 2 numbers; *The Reader* (1714), 2 numbers; *The Spectator*, 2nd series (1714), 25 numbers; *The Free-Holder, or Political Essays* (1715–16), all 55 numbers; *The Old Whig* (1719), both numbers.

OTHER PROSE WORKS: *Remarks on Several Parts of Italy* (1705); *The Present State of the War* (1708); *The Late Tryal and Conviction of Count Tariff* (1713).

BIBLIOGRAPHY. Addison's works were published with a life by THOMAS TICKELL, 4 vol. (1721; reprinted 1730, 1741, 1804); and by JOHN BASKERVILLE (1761); edited with notes by RICHARD HURD, 6 vol. (1811); and in Bohn's British Classics, 6 vol. (1854–56, frequently reprinted); and by G.W. GREENE, 6 vol. (1856). The standard modern edition of the *Miscellaneous Works* (poems, plays, and prose, omitting the periodical essays) is by A.C. GUTHKELCH, 2 vol. (1914). Addison's *Letters* were edited by WALTER GRAHAM (1941). The definitive edition of *The Spectator* is by DONALD F. BOND, 5 vol. (1965), who has also edited *Critical Essays from the Spectator* (1970). The standard biography is PETER SMITHERS, *The Life of Joseph Addison*, 2nd ed. (1968).

(D.F.B.)

Aden, Gulf of

The Gulf of Aden is the deepwater basin that forms a natural sea link between the Red Sea and the Arabian Sea. Named after the seaport of Aden, in Yemen (Aden), at its western end, the gulf is situated between the coasts of Arabia and the "horn" of Africa. To the west, it narrows into the Gulf of Tadjoura; its eastern geographic limits are defined by the meridian of Cape Guardafui (51°16′ E). In oceanographic and geologic terms, however, it extends to the eastern limits of the continental

shelf beyond the Kuria Muria Islands to the north and the island of Socotra to the south, covering an area of some 205,000 square miles (530,000 square kilometres). Its total length, measured from east-northeast to west-southwest, is 920 miles, and its mean width, measured from north-northeast to south-southwest is 300 miles.

The Sheba Ridge

Submarine relief. The dominant relief feature is the Sheba Ridge, an extension of the ridge system of the Indian Ocean, which extends along the middle of the gulf. The rough topography of the ridge includes a well-defined median valley that is continually offset by faults running approximately northeast to southwest. The largest of these faults forms the Alula-Fartak Trench, in which is found the gulf's maximum recorded depth of 17,586 feet (5,360 metres). The Sheba Ridge is flanked on both sides by sediment-filled basins that reach depths of 13,000 feet at the mouth of the gulf. To the west, the ridge gives way to a relatively shallow east–west-trending valley known as the Tadjoura Trench.

Geology. The main factor in the gulf's geologic formation is the spreading of the sea floor away from the Sheba Ridge axis. The African and Arabian continents split initially along the present continental margins either in the late Eocene Epoch (38,000,000 to 54,000,000 years ago) or else in the Oligocene Epoch (26,000,000 to 38,000,000 years ago). They have since drifted apart in a direction parallel to the gulf's faults.

The gulf is underlain by an oceanic crust, and the Sheba Ridge is characterized by shallow earthquake activity, high heat flow, fresh lavas, and a thin or absent sedimentary cover. The evolution of the gulf is also linked with the geological evolution of the East African Rift Valley and of the Red Sea.

Bottom deposits. Sediment thicknesses increase away from the Sheba Ridge toward the continental shelf, especially in the region south of the Wādī Ḥaḍramawt (a seasonal river that drains into the gulf from the Arabian Peninsula), where a thickness of about one mile is found. Brown, green, and gray muds—characteristic of materials originating on land—predominate near the coasts. The basins are filled mainly with material from the coastal margins, eroded by turbid water currents, while sediments formed in deep water far from land predominate on the Sheba Ridge. A slight amount of wind-blown material is also present.

Currents, temperature, and salinity. The intensive exchange of water between the Red Sea, the Gulf of Aden, and the Arabian Sea, as well as strong evaporation and monsoon (rain-bearing) winds that constitute part of the air flow, all assist in the formation of a complex water structure. The surface layer is highly saline, and eddies complicate its flow pattern. During the northeast monsoon from November to March, the surface temperature is fairly uniform—between 77° and 82° F (25° and 28° C). During the stronger southwest monsoon from May to September, however, horizontal temperature gradients develop, with temperatures ranging from 77° to 88° F (25° to 31° C). In the subsurface layer, at depths of from 300 to 2,000 feet, slightly less saline water flows from the Arabian Sea and passes into the Red Sea through the straits of Bab el-Mandeb. This flow is reversed below 2,500 feet in a highly saline layer that originates in the Red Sea. A low temperature and salinity layer occurs at depths of from 3,000 to 6,000 feet down to the bottom in the east but only in the depressions to the west.

Marine life. Marine life is rich in both the quantity and the variety of its species. Seasonally variable upwelling of waters in the coastal zone provides the surface layer with a considerable supply of nutrient elements, which produce an abundant growth of plankton. Sardines and mackerel abound in these areas of upwelling. The main open-sea fish are dolphin, tuna, billfish, and sharks. Whales are frequently sighted. The gulf provides a breeding ground for sea turtles, and rock lobster are abundant.

Species of fish

Fisheries. Despite a lack of large-scale commercial fishing facilities, the coastline supports many isolated fishing towns and villages. Local fishing takes place close to the shore; sardines, tuna, kingfish, and mackerel make up the bulk of the annual catches. Crayfish and sharks are also fished locally, while survey ships have occasionally pulled in exceptional catches of fish.

Prospects for the future. Future development of the gulf is difficult to predict because it has been little studied. Except for the fisheries, little is known of its resources. Full development of the fishing industry is contingent upon a better understanding of the considerable seasonal, annual, and regional variations in the movements of fish populations.

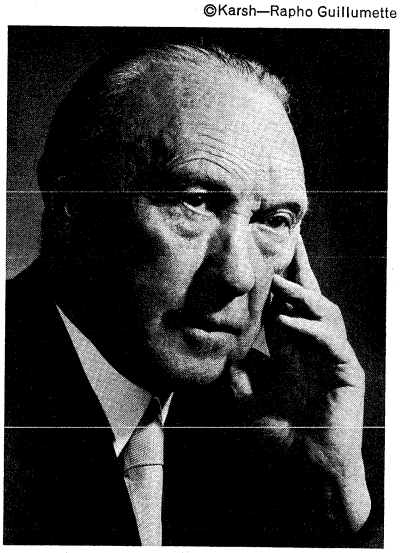
BIBLIOGRAPHY. The morphology, geology, and geophysics of the gulf and its surrounding areas are comprehensively covered in various articles in the *Phil. Trans. R. Soc.*, Series A, vol. 267, no. 1181 (1970), which also contains up-to-date bathymetric and magnetic charts of the gulf itself. The hydrological structure of the water mass is described by V.A. KHIMITSA in *Oceanology*, 8:318–322 (1968); while its physical and chemical properties are covered in the *University of California, Institute of Marine Resources Report, I.M.R. Reference 67–12* (1967). Biology is not covered by any standard text but fishery data is reported in the *Commercial Fisheries Review* (monthly).

(M.T.J.)

Adenauer, Konrad

First chancellor of the Federal Republic of Germany, Konrad Adenauer presided over the reconstruction of the western part of Germany after World War II. During his tenure of office (1949–63), West Germany regained its sovereignty, re-armed, and won a respected place in the affairs of Europe.

©Karsh—Rapho Guillumette



Adenauer.

Konrad Adenauer was born on January 5, 1876, the son of a Cologne civil servant. He grew up in a Roman Catholic family of simple means in which frugality, fulfillment of duty, and religious dedication were stressed. He studied jurisprudence and political science at the Universities of Freiburg, Munich, and Bonn. In 1906 he was elected to the Cologne City Council and, in 1917, during World War I, was chosen *Oberbürgermeister*, or lord mayor, of the city. From 1920 he served as a member of the Staatsrat (the central organ representing the diets of the Prussian provinces) and in 1928 was elected its speaker. Politically, he belonged to the Centre Party, which in Cologne was long the leading political force. The largest part of German Catholicism became politically entrenched in it. The conflict between Bismarck and the Vatican had caused the state and German Catholicism to become critically and widely separated. In the Rhineland at that time, Catholicism was submerged in the reservations of the German states, which had come under Prussian rule only in the 19th century.

Mayor of Cologne

Such influences and moods were shared with Adenauer, but they do not explain why he never seized on the opportunities offered by the Central Party to play a forma-

tive role in German politics in the Weimar Republic between 1919 and 1933. Nevertheless, under the Weimar Republic Adenauer's career remained confined to local politics; his inner aloofness and his alienation from the leading men of his own party might have contributed to this. During the Nazi regime he was twice imprisoned.

At the end of World War II, Adenauer returned to his birthplace. The Allied military authorities called him back to his old office, but, before he could begin his task of rebuilding the beautiful, old, bomb-shattered city, Adenauer was dismissed by Sir John Barraclough, British military governor of the North Rhine area, who found him politically incompetent. This second fall from power did not cause Adenauer to withdraw from public life. On the contrary, it released him for his real mission.

Even before the end of the war, a new political party was being formed—the German Christian Democratic Union (CDU)—in which Catholics and Protestants buried their long-standing differences to present a common front against Nazism and to promote Christian principles in government. This movement sprang from the insights of many that the old parties of the civil camp were obsolete and were in their death throes and that the Catholics and the Protestants must unite. This was something new for Germany, which for centuries following the Reformation had created a climate in which the two communities encountered each other often polemically or at a respectful distance. Adenauer was now able to play an important role in the formation of the new party, and in 1946 he became its chairman. He had always had a sharp eye for the possibilities and necessities of political tactics. The CDU began expanding in the four zones of the Allied occupation. As the Soviet Union began increasingly to obstruct the Allied Control Council, the Western Allies decided to give their three occupation zones a federal–state organization. Adenauer became president of the Parlamentarischer Rat (Parliamentary Council), the task of which was to work out a provisional constitution for the intended Federal Republic. The first Bundestag elections took place in August 1949. The CDU won with a narrow majority over the opposition Social Democratic Party (SPD), and Adenauer was appointed chancellor.

Chancellor

As a result of his upbringing and education, Adenauer was always opposed to Socialist ideas. He rejected the notion of an egalitarian mass society, arguing that it would not allow legitimate leadership to emerge. His leading political theme was individualism under the rule of law. He was imbued with the conviction that the state must guarantee its citizens optimal room for independent intellectual and economic development, as well as absolute protection under the law. In these views Adenauer reflected the Roman Catholic social teachings of Pope Leo XIII.

The political platform of the CDU, however, went beyond Adenauer's ideas; it advocated some programs of a Socialist nature. In the controversies that resulted from his party's social program, Adenauer restrained himself—both as chairman of the party and as a government official—from interference. Such restraint was not entirely in keeping with his nature and position.

But Konrad Adenauer was not a politician with merely private interests; he was quite pragmatic in his approach. He was willing to compromise on domestic programs with which he philosophically disagreed so that he could promote the unity of the country and give West Germany an important place in the European community.

Foreign policy

Indeed, the focus of his interest throughout his career lay in foreign affairs. He viewed the expansion of Communist rule into the heart of Europe as a direct threat to the West and its values. He had no faith in the possibility of peaceful coexistence with the Communist world and felt the need for tough opposition to any aggressive military threats from the Communist bloc. He considered as irreconcilable the principal differences between individualistic rule of law and totalitarian dictatorship and between humanistic–Christian teachings and Communist social regimentation. He therefore became a strong advocate of the politics of containment. He was not, however, an opponent of the relaxation of tensions. Under the circumstances, he felt that a policy of appeasement was utterly illusory, if not traitorous. Except for the period during which Nikita S. Khrushchev was in power in the Soviet Union, Adenauer believed the Communist bloc to be a direct military threat that could only be held in check by superior deterrent forces. As a result, he energetically supported German contributions to the North Atlantic Treaty Organization (NATO) and its nuclear arsenal, and he would have preferred the development of a European defense community. He saw in its creation one of the most promising instruments for imparting to Europe a new order and character. A prerequisite for this defense community was the reconciliation of Germany with its neighbours, especially France, and Adenauer worked for this with all his strength.

During Adenauer's chancellorship his opponents demanded that Germany be neutralized and placed in a position of non-alignment between the Eastern and Western blocs. But Adenauer and his party won all major elections because they declared that the risks to security in such a policy would be intolerable. Behind this position was a decision made by the CDU and its party leaders to guide Germany out of its historically vulnerable position between East and West by making the nation an intrinsic part of a united Europe.

To the end of his life, Adenauer was reproached, unfairly, for not having seriously desired the reunification of Germany, but he believed that it was impossible to come to terms with the Soviet Union. In addition, he had made his decision to bring German policies into close alignment with those of the West.

In 1963 Adenauer turned over the chancellorship to Ludwig Erhard, who claimed much credit for the "miracle" of Germany's economic recovery after the war.

There were no political reasons for his decision to step down from office; rather, it was his belief that at age 88—after 14 years in national government—he had grown too old to lead the nation. Thereafter, he began to write his memoirs. Nevertheless, he continued as chairman of the CDU until March 1966.

Personal characteristics

That Adenauer was able to hold his demanding office into the ninth decade of his life bespeaks his excellent physical, mental, and spiritual constitution, his first class staff, and loyal party. Equally important—and in many instances even more important—were the congenial relations he enjoyed with the American and European statesmen, particularly U.S. secretary of state John Foster Dulles and French president Charles de Gaulle. Adenauer's open and undisguised warmth toward these statesmen belies the charge that he was a cold and suspicious individual. He was merely a sober man of great sensibility who despised wishful thinking, which he held to be especially dangerous for Germany. His use of language served him in his political goals, for it was sharpened to be intelligible and convincing to the common man; and its simplicity emphasized his authority.

In his personal life, Adenauer was unpretentious and extremely disciplined. His family was gathered around him, and he was its patriarch. He was married twice and was twice widowed. He ate little, detested smoking, and loved to work in his large garden. He cultivated a deepened appreciation for painting, and several works of old masters adorned the walls of his own home as well as his official residence—the Schaumburg Palace. As a rule, his day began early and ended late, for he required little sleep.

Konrad Adenauer died on April 19, 1967. Among the friends and opponents of the past who visited West Germany to pay a last farewell were the leaders of the United States and France, as well as many other heads of state, including David Ben-Gurion of Israel.

BIBLIOGRAPHY. KONRAD ADENAUER, *Erinnerungen*, 4 vol. (1965–68), his autobiography encompassing the years 1945 to 1963; RUDOLF MORSEY and KONRAD REPGEN (eds.), *Adenauer-Studien*, vol. 1 (1971), contributions by four German historians and political scientists on the politics and personality of Adenauer; PAUL WEYMAR, *Konrad Adenauer* (1955), an authorized popular biography that provides the best description of his life.

(Eu.G.)

Adhesives

Adhesives are substances capable of holding materials together by surface attachment. Egyptian carvings dating back 3,300 years depict the gluing of a thin piece of veneer to what appears to be a plank of sycamore. Papyrus, an early, nonwoven fabric, has the fibres of reed-like plants bonded together with flour paste. Bitumen, tree pitches, and beeswax were used as sealants (protective coatings) and adhesives in ancient and medieval times. The gold leaf of illuminated manuscripts was bonded to paper by egg white, and wooden objects were bonded with glues from fish, horn, and cheese.

In the 19th century rubber- and pyroxylin-cements were introduced, but decisive advances in adhesives technology awaited the 20th century, during which natural adhesives have been improved and many synthetics have come from the laboratories. Adhesives are now essential components of plywood, automobiles, shoes, furniture, cartons, nonwoven fabrics, and a host of other products. While the natural products such as hide glue, starch and dextrin, casein, and natural gums are still important from the standpoint of cost, the synthetics have largely taken over the adhesives industry both as modifiers of natural materials and, more importantly, as high-strength, moisture-resistant additives capable of being produced in many readily usable forms.

The rapid growth of the aircraft and aerospace industries has had a profound impact on adhesives technology. The demand for a high degree of structural strength coupled with resistance to fatigue and severe environmental conditions has led to the development of high-performance materials, counterparts of which have found their way into many industrial and domestic applications.

Polymeric materials

Virtually all synthetic adhesives and certain of the natural materials are polymeric. A polymer is a substance composed of giant molecules formed by the union of from two to hundreds of thousands of simple molecules. The simple molecules that will undergo such a change are known as monomers and their union is called polymerization. The monomer molecules may be all alike, or there may be two or more varieties of monomer involved in the formation of a single polymer. Ethylene molecules can be united with themselves to form a monopolymer or homopolymer such as polyethylene. Copolymers involving a wide variety of monomer types may be formed; the synthetic rubber SBR (styrene-butadiene rubber) which is based on styrene and butadiene monomers, is one example of a relatively simple copolymer. Many polymeric materials offer the high degree of strength, and often flexibility, necessary in many adhesives applications.

A few of the more important advantages of adhesive bonding include: (1) rapid, economical joining, (2) a means of providing a new structural or composite unit, (3) more uniform distribution of stress, (4) a means for obtaining surfaces free of holes or protuberances or with unusual contours, (5) a method of forming joints sealed against moisture, and (6) an important means for minimizing potential points of corrosion so often evident in mechanical joining or even in welded metal components.

Some recent developments in adhesives and bonding include delayed-tack adhesives that after hardening can be reactivated for controlled periods of time by heating; conductive adhesives for use in thermal or electrical applications; anaerobic adhesives that set in the absence of air; hot-melt applied adhesives capable of reactivation with water; high-heat-resistant products without embrittlement tendency at very low temperatures; encapsulated adhesives activated by pressure; adhesives capable of electrodeposition on complex metal shapes; improved surgical adhesives; and adhesives for the construction, garment, and plumbing industries.

In the early 1970s adhesives were proliferating in applications and growing steadily in total volume. Synthetic resin products were increasing. Worldwide adhesive production reached 7,500,000,000 pounds (3,400,000,000 kilograms) in 1970, and promised to reach 14,000,000,000 pounds (6,300,000,000 kilograms) by 1980.

Adhesive types. Adhesives in current use may be classified into a number of groups.

Thermoplastic adhesives. Under normal temperature conditions these materials are capable of supplying the strength needed for a variety of uses and can be softened, or resoftened, by heating. They reharden on cooling without undergoing any chemical change.

Thermosetting adhesives. These polymers are used in higher demand structural applications and have the ability to form a permanently solid, heat-resistant material by cross linking, that is, the joining of long-chain polymers to each other by lateral linking.

Contact adhesives or cements. These usually are based on solvent solutions of neoprene, a relatively crystalline elastomer, and are in common use for installing hard-surfaced, counter-top laminates. Contact cements are so named because they usually are applied to both surfaces to be bonded. Following evaporation of the solvent component, the two surfaces may be joined to form a strong bond with a high resistance to shearing (sliding) forces.

Hot-melt adhesives. Products of this type contain no solvent or dispersant. Resins on which they are based are fusible as well as adhesive in character. These adhesives are applied in a molten state and, on cooling, form rigid, strong bonds. Their principal use is in packaging and lamination.

Pressure-sensitive adhesives. Materials of this group usually are used on such products as masking or mending tapes or as backings for some strippable wall coverings. Not high-strength materials, these are today generally based on synthetic elastomeric components.

Remoistenable adhesives. This group of adhesives includes the products commonly used for sealing envelopes, applying stamps or labels, or as adhesives for cartons or package-sealing tapes. Products of this type are applied from solution or dispersion and allowed to dry. The adhesive characteristic of the base material is reactivated by the application of water.

Adhesive forms. Adhesives exist in a variety of forms that include: (1) solutions in water or other solvent; (2) dispersions or emulsions, in which minute drops of liquid are dispersed in a solvent; (3) preformed sheets or tapes, sometimes reinforced with a woven or nonwoven carrier for the adhesive; (4) hot melts, where no solvent is required and the material forms a strong bond on cooling; and (5) monomeric, which polymerize rapidly in place to form a strong bond.

Theory of adhesion. During the 1960s studies were undertaken to provide a better understanding of adhesion. This work has been concerned largely with familiar physical-chemical phenomena, but a new broader theory of adhesion called the adsorption theory has emerged and is steadily gaining in favour. This theory holds that adhesives stick primarily because of intimate intermolecular contact; thus, if any two materials are placed close enough together presumably they will adhere. If this theory is true, adhesion requires no chemical bond and no mechanical dovetailing. Proponents point to a universal intermolecular force that holds the solid matter of the world together, and suggest that adhesion is a matter of physical adsorption, an attraction mechanism inherent in any surface.

The adsorption theory of adhesion contradicts the popular notion that adhesives must be sticky or tacky substances. Any material is potentially adhesive, for example, water. Water can make close molecular contact with a material; it fails to hold merely because its shear strength is low, and it evaporates. Copper can be an adhesive, but it can only be used to join materials capable of withstanding its high melting temperature. Even polyethylene, which is slippery and difficult to wet, potentially is a good adhesive if handled properly.

To best achieve the extensive and intimate contact required for adhesion, surfaces must be clean, and one of the materials must be fluid during some phase of the bonding process. For best results, the adhesive must not only be fluid, it must also spread, flowing into surface crevices rather than bridging over them.

Adhesives must either be used in a thin layer or have high cohesive strength, that is, the adhesive material itself must be strong when it sets to prevent failure of the bond during use. Maximum bond strength may be obtained by loading the adhesive joint in a preferential direction; thus, best performance with a large bond area is obtained when loading is "in shear," that is, when the loading pressure pulls the adhesive surface across the bonded surface. If the pull is "in tension," that is, pulling the adhesive directly away from the bonded surface, the stress on the adhesive is much greater.

NATURAL ADHESIVES

Natural adhesives are primarily of animal or vegetable origin. Sodium silicate (water glass) is virtually the only inorganic natural product important as an adhesive. It is also the only nonpolymeric adhesive material used in significant volume. Though the demand for natural products has declined in the second half of the 20th century, certain of these continue to be used in considerable volume, particularly in the wood and paper fields. The more important natural products are described below:

Collagen

Animal glue. The term animal glue usually is confined to glues prepared from mammalian collagen, the principal protein constituent of skin, bone, and sinew. Other types of glue obtained from animal sources are usually referred to according to the material from which they are derived, for example, casein glue. Soybean glue, in spite of its vegetable origin, is usually included in this broad category because of its protein nature.

When treated with acids, alkalies, or hot water, the insoluble protein collagen slowly becomes soluble. If the original protein is pure and the conversion process is mild, the high molecular weight product is called gelatin and may be used for edible or photographic products. The lower molecular weight material produced by more vigorous processing is normally less pure and darker in colour and is called animal glue.

Animal glue traditionally has been used in wood joining, book bindery, sandpaper manufacture, heavy gummed tapes, and similar applications. In spite of its advantage of high initial tack (stickiness), much animal glue has been modified by or entirely replaced with synthetic products.

Casein glue. This product is made by dissolving casein, a protein obtained from milk, in an aqueous alkaline solvent. The degree and type of alkali influences product behaviour. In wood bonding, casein glues generally are superior to true animal glues in moisture resistance and aging characteristics. Casein also is used to improve adhering characteristics of paints and coatings.

Blood albumen glue. Glue of this type can be made from either fresh slaughterhouse blood or dried soluble blood powder to which water has been added. Addition of alkali to albumen-water mixtures improves adhesive properties. A considerable quantity of glue products from blood is used in the plywood industry.

Vegetable adhesives. Vegetable adhesives are those that are soluble or dispersible in water and are produced or extracted from natural sources by relatively simple processes. Other adhesives of vegetable origin, such as rubber cements, nitrocellulose, and ethyl cellulose lacquer cements are extracted or produced from plant sources; but these are not water soluble or water dispersible and are not usually called vegetable adhesives.

Vegetable adhesives are obtained from plant sources throughout the world. Starches (and dextrins), which constitute the principal source of adhesives of this type, may be extracted by commercial means from corn, wheat, potatoes, or rice. Their major uses are in corrugated board, packaging, and as a wallpaper adhesive.

Natural gums

Substances known as natural gums extracted from their natural sources also are used as adhesives. Agar, a marine-plant colloid (suspension of extremely minute particles), is extracted by hot water and subsequently frozen for purification. Algin is obtained by digesting seaweed in alkali and precipitating either the calcium salt or alginic acid. Gum arabic is harvested from acacia trees artificially wounded to cause the gum to exude. Chitin, although produced principally from the skeletal material of invertebrates, is loosely classified as vegetable. Synthetically produced carbohydrate adhesives are made from dextrans, in turn produced by the fermentation of sucrose. Most gums are used chiefly in water-remoistenable products.

SYNTHETIC RESIN ADHESIVES

Adhesives based on synthetic resins excel in versatility and performance. A wide range of base polymers may be synthesized for adhesive and other uses in a constant supply of uniform products. Synthetic resins can be modified in many ways and are often combined to obtain the best characteristics for a particular application. Synthetic adhesives may be divided into two main categories: thermoplastic, and thermosetting. A third category, which sometimes straddles the two, is the class of rubber- or elastomer-based materials (see also PLASTICS AND RESINS).

Thermoplastic resin adhesives. A number of different resins may serve as thermoplastic adhesives and may be used in the form of solutions, dispersions, or solids.

Vinyl resins. The vinyl resins, all of which contain a vinyl ($CH_2 = CH-$) group in their molecules, comprise the largest class of thermoplastic adhesives. Many modifications are available; the adhesive bonds may be formed either by fusing with heat or by evaporation of solvent. Vinyl acetate, obtained by the reaction of acetylene, a gaseous hydrocarbon, and acetic acid, the chief acid of vinegar, with a suitable catalyst, can be polymerized to form the white emulsions or water dispersions often sold for household use and consumed industrially in large quantities for bonding wood and paper products.

Other polyvinyl resins are used in laminating shatterproof glass, in bonding plastic sheet to metals, in various construction applications, and in bonding upholstery materials to furniture frames.

Styrene, synthesized from ethyl benzene, may be polymerized to form polystyrene, which is used either in solution or aqueous dispersion to bond plastic tile to wood, plaster, or concrete materials.

Acrylates, resins derived from acrylic acid, a liquid organic acid that polymerizes readily, are displacing rubber and rosin in pressure-sensitive adhesive tapes.

The cyanoacrylates, monomers that polymerize spontaneously in place, form excellent bonds almost instantly to a variety of surfaces. Materials of this group have been used as tissue adhesives in surgical procedures.

Cellulose derivatives. Certain esters and ethers, compounds derived from cellulose, a complex carbohydrate that comprises the main part of plant cell walls and yields many fibrous products, have been used extensively as adhesives.

Cellulose acetate and cellulose nitrate are dissolved in solvents to form quick-drying, general-purpose, household adhesives; they are excellent for wood and paper and are resistant to moisture, but their adhesion to metal and glass is poor. Nitrocellulose was used to bond some of the early nonstitched footwear, but its use for this purpose is now limited.

Other thermoplastics. Polyamides, chemically similar to nylon, or modifications of them are used as hot-melt adhesives, as are rosin, rosin derivatives, ethylene, and vinyl acetate copolymers. Many other thermoplastic polymers also can be adapted for particular hot-melt applications.

Infusible and insoluble bonds

Thermosetting resin adhesives. Thermosetting adhesives are based on synthetic resins that, on hardening, or curing, undergo an irreversible chemical and physical change to become substantially infusible and insoluble. The type of chemical reaction responsible for the change varies with the resin type. Generally, thermosetting adhesives retain their strength better at elevated temperatures than do the thermoplastic adhesives. The number of resins in this category presently in use or potentially useful is large. Some of the major classes are listed below:

Urea–formaldehyde resins. Urea (a nitrogenous compound, the chief solid constituent of the urine of mam-

mals, synthesized from carbon dioxide and ammonia) and formaldehyde (a colourless, irritating gaseous compound made by oxidation of gaseous hydrocarbons) are made to undergo a condensation reaction, that is, a chemical change in which they combine, eliminating water and forming a polymer. The polymer molecules can be further cross-linked to form bonds that are resistant to moisture and heat. These urea–formaldehyde resins are in common use in the plywood industry.

Melamines. Melamine, an alkaline, white, crystalline, organic compound, reacts with formaldehyde in a way analogous to that of urea. The melamine and formaldehyde form a hard, clear, water-and-heat resistant product that is commonly used in bonding industrial and decorative laminates for door liners and counter tops.

Phenol–formaldehyde adhesives. Phenol, a poisonous crystalline organic acid, on condensation with formaldehyde, forms a phenolic resin that was one of the first thermosetting adhesives produced. It has been used in plywood, in structural and decorative laminates, in conventional glues, in adhesives for high-strength aircraft structural uses, and in other adhesive formulations to obtain additional resistance to water or heat, or to contribute hardness and rigidity.

Resorcinol–formaldehyde adhesives. Resorcinol, a crystalline phenol, reacts with formaldehyde to form a glue capable of setting without heat like the less costly phenolic resin, but unlike the phenolic glue that cures, or sets, under acid conditions, the resorcinol–formaldehyde adhesive cures under neutral or near neutral conditions. Resorcinol–formaldehyde adhesives are valuable in producing water-and-weather resistant wood bonds for structural and marine applications.

Epoxy resin adhesives. Epoxies are usually furnished in two parts, a resin containing an epoxide group (an oxygen atom joined to two other atoms that are already connected with each other), and a catalyst consisting of an amine or other compound that acts as a curing agent. These two parts are mixed together just before use, and hardening is rapid. Epoxies that are designed to harden at room temperature must be used within a short time after mixing, known as the pot life, often a half hour or less. Epoxies that need to be cured at elevated temperatures may be kept for days or even months at room temperature before they harden, but will harden quickly when heated. Epoxies may be formulated to form a hard, or a slightly flexible bond. Extremely strong bonds may be obtained that are resistant to moisture and high temperatures. Both porous and nonporous materials, including metals, may be bonded together with excellent adhesion. Since epoxies harden by chemical action and are made without solvents, they shrink very slightly on hardening; this quality adapts them for use on poorly fitted joints or where a considerable buildup is required. Their excellence as electrical insulators makes them useful in electronics and aircraft industries. Epoxy resins are used as surface coatings, and as laminating adhesives in making plastic tanks and pipes. They are used in formulations along with other adhesive types, for example those based on phenolics. There is an extremely large potential number of resins containing reactive epoxide groups.

Polyimide resins. Polyimide resins are polymers usually consisting of aromatic nuclei coupled by repeating imide linkages formed by condensation of a diacid or a diamine. These recently developed products are being used increasingly as binders for glass- or graphite-fibre-reinforced composite structures. High priced and relatively difficult to handle, these adhesives in the early 1970s were used only in applications where extreme resistance to heat is required along with good mechanical properties and chemical resistance. In these specialized applications, the polyimides tend to replace the more conventional epoxy, phenolic, and polyester adhesives.

Polyester resins. Polyester resins are polymers whose molecules contain many ester groups (–COO–). Esters are organic compounds analogous to salts in inorganic chemistry. These adhesives are used as binders for glass fibres in the formation of structural panels for luggage, boats, automobile bodies, and machine housings. Like other thermosetting resins, the polyesters harden by chemical action rather than by loss of solvent or dispersant.

Rigid or flexible bonds

Polyurethane-based adhesives. Polyurethanes are polymers containing repeating units of urethane structures which are a reaction product of di- or polyisocyanates with compounds containing active hydrogen. Polyurethanes also form many flexible and rigid foams, and coatings of many uses. Soluble polyurethane elastomers may be used as adhesives. Room-temperature curing is possible, but heat cures give superior bond strengths. Products can be formulated from starting components to yield a wide range of flexibility or rigidity. Polyurethanes have been used in bonding other rubberlike materials and also in laminating fabrics and foams in the garment industry.

Elastomeric adhesives. Elastomeric adhesives are based on natural as well as a variety of synthetic rubber materials. These include polymers such as SBR nitriles (made from the monomers butadiene and acrylonitrile), butyl rubber, neoprene (a synthetic rubber polymer), EPDM (made from monomers ethylene, propylene, and a diene), and others.

Various polymeric organic silicon compounds form adhesives that are rubberlike and retain good flexibility at low temperatures, yet are capable of withstanding temperatures as high as 500° F (260° C) for extended periods of time. These silicone elastomers usually lack high strength and do not adhere to all substances, but they can be modified to increase adhesive characteristics.

BIBLIOGRAPHY. R.L. PATRICK (ed.), *Treatise on Adhesion and Adhesives*, vol. 1, *Theory*, vol. 2, *Materials* (1967–69), a projected three-volume series; R. HOUWINK and G. SALOMON (eds.), *Adhesion and Adhesives*, 2nd rev. ed., vol. 1 (1965), a discussion of adhesion phenomena, including information on both natural and synthetic adhesives; L.H. SHARPE, H. SHONHORN, and C.J. LYNCH, "Adhesives," *Int. Sci. Tech.*, 28:26–37 (1964), a popular treatment of adhesion theory; C.W. COOPER and W.R. DUNNAVANT, "Future Trends and Developments in Adhesive Bonding," *Adhes. Age*, 13:39–42 (1970), a discussion of the advantages of adhesive bonding, recent developments in adhesive applications, and some predictions for the future; I. SKEIST (ed.), *Handbook of Adhesives* (1962), a popular handbook containing both theoretical and practical information on virtually all adhesive types; UNITED NATIONS, ECONOMIC COMMISSION FOR EUROPE, *Market Trends and Prospects for Chemical Products*, 3 vol. (1969), an economic report summarizing chemical product information.

Administrative Law

Administrative law derives from the need to create and develop a system of public administration under law, a concept that may be compared with the much older notion of justice under law. Since administration involves the exercise of power by the executive arm of government, administrative law is of constitutional and political importance as well as of juridical significance.

The law regulating public administration

There is no universally accepted definition of administrative law that fits precisely the diverse systems obtaining in the many countries in which it exists in one form or another. A rational method of legal classification, however, defines each branch of the law according to the subject matter that it regulates; and thus administrative law is the law regulating public administration. This includes the organization, powers, duties, and functions of public authorities of all kinds engaged in administration; their relations with each other and with citizens and nongovernmental bodies; legal methods of controlling public administration; and the rights and liabilities of officials. The line between constitutional and administrative law is hard to draw, and over much of the ground these two branches of public law are complementary. The organization of a national legislature, the structure of the courts, the characteristics of a cabinet, and the role of the head of state are generally regarded as matters of constitutional law, whereas the substantive and procedural provisions relating to central and local governments and judicial review of administration are clearly matters of

administrative law. But some matters, such as the responsibility of ministers, cannot be exclusively assigned to either administrative or constitutional law. Some French and American jurists regard administrative law as including parts of constitutional law. This is one way of avoiding the dilemma.

The huge mass of substantive law relating to public health, education, housing, and other public services could logically be regarded as part of the corpus of administrative law; but from a practical point of view its sheer bulk makes the subject too unwieldy to master in a systematic manner. It is better, therefore, to consider these specialized bodies of law as ancillary branches (such as the law of taxation, the law of public health) to be used freely when appropriate in illustrating particular powers, functions, practices, or procedures applicable to administrative law.

GROWTH AND CHARACTERISTICS OF ADMINISTRATIVE LAW

Modern expansion of functions of the state

The growth in the functions of the state has been a worldwide phenomenon during the 20th century. It is to be found in the more developed and in the less developed countries; in both old and new states; in democratic, authoritarian, and totalitarian regimes; in the Communist countries of eastern Europe and in the mixed economies of the West. The movement is far from having reached its zenith. With each addition to the functions of the state, additional powers have been acquired by the administrative organs concerned, which may be central ministries, local, provincial, or regional governments, or special agencies created for a particular purpose.

Distinctions between public administration and private action. Many functions of government have no counterpart in private activity. Traffic control, fire-protection services, the arrest of criminals, smoke abatement, the construction or repair of highways, the provision of currency, town and country planning, customs and excise—these and innumerable other services are peculiar to government. In carrying out such functions, the executive organs are assumed to represent the collective will of the community and to be acting for the common good. For this reason they are given powers not normally conferred on private persons. They are authorized to require citizens to surrender their property rights and to restrict their freedom of action in many different ways, ranging from the quarantining of sick or suspected persons to the payment of taxes. To take another example, the postal laws are much more favourable to the post office and more severe to the customer than the law applying to common carriers. Again, a public authority that is building houses or clearing a slum is in a different—and much stronger—legal position than is a speculative builder or a private developer who is engaged in similar operations.

The result of the distinction between public administration and private action is that administrative law is quite different from private law regulating the actions, interests, and obligations of private persons. Civil servants do not serve under a contract of employment but have a special status. Taxes are not debts, nor are they governed by the law relating to the recovery of debts by private persons. The interrelations between executive organs and their relations with private individuals are usually regulated by compulsory or permissive powers conferred upon the executive organs by the legislature.

Internal and external aspects of administrative law

Administrative law is not all of a kind. The law regulating the internal aspects of administration is different from that which relates to its external aspects. The former comprises such matters as the relations between the government and its officials, the relations between a local authority and its committees, or between a central department and a local authority. The external aspects of public administration, on the other hand, deal with the relations between the administration and private persons or interests. In practice the internal and external aspects are often closely interwoven, and legal provisions of both kinds exist side by side in the same statute. Thus, a law dealing with education may modify the administrative organization of the education service and also regulate the relations between parents and the school authorities.

Another distinction can be made between the type of administrative law that consists of a command addressed by legislation to the citizen, requiring him to act or to refrain from acting in a certain way, and the type that consists of a direction addressed to the administrative authorities telling them how they must or may act. When a rule of administrative law takes the form of an unconditional command addressed to the citizen, a criminal penalty is usually attached to its violation. The application of the penalty is entrusted to the criminal courts, which can review any administrative action of which complaint is made, in the course of the enforcement procedure.

The need for legal safeguards over public administration. Statutory directions addressed to the executive authorities may impose absolute duties, or they may confer discretionary powers that authorize the administrative organ to take action of a specified kind in certain circumstances if it thinks fit. There are many variations of these types. Such legislation may give general directions that the executive bodies are to follow in establishing or operating a system of factory inspection or a policy of slum clearance or making a town-planning scheme. The statute lays down the conditions in which it will be lawful for the administration to act and clothes the authorities with the necessary powers, many of which confer large discretions. In legislation of this type, the action of the executive is not confined merely to carrying out the directions of the legislature. The executive often shares in the law-making process by being empowered to issue regulations or ordinances dealing with specified matters not regulated by the statute. This may be regarded as part of the ordinary process by which legislative powers are delegated by the legislature or as an essential feature of modern government, in which many matters are too technical, or too detailed, or too subject to frequent change, or too dependent on unforeseen emergencies to be included in the main body of legislation, which is less easy to change than regulations.

Problems of executive power

Whatever the source of the rule-making power of the executive may be, safeguards are necessary to ensure that it is properly exercised. The regulation must not exceed the delegated powers; its provisions must conform with the aims of the parent statute; prior consultation with interests likely to be affected should take place whenever practicable; and the regulations must not contravene relevant constitutional rules and legal standards. In some states regulations are scrutinized by the council of state before they come into force; in others the parliamentary assembly acts as the watchdog; while yet another method is review by the ordinary courts.

In most countries the executive arm of government possesses certain powers that are not derived from legislation, customary law, or a written constitution. In England the prerogative powers of the crown are now exercised by ministers in regard to such matters as making treaties, declaring war and peace, pardoning criminals, issuing passports, and conferring honours. In Italy, France, Belgium, and other European countries, certain acts concerning the higher interests of the state are recognized as *actes de gouvernement* and are thereby immune from control by any court or administrative tribunal. In the German Empire (1871–1918) the principle that an administrative act carried its own legal validity was accepted at the end of the 19th century by leading jurists. This led to the notion that administration was only loosely bound to the law. This doctrine was rejected in the Federal Republic of Germany (1949), and efforts were made to reduce the area in which the executive was free to act outside the controlling influence of administrative law.

Public administration consists of the execution of public policy by public authorities. It represents, therefore, the means by which the purposes of the state are promoted and realized in action. Administrative law is the legal framework within which public administration is carried on. One of its principal objects is to ensure efficient, economical, and just administration. A system of administrative law that impedes or frustrates administration

would clearly be bad, and so, too, would be a system that results in injustice to the individual. But to judge whether administrative law helps or hinders effective administration or works in such a way as to deny justice to the individual involves an examination of the ends that public administration is supposed to serve, as well as the means that it employs.

In this connection only the broadest generalities can be attempted. It can be asserted that all states, irrespective of their economic and political system, or of their stage of development, are seeking to achieve a high rate of economic growth and a higher average income per person. They are all pursuing the goals of modernization, urbanization, and industrialization. They are all trying to provide the major social services, especially education and public health, at as high a standard as possible. The level of popular expectation is much higher than in former ages. The government is expected not only to maintain order but to achieve progress. There is a widespread belief that wise and well-directed government action can abolish poverty, prevent severe unemployment, raise the standard of living of the nation, and bring about rapid social development. People in all countries are far more aware than their forefathers were of the impact of government on their daily lives and of its potentialities for good and evil.

Problems of bureaucracy

Bureaucracy and the role of administrative law. An inevitable consequence of the expansion of governmental functions has been the rise of bureaucracy. The number of officials of all kinds has greatly increased, and so too have the material resources allocated to their activities, while their powers have been enlarged in scope and depth. The rise of bureaucracy has occurred in the Communist countries, in the dictatorships and Fascist regimes, and in the political democracies. It is as conspicuous in the new states of Africa and Asia as among the highly developed countries of western Europe or North America. A large, strong, and well-trained civil service is essential in a modern state, irrespective of the political character of its regime or the nature of its economy.

Fear of the maladies that tend to afflict bureaucracy has produced a considerable volume of protest in some countries; and, even in those where opposition to the government or the party in power is not permitted, criticism and exposure of bureaucratic maladministration are generally encouraged.

Bureaucratic maladies are of different kinds. They include the overdevotion of officials to precedent, remoteness from the rest of the community, inaccessibility, arrogance in dealing with the general public, ineffective organization, waste of manpower, procrastination, an excessive sense of self-importance, indifference to the feelings or convenience of citizens, an obsession with the binding authority of departmental decisions, inflexibility, abuse of power, and a reluctance to admit error. Many of these defects can be prevented or cured by good management and careful training. A whole range of techniques is now available for this purpose, including public relations, work–study programs, organization and management (or methods), operational research, and the social survey.

Administrative law has a valuable contribution to make as an instrument for controlling the bureaucracy. In social-democratic regimes, political and judicial control of administration are regarded as complementary, but distinct. The former is concerned with questions of policy and the responsibility of the executive for administration and expenditure. The latter is concerned with inquiring into particular cases of complaint. Administrative law does not include the control of policy by ministers or the head of state. In the people's democracies of eastern Europe, however, this distinction does not exist. The control exercised by an elected council or a presidium over a similar body at the next lower level of government is regarded as a form of legal control over administration. Internal methods of control are also regarded as falling within the ambit of administrative law. These include an appeal from the decision of an official within the same organization or an appeal to a higher administrative unit.

Socialist fusion of political and legal control

A distinctive feature of the Socialist countries of eastern Europe is that there is no definition of the powers of governmental organs at different levels. They are all assumed to be unlimited in scope but always subject to an equally unlimited right of intervention and restraint by the corresponding organ at the next higher level of government. This contrasts with the explicit definition of powers at each level of government that is found in the Western-type democracies. The admixture of political and legal control results in administrative law having a loose and imprecise meaning in the people's democracies of eastern Europe.

JUDICIAL REVIEW OF ADMINISTRATION

Judicial review of administration is, in a sense, the heart of administrative law. The aspects of an official decision or an administrative act that may be scrutinized by the judicial process are the competence of the public authority; the extent of its legal powers; the adequacy and fairness of the procedure; the evidence considered in arriving at the administrative decision and the motives underlying it; and the nature and scope of the discretionary power. An administrative act or decision can be invalidated on any of these grounds if the reviewing court or tribunal has a sufficiently wide jurisdiction. There is also the question of responsibility for damage caused by the public authority in the performance of its functions.

The nature and extent of judicial review of administration, as well as the institutions by which it is undertaken, vary greatly among the countries of the world. History, politics, and constitutional theory have all played a part in determining their position. There are, broadly, three major systems that can be identified (though it should be recognized that around the world there are countless variations on these prototypical systems). These three systems may be described as the common-law model; the French, or council of state, model; and the procurator model.

The common-law system. *Origins.* The common-law system originated in England, where the constitutional struggle in the 17th century played a decisive part in determining the relations between the courts and the executive. The Stuart kings came into open conflict with the judges over their right to decide questions affecting the royal power and even to pronounce an independent judgment in cases in which the king had an interest. Francis Bacon, in his essay *Of Judicature* (written in 1612), put the royalist point of view when he declared that the judges should be "lions, but yet lions under the throne." "It is a happy thing in a state," he wrote, "when kings and states do often consult with judges; and again, when judges do often consult with the king and state: the one, when there is matter of law intervenient in business of state; the other, when there is some consideration of state intervenient in matter of law." The subordination of the judicature to the royal will was strongly resisted by Chief Justice Sir Edward Coke, Bacon's great rival, who refused to comply with James I's wishes in a number of cases in which the royal prerogative was involved. The King harangued the judges more than once on their duty to respect the royal prerogative and power.

Independence of the judiciary

In the constitutional conflict that took place later in the 17th century, the judges and the lawyers made common cause with Parliament against Charles I. An essential feature of the settlement was to establish the independence of the judges. Henceforth there was to be one system of law to which all would owe obedience. Certain consequences resulted from this conflict. First, the executive possessed no inherent powers other than the carefully defined and limited prerogatives of the crown, recognized by the law. Second, the judges were expected to protect the subject against the executive. A more intangible consequence was the belief that "government" and "law" were often thought to be opposed to one another. The conflict between James and the judges survived in an antagonism between the legal profession and the executive, particularly the civil service.

These developments established the principle that the executive should never interfere with the judiciary in the

exercise of judicial functions. This was, indeed, almost the only strict application in England of the doctrine of the separation of powers. On the other hand, it was regarded as right and proper that the judiciary should interfere with the executive whenever a minister or a department was shown to have acted illegally. In this way the concept of the rule of law came gradually to be identified with the idea that the judges, in ordinary legal proceedings in the ordinary courts, could pronounce upon the lawfulness of the activities of the executive. Any attempt to divide the seamless web of the law, any suggestion of a distinction between public and private law, appeared to the generations brought up on this traditional view to destroy its universality and its power to keep the executive within bounds.

The principle that all public authorities are liable to have the lawfulness of their acts and decisions tested in the ordinary courts was applied in all the countries in which the common law was transplanted. In the United States the same conception prevailed, despite the much stricter interpretation given to the doctrine of the separation of powers by the Founding Fathers and the embodiment of the doctrine in the federal Constitution and the state constitutions. A complete separation of powers was not considered to be feasible by the framers of the Constitution, and they therefore modified the doctrine by adding to it the notion of checks and balances, whereby each of the three branches of government would be checked and balanced by the others. This actually strengthened the power of the courts to review the actions of the executive. Elsewhere in the common-law world, the role that the courts were expected to play in reviewing administration was applied without any public debate concerning the separation of powers or the need to protect liberty and prevent despotism by a system of checks and balances.

American checks and balances

Modification of the common-law system. The common-law system has been extensively modified in the course of the 20th century. It never corresponded with the realities of the situation in England because, prior to the Crown Proceedings Act, 1947, it was not possible to sue ministers and their departments in tort because of the legal doctrine that "the king can do no wrong." Moreover, the development of the social services has been accompanied by the creation of a large number of administrative tribunals to hear and determine disputes between a government department and a citizen. The jurisdiction of these tribunals is of a specialized and narrowly circumscribed character and relates to such functions as social insurance and social assistance, the National Health Service, rent control, assessment of property for local taxation, the compulsory acquisition of land by public authorities, and the registration of children's homes. There are about 2,000 of these tribunals, and, although since 1958 a permanent Council on Tribunals appointed by the lord chancellor exercises a general supervision over them, they remain an unsystematic and uncoordinated movement. The reasons for their existence are that they provide a method of administrative adjudication that is far cheaper, more informal, and more rapid than that offered by the courts; the members are persons possessing special knowledge and experience of the activity in which the disputes arise; they do not have to follow the strict and complex rules of evidence that prevail in the courts; and it is possible to introduce new social standards and moral concepts to guide their decisions. These tribunals have acquired general approval for the quality and impartiality of their work. An appeal on a question of law lies in most instances from the decision of an administrative tribunal to the High Court of Justice. There is still no comprehensive administrative jurisdiction in England to provide an opportunity for judicial review over the whole field of executive action and decision.

British development of administrative tribunals

In Australia a similar movement has taken place with the growth of a large number of administrative tribunals covering many different spheres of public administration, such as the regulation of industrial conditions; many branches of the public service; the award of pensions, allowances, and other state grants; the supervision of social conditions (in regard to town planning, the censorship of films, fair rents, and so on); the licensing of occupations calling for special skill or public responsibility; the supervision of trade, transport, and marketing; the assessment of taxes, rates, or duties; the protection of industrial design, patents, and copyrights; and compensation for interference with private-property rights in the public interest. As in England, the growth of these tribunals has been sporadic, and no attempt has been made to introduce a systematic or comprehensive administrative jurisdiction. A similar situation has occurred in some of the other Commonwealth countries.

In the United States the courts have been much more far-reaching and comprehensive in reviewing administration than in England. Nevertheless, much adjudication is now performed by public authorities other than the courts of law. The movement toward administrative tribunals began with the Interstate Commerce Act (1887), establishing the Interstate Commerce Commission to regulate and control railways and other carriers. The law introduced a new type of federal agency outside the framework of the executive departments, with a large measure of independence from the president. This was followed by a number of other regulatory commissions of a similar type, such as the Federal Trade Commission, the Federal Power Commission, the Federal Communications Commission, the Securities and Exchange Commission, the Civil Aeronautics Board, and the National Labor Relations Board. These bodies have had administrative, legislative, and judicial functions delegated to them by Congress, and the doctrine of the separation of powers can no longer be successfully invoked to challenge the constitutionality of such legislation. The regulatory commissions are often described by American jurists as administrative tribunals.

American development of regulatory commissions

Thus, in the United States, as in other parts of the Anglo-American common-law world, the concept of the exclusive exercise by the ordinary courts of all judicial powers and of the absence of special administrative tribunals has been substantially modified by these developments.

The council of state system. *The French system.* In France the separation of powers was given a place of honour in the Declaration of the Rights of Man and of the Citizen (1789), but a different interpretation was given to the doctrine. In the French view, if a court were permitted to review an administrative act or decision, it would contravene the separation of powers as much as if the executive could override the decision of a court. Just as an appeal from a court lies to a higher court, so an appeal from an administrative authority should lie to a higher administrative authority. Only thus would the true separation of powers be observed.

Herein lies the explanation of administrative law as a system of law separate from the body of law administered in the courts. A law of August 1790 declared that the judiciary was distinct from and would always remain separated from the executive. It forbade judges, on pain of dismissal, to interfere in any way with the work of administrative bodies. In October 1790 a second law stated that under no circumstances should claims to annul acts of administrative bodies fall within the jurisdiction of the courts. Such claims should be brought before the king as head of the general administration.

French conception of separation of powers

Napoleon laid down the basic organization of the modern Conseil d'État (Council of State). Among the functions accorded to it by the constitution of the year VIII (December 1799) was that of adjudicating in conflicts that might arise between the administration and the courts and on any matters previously left to the minister's discretion that ought to be the subject of judicial decision. In 1806 a decree created a Judicial Committee of the Council to examine applications and report thereon to the General Assembly. These were the forerunners of the administrative jurisdiction that was not clearly established until May 24, 1872, when a law delegated to the Conseil d'État the judicial power to make binding decisions and recognized the Conseil as the court in which claims against the administration should be brought.

The Conseil d'État is and always has been part of the administration. It has for long had the task of giving legal advice to the government on bills, regulations, decrees, and administrative questions. It is this that long led foreign jurists into believing that, when sitting as a court, its decisions would inevitably be biassed in favour of the executive. Nothing could be further from the truth, and today the Conseil is universally recognized as an independent court that provides French citizens with exceptionally good protection against maladministration. Suits that are directed against the French administration are heard in the Section du Contentieux, or Judicial Division.

French system of administrative courts

The Conseil d'État is the final authority in administrative disputes. Owing to the immense volume of work falling on it, the former prefectoral councils were transformed in 1953 into administrative tribunals of first instance, and the professional qualifications and career prospects of their members were improved. There are 22 of these tribunals in France, together with a special one for Paris. The great majority of cases go before them, and the Conseil d'État is the court of first and last instance only rarely when it is specially designated for that purpose.

If difficulty or doubt arises as to whether a case falls within the administrative jurisdiction or that of the ordinary courts, the question is resolved by the Tribunal des Conflits. This is a court specially established for the purpose, consisting of five judges from the Cour de Cassation (the highest civil court) and five from the Conseil d'État. The minister of justice, in his capacity as keeper of the seals (*garde des sceaux*), may preside and cast a tie-breaking vote.

Several other countries have established councils of state as a result of French influence. Among them are Italy, Greece, Belgium, Spain, Turkey, Portugal, Egypt, and Lebanon. It must be stated, however, that in no other country has a council of state acquired the high status, powers, authority, or prestige possessed by that of France.

The German system. Germany traditionally has had no council of state per se, but the Federal Republic of Germany today does have a fully articulated system of special administrative courts. In the states, or *Länder*, there are lower administrative courts and superior administrative courts, and for the federation there is the Federal Administrative Court, which acts mainly as a court of appeals from the superior administrative courts in the *Länder* and even from the lower administrative courts in certain circumstances. The Federal Administrative Court serves also as a court of first and last instance in public law disputes between the federation and the *Länder* or between two or more *Länder;* it hears petitions by the federal Cabinet on declarations that an association is prohibited under the Basic Law of the Federal Republic; petitions against the federation in matters concerning the diplomatic or consular service; and cases concerning the business of the Federal Intelligence Service.

German system of administrative and civil jurisdictions

A *Land* administrative court possesses jurisdiction concerning the acts of the *Länder* administrative authorities and also complaints against officers of the federal government located in the *Länder*. Some of the highest federal organs are exempt from the *Länder* courts. Few cases go beyond the *Länder* supreme administrative courts.

Recourse to an administrative court is available for public law disputes unless the matter has been assigned to another court by federal legislation. The Administrative Courts Code holds that property claims arising from services for the common good and restitution claims arising from violation of duties under public law shall be heard by the ordinary courts. In other words, the German system is complicated by the rule that only the ordinary civil courts can award damages against an official or the executive arm of government. Hence the distinction between the ordinary courts and the administrative courts depends on the remedy sought and not on the subject matter of the dispute or the nature of the parties. The jurisdiction of the administrative courts in West Germany is thus less comprehensive and clear-cut than in France.

The procurator system. The third system for ensuring administrative legality is the Procuracy, an institution founded in Russia by Peter the Great in 1722, who intended it to be the "eye of the tsar." Catherine II issued a directive in 1764 stating that the procurator general and his staff were to supervise the execution of the laws in the provinces, to ensure justice, and prevent abuses. This was designated general supervision. In 1864 the Procuracy was relieved of its responsibility for supervising administration, and its functions were confined to judicial matters, such as acting as public prosecutor in all criminal cases and conducting them on behalf of the government and the law.

The Procuracy was abolished in November 1917 but revived in 1922. The Soviet constitution charges the procurator general with the general duty of supervising the observance of the law by all ministries and institutions subordinate to them as well as by individual officials and citizens. The procurator general is appointed by the Supreme Soviet for seven years. He appoints subordinate procurators at all administrative levels from union republic to district and town.

Fluctuations in the functions of the Soviet procurator

The functions actually performed by the procurator have undergone many fluctuations and vicissitudes since 1922. His duty to act as a public prosecutor has remained stable but sometimes has been exercised mainly to enforce party policies or programs on recalcitrant citizens rather than to punish public officers or authorities for breaches of the law. Sometimes the task of general supervision has been emphasized; at other times it has been abandoned in favour of more urgent tasks connected with economic policy or the conduct of World War II. The present position of the Procuracy is laid down in a union law of May 24, 1955, and a decree of the Presidium of the Supreme Soviet made on April 7, 1956.

The procurator is not the president of a court or a tribunal but a watchdog of legality. His organization comprises a department for general supervision; a bureau of investigation for the supervision of preliminary inquiries in criminal matters; a department for the supervision of investigations carried out by the State Security Service (KGB); departments to supervise criminal and civil proceedings in the courts; a department to supervise prisons, compulsory-labour centres, and the like; and departments for statistics, administration, and research.

General supervision is defined by Soviet writers on administrative law as meaning supervision by the procurators over legality in administration. Procurators are expected to see that the laws are strictly observed, to oppose their violation by anyone whatsoever, to protect the citizens, and ensure that they fulfill their duties. The law of May 24, 1955, requires the Procuracy to ensure strict conformity of the regulations or decisions issued by ministries, departments, and their subordinate establishments and enterprises and cooperative and other public organizations with the constitution and laws of the U.S.S.R. and the union and autonomous republics and the decrees of the Council of Ministers of the U.S.S.R. and the republics. They are also to ensure strict execution of the laws by officials and citizens. The procurator is concerned solely with the legality, not with the efficiency or expediency, of administrative action.

Since the Procuracy is not a court, it cannot make a binding decision. The normal procedure (apart from cases of dereliction involving a criminal prosecution) is for the procurator to protest against any illegality that he may detect or that is brought to his notice or to initiate disciplinary action against an erring official. Every citizen has a right to lodge a complaint with the procurator, and denunciation and exposure of bureaucratic abuses are encouraged by the Communist Party. The fact that the Procuracy cannot make a binding decision does not necessarily prevent it from being an effective organ for securing administration according to law. Neither the French Council of State nor the courts in any country can enforce judgments against the central gov-

ernment, but this does not prevent the decisions of the council or declaratory judgments of the courts from being observed almost as a matter of course.

The other Communist regimes of eastern Europe have established procuracies based on the Soviet model. In Poland an additional institution to maintain administrative legality is the Supreme Chamber of Control, which is independent of the government and subordinate only to the legislature and the Council of State, a political body quite different from the French model. The functions of the Supreme Chamber of Control involve exercising general supervision over public administration, taking into account legality, economy, and opportuneness.

ADMINISTRATIVE PROCEDURE

Importance of administrative procedure

The administrative state is characterized not only by public administration but by the elaboration of what is termed the administrative process. The essence of the process is an orderly and ascertainable procedure that may be prescribed by legislation or regulations or laid down by the courts or specified by the administrative authority itself or by a higher level of government.

Administrative procedure is of great importance in any system of administration according to law. Well-considered procedure enables responsibility to be fixed on a responsible officer or body at each stage of the administrative process. It can safeguard the rights of citizens and protect the executive against the criticism of having acted in an arbitrary manner. It can ensure regularity and consistency in the handling of individual cases. Much depends, however, on the quality and purpose of the procedural requirements.

Codes of administrative procedure. Some states have general codes of administrative procedure embodied in legislation. Among them are Austria, Czechoslovakia, Poland, Yugoslavia, Spain, and the United States. The Yugoslav law regulating the general administrative procedure is a lengthy statute containing 393 sections. It enacts that the organs of the state administration and other state organs have to proceed according to this law whenever, applying rules directly, they decide in administrative affairs on the rights, obligations, or legal interests of individuals, corporate bodies, or other parties. The law deals in great detail with the right of appeal to a higher administrative body or, exceptionally, to another agency at the same level of government. Appeals from a people's committee go to the corresponding committee at the next higher level, while appeals from an assembly go before the next higher assembly. Each enactment is expressed in clear and precise language.

In addition to an appeal to the next higher level of authority in the administrative hierarchy, the Yugoslav code of procedure provides for an appeal on a question of law to the courts. Judicial control of the legality of administrative actions is vested in the supreme courts, which have separate divisions to hear administrative litigation. The intention is that decisions on the merits of administrative acts will normally be dealt with by the administrative hierarchy, since they are experts in public administration and familiar with its tasks and needs, while the legal aspects will normally be decided on appeal to the courts. Only in exceptional cases will the Supreme Court have power to consider the merit of the act.

Apart from the few states that have separate codes of this kind, there is in all other developed systems a mass of administrative law prescribing procedure. Much of it is to be found in the laws and regulations governing particular functions of government, such as, for example, taxation, public health, education, and town planning.

Scope of procedural rules

It is not possible to summarize the immense variety and scope of procedural requirements. Among the matters usually dealt with are the manner in which the administrative machinery is to be set in motion for a specified purpose, a reference to the relevant legal enactments, information about the rights of interested persons, a statement of how, when, and where an appeal may be lodged, any time limits that must be observed, the conditions to be satisfied by objectors, and the right of legal representation. The leading treatise on U.S. administrative law contains many chapters devoted entirely to such topics as rule making, requirement of opportunity to be heard, adjudication procedure, examiners, bias, evidence, official notice, findings, reasons, and opinions, all of which are wholly, or in part, procedural matters at some stage of the administrative process.

Natural justice and administrative procedure. In the common-law systems, the doctrine of natural justice has had a substantial influence on administrative procedure. The doctrine has evolved into two rules: one is that a man may not be judge of his own cause; the other (summed up in the maxim *audi alteram partem*, "hear the other side") is that a person shall not be dealt with to his material disadvantage, whether of person or property, or removed from or disqualified for office, without being given adequate notice of what is alleged against him and an opportunity to defend himself.

An indirect result of the second rule of natural justice is the widespread adoption of the public hearing by government departments (and in the U.S. by the regulatory commissions) in deciding matters involving individual or corporate rights. In Great Britain a public inquiry is now a recognized device for informing ministers and their departments on particular matters before a decision is reached or an appeal decided in a great variety of spheres, such as town planning, the creation of a new town, slum clearance, and compulsory acquisition of land.

In 1957 the Franks Committee was appointed in Great Britain by the Lord Chancellor to consider and make recommendations on the constitution and working of administrative tribunals and administrative procedures such as the holding of an inquiry or hearing by or on behalf of a minister. The committee declared that the work of administrative tribunals and of public inquiries should be characterized by openness, fairness, and impartiality, and their report applied these aims in great detail. Its recommendations were largely accepted and resulted in the Tribunals and Enquiries Act of 1958.

Although good administrative procedure is of great value, it is possible to exaggerate its importance. When that happens, the main test that an administrative decision subject to judicial review has to satisfy is whether the correct procedure was observed, with little or no attention being given to the substance of the decision.

Appropriateness of judicial review

Scope of judicial review. Judicial review, whether conducted by administrative tribunals or the ordinary courts, is the most appropriate method of inquiring into the legal competence of a public authority, the fairness and adequacy of the procedure it follows, the evidence on which it relied in a particular decision, the motives or considerations on which it was based, and the extent of the grant of any discretionary power it may possess. It is less effective as a method of inquiring into the wisdom, expediency, or reasonableness of administrative acts, and courts and tribunals are unwilling to substitute their own decisions for that of the responsible authority.

The extent to which discretionary powers are liable to be circumscribed by judicial review varies greatly. Sweden and France have gone furthest in subjecting the exercise of all discretionary powers, other than those relating to foreign affairs and defense, to judicial review and potential limitation. Moreover, in both these countries, the administrative tribunals will consider the merits of a decision.

It is of course impracticable to subject every administrative act or decision to investigation, for this would slow up the work of government to an unacceptable extent. The complainant must, therefore, always make out a prima facie case showing that some form of maladministration has occurred.

Administrative tribunals and courts can annul unlawful decisions, allow appeals, declare the law, and award damages. They cannot compel the state to act in a particular way because sanctions cannot be imposed on the government, which itself controls the use of force. Such remedies as an injunction, an order for specific performance, or an order for mandamus will not lie against

the central government. These inhibitions are of less practical importance than might be thought because nearly all governments (including those that have come to power by revolution) are eager to proclaim the lawfulness of the regime and seldom disregard the decisions of an authorized court or tribunal.

THE OMBUDSMAN

The type of official known in Scandinavia as an ombudsman must now be included within the scope of administrative law. The office originated in Sweden when the post of ombudsman for civil affairs was created by the constitution of 1809. The holder of the office is appointed by Parliament; he enjoys complete independence and is responsible only to the law.

Independence of the Swedish ombudsman

The Swedish ombudsman is entrusted with a general supervision of the entire sphere of civil affairs, including the judicature, the police, prisons, and the public administration, both central and local, but excluding ministers. He can act as a public prosecutor; as a receiver of complaints from aggrieved citizens; as an inspector undertaking careful investigations into such institutions as jails, mental hospitals, institutions for delinquent children, and retreats for alcoholics to discover if they are being administered in accordance with the law. In the past the ombudsman was mainly concerned with errors in the courts, including procrastination, improper conduct, and similar instances; currently the bulk of his work is directed to the civil administration. (A separate ombudsman supervises military affairs.)

Apart from occasional prosecutions, the ombudsman has little positive power other than the right to inspect and to demand the fullest information. But he is entitled to comment, to criticize, and to make recommendations as to the correct interpretation of the law. He can recommend that the government pay compensation to a complainant. He can propose changes in the law. He can expose maladministration.

Only ministers and the king are immune from the scrutiny of the ombudsman.

The idea of an official appointed for the express purpose of receiving and inquiring into complaints of citizens has taken root in several countries. The other Scandinavian countries have followed the Swedish example, and ombudsmen have also been appointed in, for example, New Zealand, the United Kingdom, West Germany, Israel, several U.S. states, and certain Canadian provinces and Australian states. A number of cities have appointed municipal ombudsmen, and in Great Britain a commission serves as a complaints mechanism for local governments.

Some specialized ombudsmen have been appointed in the U.S. to safeguard the rights of prisoners to medical treatment, and in Britain an ombudsman has been appointed to inquire into complaints concerning the National Health Service. In Israel the police have an office of public complaints and a military ombudsman. There are considerable differences in the powers and status of the ombudsman in these countries.

The ombudsman should be regarded as a part of the system of administrative law for scrutinizing the work of the executive. He is the appointee of the supreme legislative assembly and not of the executive. He enjoys a large measure of independence and personal responsibility and

Ombudsman as guardian of law

is primarily a guardian of the law. His function is to safeguard the interests of citizens by ensuring administration according to law, discovering instances of maladministration, and eliminating defects by various methods that include bringing pressure to bear on the responsible authority, publicizing a refusal to rectify injustice or a defective administrative practice, bringing the matter to the attention of the legislature, and instigating a criminal prosecution or disciplinary action.

There is no doubt about the value of the institution in the states in which it has been established. Part of the ombudsman's usefulness lies in his ability to reassure citizens who believe they have been unjustly treated that careful inquiry into their complaints shows their suspicions to be groundless.

CONSEQUENCES OF LEGAL SCRUTINY OF ADMINISTRATION

Most of this survey has been concerned with the control of public administration by methods other than those regarded as political, such as popular elections, legislative control of the executive, pressure groups, etc. The justification for this emphasis is that, as the powers of the administrative state increase, the need to safeguard the rights and interests of the citizens both individually and collectively becomes greater also. The safeguards concern both the rights of the individual in his dealings with the powerful machinery of the contemporary state and the interest of the community in ensuring that administration is carried on according to law. None of the methods of legal scrutiny, control, and prescription described above is intended to undermine the strength, effectiveness, and vigour of the administrative state. They are designed rather to reveal its shortcomings and to find remedies for its defects.

Administrative law as a supplement to political controls

Administrative law can make the activities of the modern administrative state more acceptable to the citizens by preventing or correcting maladministration, by laying down standards of behaviour for ministers and officials, by insisting on just procedures, by refusing to permit discretionary power to be used for unauthorized purposes, by ensuring that equality of treatment is a condition of legality, and in various other ways. By becoming more acceptable to the citizens, public administration can also become more efficient.

BIBLIOGRAPHY. There are no works covering the whole subject of administrative law in its differing forms in many countries. The position in a number of European states is well described in B. CHAPMAN, *The Profession of Government: The Public Service in Europe* (1959). The general problem of controlling the bureaucracy is briefly considered in F. MORSTEIN MARX, *The Administrative State* (1957); and W.A. ROBSON, *The Governors and the Governed* (1964). SIR JOHN MARRIOTT, *The Crisis of English Liberty* (1930), describes the constitutional struggle in England between the king and the judiciary in the 17th century. For the present system in England, see H.W.R. WADE, *Administrative Law*, 3rd ed. (1971); J.A.G. GRIFFITH and H. STREET, *Principles of Administrative Law*, 4th ed. (1967); S.A. DE SMITH, *Judicial Review of Administrative Action*, 2nd ed. (1968); and J.F. GARNER, *Administrative Law*, 3rd ed. (1970). W.A. ROBSON, *Justice and Administrative Law*, 3rd ed. (1951), is a standard work on the rise and purpose of administrative tribunals. The leading work on American administrative law is K.C. DAVIS' monumental *Administrative Law Treatise*, 4 vol. (1958); the complex question of the control of administrative discretion in the United States is examined by K.C. DAVIS, *Discretionary Justice* (1969). A readable account of the American system may be found in B. SCHWARTZ, *An Introduction to American Administrative Law*, 2nd ed. (1962). See also W. GELLHORN and C. BYSE (eds.), *Administrative Law: Cases and Comments*, 4th ed. (1960). The French system is favourably appraised in C.J. HAMSON, *Executive Discretion and Judicial Control* (1954); a reliable description of its principles is contained in L. NEVILLE BROWN and J.F. GARNER, *French Administrative Law* (1967); a comparison between the Anglo-American system and the French is given by B. SCHWARTZ in *French Administrative Law and the Common-Law World* (1954). Leading treatises by eminent French jurists are G. VEDEL, *Droit administratif*, 2nd ed. (1961); M. WALINE, *Droit administratif*, 4th ed. (1968); and M. LONG, P. WEIL, and G. BRAIBANT, *Les Grands Arrêts de la jurisprudence administrative*, 5th ed. (1969). For Australia, see D.G. BENJAFIELD and H. WHITMORE, *Principles of Australian Administrative Law*, 3rd ed. (1966). On West Germany, E. FORSTHOFF, *Lehrbuch des Verwaltungsrechts*, 9th rev. ed. (1966); and H.J. WOLFF, *Verwaltungsrecht*, 4th rev. ed. (1961), are reliable works. The procuracy can be studied in G.G. MORGAN, *Soviet Administrative Legality* (1962); and more briefly in L.B. SCHAPIRO, *The Government and Politics of the Soviet Union*, 2nd rev. ed. (1967). Three wide-ranging books on the ombudsman are a symposium edited by DONALD C. ROWAT, *The Ombudsman: Citizen's Defender*, 2nd ed. (1968); and WALTER GELLHORN'S two volumes, *Ombudsmen and Others* (1966) and *When Americans Complain* (1966). An excellent work on Scandinavia is NILS HERLITZ, *Nordisk offentlig rätt*, 2 vol. (1958; Eng. trans., *Elements of Nordic Public Law*, 1969). The only accessible text in English of the Yugoslav procedural code is a translation of LEONIDAS PITAMIC, *Zakon o opštem upravnom postupku* (1957; Eng. trans., *The Yugoslav Law on General Administrative Procedure*, 1961).

(W.A.R.)

Adriatic Sea

Constituting an arm of the north central Mediterranean Sea, the Adriatic Sea thrusts northwestward between the eastern Italian coast and that portion of the western coast of the Balkan Peninsula occupied by Yugoslavia and Albania. It has been of great importance in the historical development of Mediterranean Europe and is of considerable scientific interest in itself.

There is a striking contrast between its two shores. The Italian coast is relatively straight and continuous, having no islands, whereas that of the Balkan Peninsula is full of both large and small islands, generally oblong in shape and running parallel to the continental coast. Many tortuous straits also form inlets between the islands similar to those of the Norwegian fjords and make the coastline very intricate. In overall dimensions, the Adriatic Sea can be viewed as a gigantic canal, about 500 miles long, penetrating far into the continental land mass from the Ionian Sea, which it leaves through the Strait of Otranto. Its average width is 100 miles, and its total area is 50,590 square miles (131,050 square kilometres).

Modern scientific study of the Adriatic has been carried out mainly under the auspices of a dozen or so Italian scientific institutes and their Yugoslavian equivalents. Each nation has a number of research vessels at its disposal.

The physical background. The depths of the Adriatic near to its shores bear a close relationship to the physiography of the adjacent coasts. Wherever such coasts are high and mountainous, the nearby sea depths are considerable, as in the case of the Istrian and Dalmatian areas of Yugoslavia. Where, on the contrary, low and sandy shores are found, the nearby sea is shallow, as in the case of the sea in the vicinity of Venice or, farther south, near the delta of the Italian Po River. Generally speaking, depths are small all along the Italian coast. The maximum depth of the Adriatic Sea is situated south of the central area, reaching 4,035 feet (1,324 metres). The average depth is 1,457 feet (444 metres).

Two sea bottom types

The Adriatic has two types of rather special sea bottoms, difficult to arrange in a rigorous classification but very common in the Mediterranean, namely, inlet-derived sediments and heat-altered sediments of the sea bottom proper. In general, the seabed consists of a yellowish mud and sand, containing fragments of shells, fossil mollusks, and corals. The main winds prevailing in the area are the bora, a strong northeast wind that blows from the nearby mountains into the sea, and a southeasterly wind named the sirocco that is less troublesome from a navigational point of view. During the six winter months, bora and sirocco alternate, with or without an interval of a few days calm. The tides of the Adriatic, which have been intensively studied, follow a complicated pattern, sweeping into the region from the south and being linked with those of the Ionian Sea.

The tidal range is about three feet, in contrast to the general Mediterranean tidal range of about 0.9 foot. The surface currents are chiefly influenced by the blowing winds, with currents spurred by north winds reaching a speed of four miles per hour.

Temperatures in the surface layers of the sea reach 75°–77° F (25° C) during the month of August, and the minimum readings, some 50° F (10° C), are usually reached during January and February. In the northern Adriatic, river mouth temperatures are even lower because the waters are cooled by melting ice and snow. At greater depths (820–980 feet) the maximum temperatures fluctuate around 57° F (14° C) while minimum temperatures are about 52° F (11° C).

Salinity in the south Adriatic is 38–39‰ (parts per thousand), decreasing further north to 36–37‰, and falling to 33‰, or even as low as 25‰, in the North Adriatic, as a result of the inflow of fresh waters.

Marine life. The Adriatic Sea, like the Mediterranean in general, is very deficient in life, and nutrient content, as indicated by the amount of phosphates and nitrates, is very low.

Three main areas of marine life may nevertheless be recognized. In the northern Adriatic area important winter cooling and a lowered salinity further impoverish the typical Mediterranean marine life. In the middle Adriatic area, life is much richer than further north, while the southern Adriatic area comprising the coastal regions on both sides of the central area has its own distinctive forms of life.

BIBLIOGRAPHY. M. BERNARD, "Recent Advances in Research on the Zooplankton of the Mediterranean Sea," in H. BARNES (ed.), *Oceanography and Marine Biology*, vol. 5, pp. 231–255 (1967), various investigations considered from the points of view of the particular disciplines or methods employed; H. GAMULIN-BRIDA, "The Benthic Fauna of the Adriatic Sea," *ibid.*, pp. 535–568, an examination of the origin, biogeography, and units of the animal life at the bottom of the Adriatic Sea; M. BULJAN, "The Nutrient Salts in the Adriatic Waters," *Acta Adriatica*, 5:233–238 (1953), an examination of phosphates and nitrates, mainly in the Adriatic Sea; M. ZORE-ARMANDA, "Temperature Relations in the Adriatic Sea," *ibid.*, 13:1–51 (1969).

(A.D.H.)

Adult Education

In the modern period education has come to be strongly identified with organized preparation of the young in schools and colleges for both their personal development and their subsequent social and economic roles in society. Throughout recorded history, however, it has been recognized that to deny oneself of later adult education is to deny oneself of a great part of education that is meaningful only to grown, experienced men and women. In brief, the term adult education can thus be defined as the organized provision of facilities to enable men and women to enlarge and interpret their experiences as adults. Such facilities may be provided—and, in different places and at different times, have been provided—by voluntary action and independent agencies, by state and municipal authorities, and by commercial enterprises.

A more explicit definition is contained in a 1970 report of the National Institute of Adult Education (England and Wales), which stated:

> In the widest sense adult education is any kind of education for people who are old enough to work, vote, fight and marry and who have completed the cycle of continuous education, [if any] commenced in childhood. They may want to make up for limited schooling [or for no schooling], to pass examinations, to learn the basic skills of trades or professions or to master new working processes. They may turn to it because they want to understand themselves and their world better and to act in the light of their understanding, or they may go to classes for the pleasure they can get from developing talents and skills—intellectual, aesthetic, physical or practical. They may not even "go to classes": they may find what they want from books or broadcasts or take guidance by post from a tutor they never meet. They may find education without a label by sharing in common pursuits with like-minded people. (National Institute of Adult Education, *Provision for Adult Education*, 1970.)

Varieties of adult education

Adult education thus comprehends such diverse modes as independent study consciously pursued with or without the aid of libraries; broadcast programs or correspondence courses; group discussion and other "mutual aid" learning in study circles, colloquia, seminars or workshops, and residential conferences or meetings; and full- or part-time study in classes or courses in which the lecturer, teacher, or tutor has a formal leading role. Types of adult education can also be classified as follows:

1. Education for vocational, technical, and professional competence. (Such education may aim at preparing an adult for a first job or for a new job, or it may aim at keeping him up to date on new developments in his occupation or profession.)
2. Education for health, welfare, and family living. (Such education includes all kinds of education in health, family relations, consumer buying, planned parenthood, hygiene, child care, and the like.)
3. Education for civic, political, and community competence. (Such education includes all kinds of education relating to government, community development, public and international affairs, voting and political participation, and so forth.)

4. Education for "self-fulfillment." (Such education embraces all kinds of liberal education programs: education in music, the arts, dance, theatre, literature, arts and crafts, whether brief or long term. These programs aim primarily at learning for the sake of learning rather than at achieving the aims included in the other categories.)

5. Remedial education: fundamental and literacy education. (Such education is obviously a prerequisite for all other kinds of adult education and thus, as a category, stands somewhat apart from the other types of adult education.)

In reference to the fifth category, adults frequently need to compensate for inadequacies of earlier education. If they are not remedied, they inhibit recourse to modes of education that are "adult"—adult, that is, in terms of sophistication in modern society and not in terms of age. Such education is required most extensively in societies changing rapidly from a subsistence to an industrial economy and concurrently changing politically and socially. Mass literacy acquires a new importance in these nations of Asia, Africa, and Latin America, and the establishment of universal primary education becomes a social imperative. To prevent a "generation gap" in reading skills and education while an effective school system is being created for the young, governments must attempt to provide parallel facilities for adults. Even in countries with mature systems of childhood education, however, opportunities for higher or even sometimes secondary education are unequal among various regional, occupational, and social groups. Hence there are adult programs for completing high school or preparing for examinations normally taken at the end of secondary school. Even in so highly developed a country as Sweden it has been considered necessary in recent years not only to try to extend secondary education to more children but also to offer equivalent second chances for adults. The implication is that "remedial" adult education is needed in societies in all stages of development and not merely in the so-called developing countries.

BACKGROUND OF MODERN ADULT EDUCATION

Education as a lifelong process

The idea of education as a lifelong process has been expressed by philosophers and educators throughout the centuries. It is explicit in the thought and writings of Confucius, Plato, Sir Thomas More, Comenius, and Benjamin Franklin, to cite only five examples widely separated in time and place. Attempts to give reality to such aspirations and to provide education for large numbers of adults, however, cannot be reasonably dated before the Industrial Revolution—and its accompanying social revolutions—of the late 18th and the 19th centuries. Changes in the ways of living—both occupationally and politically (as the franchise was extended)—produced a need for further education and re-education of adults. In the early and middle years of the 19th century these socioeconomic changes were strongest in Great Britain and the United States; and it is in these countries that adult education received greatest emphasis, although significant developments also occurred in Scandinavia. Even today, the provisions for adult education in such lands as France, Italy, and Latin America are relatively small compared to what they are in the countries that pioneered further education.

Early programs. The rise of organized adult education in Great Britain is generally dated from the late 1790s, when the first adult school and the foundations of the first mechanics' institute were established, in Nottingham and Glasgow, respectively. In the next 50 years adult schools created generally by religious people to reduce adult illiteracy and mechanics' institutes designed to spread among artisans knowledge of science and its applications in industry, all dependent on voluntary effort and voluntary finance, were founded in many parts of England. Both declined later in the 19th century, although the adult schools revived and were active in the 20th, and the institutes often became technical colleges or reading and lecture societies.

In the 1840s the early cooperative societies in Great Britain undertook adult education work that still continues, and the industrial and social conditions that inspired Christian socialism and other humanitarian movements also prompted the founding of people's and workingmen's colleges. The first of these was established in Sheffield in 1842, but the most influential was the London Working Men's College (1854). Others were short lived, and none was residential until Ruskin College was founded at Oxford in 1899 to equip workingmen for their responsibilities as citizens and trade unionists.

Meanwhile, in the United States, the first half of the 19th century was a period of national adjustment and the rich development of educational institutions, each one unique and often unrelated to the others but all providing new channels of learning. Most important were the lyceums, local associations of men and women who had had some formal education and were concerned both with their own further learning and with the creation of a public school system; their education came chiefly through hearing debates and lectures on timely topics. Paralleling the growth of the lyceums and to some extent developing out of them were libraries, museums, mechanics' institutes, lecture series, endowed institutions, publicly supported evening schools, and a host of other institutions, many of which became permanent.

The second half of the 19th century, particularly during and after the Civil War, was another period of America's readjustment and therefore another era of growth for organized adult education. Both the land-grant colleges and the U.S. Department of Agriculture were created, thus laying the groundwork for their later collaboration in the teaching of scientific agriculture and home economics to farmers and their wives. Institutions developed earlier gained strength; the most notable development of this sort perhaps was the growth of the public library, greatly stimulated by the benefactions of the steelman and philanthropist Andrew Carnegie. Sporadic university lecture series for the general public evolved into systematic programs of "university extension"—educational courses for adults who are not full-time students. (The idea that a university owed responsibility to a wider constituency than its enrolled students was first officially endorsed, however, in England at the University of Cambridge in 1873.) A new variation on the lyceums was the "chautauqua," a nationwide movement that was inspired by the example of the annual summer-camp meetings of teachers, social workers, and others at Chautauqua Lake in New York State, where lectures, study workshops, and popular entertainment were offered for moral and intellectual improvement. (The popularity of the chautauquas did not decline until the mid-1920s.)

National variations. Although the British and American programs in adult education were the ones to influence most strongly the development of adult education elsewhere in the world, the folk high schools in Scandinavia represent a noteworthy national development. These residential schools for young adults were first founded in Denmark in the middle of the 19th century to foster a sense of cultural unity, and they became a powerful instrument of national regeneration after the country's defeat by Germany in 1864. The man whose personality and writings stimulated their development in Denmark as well as the other Nordic nations was N.F.S. Grundtvig, who has been called the "father of the folk high school" and whose interests were unquestionably more political and nationalist than industrial or commercial. A theologian as well as an educationist, he stressed the need for a thorough knowledge of the mother tongue and of national and biblical history, and he put forward the idea that the years after 18 are those of real awakening to life. Such was the inspiration for the folk high schools in which young people of every class were to be encouraged to educate themselves and one another.

Past inadequacies of adult education

Even when such efforts in Scandinavia, Britain, the United States, and elsewhere were aimed at the educationally least privileged; however, several factors at first limited their success. The planners took insufficient account of the inadequacies of elementary education, and, in the absence of public subsidies or other aid, the programs were frequently priced beyond the capacity of

working men and women. They tended to appeal more effectively to those with already established educational appetites and to be eyed with suspicion by articulate workers attempting to institutionalize political and economic protest through trade unions, cooperative societies, and independent political parties.

Moreover, even though its importance has often been publicly affirmed, adult education has seldom in the past engaged the attention of legislatures. The most notable exceptions have been in the Scandinavian countries, in the eastern European countries under Communist regimes, and most recently in some of the newly emergent nations, such as Tanzania. In free enterprise economies and pluralistic societies, the interplay of philanthropy and social reformism and the promotion of particular religious, political, and social interests have all contributed to the creation of adult learning opportunities. But government action has largely resulted from the efforts of these independent pressure groups seeking official recognition and public financial support; thus specific legislative acts have usually not been concerned with a general theory or broad program of adult education so much as with the solution of particular problems brought forward by special interests. (Such a pragmatic or problem-oriented approach is illustrated by the setting up of agricultural experimental stations in the United States under the Hatch Act of 1887, which provided the basis for federal extension services in agriculture and home economics.)

The Scandinavian countries are unique in the numbers of legislators who claim personal experience of adult education, particularly in the folk high schools, and who are thus favourably disposed toward adult education. They are also the countries in which democratic practices are most deeply imbedded in a wealth of popular movements to which the state has come to entrust relatively large amounts of money for the development of adult education. One of the most recent examples is the Danish legislative act of June 1968 concerning "leisure time instruction." This act provides aid for youth schools, leisure-time activities for children and young people, leisure-time instruction for adults, supplementary cultural arrangements (that is, lectures, art exhibitions, concerts), university extension activities, correspondence schools, leisure-time instruction for seamen. Although the act specifies in detail the conditions that must be fulfilled by voluntary organizations and municipalities in order to qualify for state financial support, there is no attempt to define curricula or to restrict educational and ideological objectives. Similar "democratic" programs exist in all the Scandinavian countries.

In Great Britain, which continued in the early part of the 20th century to be a more distinctly hierarchic or class-structured society, adult education was frequently a case of the "privileged" accepting a responsibility for helping the "underprivileged." Some individuals with an elite education in independent schools and universities were willing to apply their education, example, and moral suasion to help the working classes to help themselves. Thus the Workers' Educational Association (WEA), founded in 1903 "to promote the higher education of working men and women," drew upon the services of tutors from the universities—in effect, broadening the activities of university extension. Emphasis on continuous academic study in small groups under the guidance of highly qualified tutors largely replaced the earlier tradition of lectures to large audiences. From almost the beginning the national board of education offered grants toward the teaching costs of such classes, and especially after World War II there was a rapid and substantial increase in the work of the WEA and of the university extension services based on increased grants from the new Ministry of Education (today entitled the Department of Education and Science). In addition there were established a number of new institutions concerned with adult education. Prominent among these were some 30 new residential colleges, some provided by universities, a small number by voluntary effort, but the majority by local education authorities. They all differed from Ruskin College and several others established earlier in that they provided only short courses, lasting from a weekend to a few weeks or occasionally two or three months. At the same time, local education authorities were increasingly fostering liberal studies in technical colleges and other institutions.

Traditions of voluntary participation and curricular freedom

Two traditions in adult education thus established themselves in the countries that pioneered adult education: the tradition of voluntary participation and the tradition of freedom from state control in matters of curricula and educational goals (even though the state might fund the programs). This contrasts with practices, for instance, in the U.S.S.R., which has put heavy emphasis on adult educational programs. According to an anonymous but typical text, *Education of Adults in the U.S.S.R.*, dated 1960, "The transformation of society into a Communist society now making headway in the U.S.S.R. involves the education of the wide masses of people, the development of their consciousness and the raising of their political and production activities." The provisions made to achieve these ends are impressive, but the key words here are *masses* and *society*—ignoring *individuals*. Adult education is promoted to enable people en masse to understand and adjust to change, not to enable them to exercise individual judgment about the character and direction of change.

Another kind of of contrast exists between the Scandinavian and Anglo-American countries on the one hand and France and other Latin nations on the other, where the Napoleonic aftermath of central direction of education is most marked. The public education authorities have taken little notice of adult education. As one English observer has noted:

> Those accustomed to the provision made for non-vocational adult education in the English-speaking countries have long found French practice in this sphere so foreign as to be incomprehensible. . . . In France, educational opportunities for adults are highly decentralised and informal, in the hands of a vast number of private societies, large and small, run by unpaid, part-time labour, whose records are often sketchily kept and whose finances are complicated. . . . The scene is further obscured because adult education is not considered as an independent branch of education as we tend to see it and the term itself is little used. For the French it forms only a part of a wider activity, called *éducation populaire* or *culture populaire*, which in its widest sense is defined as any communal non-vocational activity of a social, recreational, sporting or intellectual nature, carried on outside home, school, or work, which enriches and develops the personality. (From Colin J. Titmus, *Adult Education in France;* Pergamon Press Ltd., [1967].)

Irrespective of modes of provision and direction, however, adult education is also affected by the quantity and quality of education available in childhood and youth. Quite simply, the more education that people receive the more they are likely to want and the more likely they are to actively seek information about its provision. Raising the overall school-leaving age and increasing the proportions of young people proceeding directly to higher or continuative education are clearly associated with a rising demand for adult education for personal and social reasons. The same kind of increasing demand is associated with high economic productivity and the accelerated technological obsolescence that accompanies it. This sets in motion the need for vocational retraining in adult life that must be included in the modern concept of adult education.

AGENCIES AND INSTITUTIONS

Any classification of agencies and institutions involved in adult education must necessarily be arbitrary, given the great variety found not only among nations but within single nations. The following are the general types.

Varieties of schools and programs

The *folk high schools*, first established in Denmark and now found in all Scandinavian countries, are residential schools in which young adults who have completed formal schooling and usually have had some subsequent work experience pursue at least several months of study. The study aims at furthering both moral and intellectual development and instilling an understanding of local and

national traditions and conditions, thus still reflecting Grundtvig's original concept. Although at first they were independent or separate institutions, they are now frequently promoted or supported by communal boards of education. Although rarely exported with success in their pure form, the folk high schools have influenced the development of residential forms of adult education in countries as diverse as Canada, Kenya, India, and The Netherlands.

Nonresident adult education centres, the most widely distributed specialized institutes for adult education, are represented by "workers academies" in Finland, "palaces of culture" in the U.S.S.R. and some other Communist countries, "workers' universities" in Yugoslavia, "peoples' high schools" in West Germany and Austria, "adult education centres" in Great Britain, and "people's universities" in The Netherlands, Italy, and Switzerland. The distinguishing characteristics of these institutions are that they are independent of the general education authorities, at least in terms of programming; that student attendance is voluntary and part-time; and that teachers and administrators are either volunteers or professionals offering mainly part-time services. Traditionally these schools do not prepare students for examinations or offer training in advanced vocational skills. Typically the curriculum includes instruction in practical and domestic crafts, fine arts, music and drama, familial and social problem solving, and modern languages, as well as instruction designed to reinforce primary and secondary education. The workers' universities in Yugoslavia probably go furthest in trying to fuse these elements into an overall program to equip workers to play a role in the self-management of industrial and commercial enterprises.

Agricultural extension services, though almost wholly an American development, are conducted on a scale great enough to rate separate mention. The extension service of the U.S. Department of Agriculture conducts agricultural, home economics, and even public affairs programs in every county in the United States. It has had special significance in developing "demonstration" as a method of adult education, in emphasizing the adoption of new practices, and in providing a possibly appropriate model for newly developing countries.

The *open university*, a new British institution, is significant for its new dimension and sharp break with previous degree programs for adults. In some educationally advanced countries—such as Australia, New Zealand, Canada, and the United States—adults have long had opportunities to pursue part-time education leading to the award of university degrees, but these programs have usually been carbon copies of programs offered to regular undergraduates. The open university, in theory at least, aspires to a kind of universal higher education. It is intended to service only mature or older adults studying part-time; it has no standardized entry requirements; and it attempts to combine various educational technologies and techniques—correspondence instruction, mass-communication media, personal counselling, and short-term residential courses.

Commercial enterprises have provided correspondence courses or class instruction (part- or full-time) for adults who are usually seeking some form of vocational qualification (but who may, however, be simply seeking "self-improvement," as in speed-reading programs). Such schools may be licensed or supervised by state agencies (as in Sweden and The Netherlands), or they may be self-policing through associations offering accreditation. Some schools are nonprofit organizations. One interesting development has been Switzerland's *Klub-Schulen* ("club schools"), supported through the philanthropy of a commercial concern, the Migros trading organization, but operated on a nonprofit basis; the programs, furthermore, have no particular vocational emphasis.

Extension services include both public school programs for adults and the university extensions mentioned earlier. The school programs are administered by the public school systems, and they are popularly termed night schools because ordinarily they are housed in the same school buildings used in the daytime for school-age youth and also because some of the same teachers are often involved. (Much of the teaching, however, is also done by subject specialists not employed as schoolteachers.) Though often originating in efforts to remedy or supplement inadequate childhood education, many of these programs now cater to the same range of interests served by the "nonresident adult education centres" cited above. They often retain elements of vocational preparation at a less specialized level, generally for younger adults—for example, in commercial and trade skills.

The difficulties inherent in trying to provide adequate services to adults as a kind of marginal addition or afterthought to the education of children have led some local education authorities in Britain and the United States to design schools that can be used specifically for adults as well as children. The "village colleges" in Cambridgeshire and the Mott Foundation program in Flint, Michigan, involve a pattern of separate staffing for adult and child education and of comprehensive planning that enables a very large adult program to be provided within the school system. Such developments require the recruitment and training of larger cadres of full-time adult education teachers than have been customary in the past. The further growth and sophistication of public provision is likely to be a marked feature of adult education in the future.

The extension services offered by institutions of higher learning are of two broad types. The British tradition, influential in most Commonwealth countries and former colonial territories, has emphasized the provision of noncredit courses of "liberal" studies. The North American tradition, found in countries influenced by the United States and Canada, places a larger emphasis on credit programs duplicating courses offered to regular undergraduates; such programs are offered via television or correspondence or in separate "downtown" (urban) colleges. Both traditions seem in the process of modification—the British in the direction of offering more credit-earning and vocation-related refresher courses, the North American toward a wider acceptance of the provision of general liberal studies for the public at large and for specialized vocational groups. It is everywhere apparent that universities are assuming more responsibility for the continuance and renewal of education for the highly educated.

Organizations, libraries, and museums servicing adult activities

In addition to the various schools or services listed above, there are countless organizations whose main purposes may not be adult education but that offer some kind of instruction or leisure-time activities for adults. They include such bodies as the Young Men's Christian Association, the Young Women's Christian Association, political parties and labour unions (often operating through especially constituted "workers educational associations"), women's organizations like the Norwegian Housewives League, and temperance organizations like the International Order of Good Templars. Such a list could be almost endless. Other agencies for which adult education is a related rather than a primary function are libraries, museums, botanical gardens, and the like. Not only do these agencies provide the means of individual self-education but also they frequently promote group activities or put their accommodations and resources at the disposal of adult education agencies. Finally there should be mentioned the advisory and instructional services offered by various social and welfare agencies in the fields of health, safety, marital guidance, family planning, and so forth.

FUTURE PERSPECTIVES

Acceleration in the rate of change in human affairs is so much a cliché of contemporary discussion as not to require elaboration. It is summed up in the aphorism attributed to American anthropologist Margaret Mead that "no man today will die in the world in which he was born." The inescapable consequence is the need for continual renewal of learning in all forms of employment and in various phases of life. As one example of the problems involved, in those fields of advanced technology (such as aerospace and nuclear engineering) in which it

is impracticable to duplicate some operational plant in a separate educational milieu, education may have to be integrated with the productive process itself. Moreover, initial education in many fields is now so specialized and demanding on the time and energy of the young that a broad, general education may be impossible unless the whole of life, not merely the formative years, is accepted as time for learning.

Encouragement is given to these trends and to more widely ranging concepts of adult education by professional associations in adult education—associations whose objects are to promote cooperation between different schools and agencies, to foster enquiry and research, and to disseminate information. With differing structures and varying levels of development such organizations exist in Australia, Canada, Denmark, England and Wales, Finland, India, Ireland, The Netherlands, New Zealand, Norway, Scotland, Sweden, Switzerland, and the United States. Intercommunication beyond national boundaries has been encouraged by the United Nations Educational, Scientific and Cultural Organization (UNESCO) and by such regional bodies as the European Bureau of Adult Education, the African Adult Education Association, and the Asian-South Pacific Association for Adult Education. Still largely dependent on the personal efforts of a few individuals, these bodies are an index to the growing awareness of the importance of adult education throughout the world.

BIBLIOGRAPHY. Much of the literature of adult education is to be found in journal articles and in pamphlets, frequently issued in small editions and only accessible in specialized libraries, although sometimes now available on microfilm or microfiche. Some important general works, however, are: THOMAS KELLY (ed.), *A Select Bibliography of Adult Education in Great Britain*, 2nd ed. (1963), a bibliography of works published to the end of 1961; and *A History of Adult Education in Great Britain*, 2nd ed. (1970), on the development of both formal and informal means of adult education from the Middle Ages to the mid-20th century; J.F.C. HARRISON, *Learning and Living, 1790–1960: A Study in the History of the English Adult Education Movement* (1961), a text starting with the experience in England as the first industrial society and ending with the post-World War II search for social relevance; JOHN LOWE, *Adult Education in England and Wales: A Critical Survey* (1970), a recent general work on the English educational system; EDMUND DE SCHWEINITZ BRUNNER *et al.*, *An Overview of Adult Education Research* (1959), an inventory of research in nonvocational adult education made by the Bureau of Applied Social Research, Columbia University, that attempts to achieve valid generalizations applicable to the broad field of adult education; C. HARTLEY GRATTAN, *In Quest of Knowledge* (1955), a short sketch that concentrates on the western European background from primitive man to the Industrial Revolution and then focusses on the American experience, with a brief résumé of that in Britain; CYRIL O. HOULE, *The Inquiring Mind* (1961), a comprehensive study of the adult student; and J.R. KIDD, *How Adults Learn* (1959), a synthesis of theory and practice in the field of adult education; MALCOLM S. KNOWLES, *The Adult Education Movement in the United States* (1962), an informed, comprehensive study, containing a bibliography of general sources; COOLIE VERNER and A. BOOTH, *Adult Education* (1969), a well-documented basic text.

(E.M.H.)

Adventists

Adventism is a term that has come to refer to a belief in the Second Coming of Christ, at which time he will separate the saints from the wicked and inaugurate his millennial (1,000-year) Kingdom. Though usually associated with the Seventh-day Adventists, the largest organized modern Adventist denomination, Adventism is deeply rooted in Hebrew and Christian prophetism, messianism (belief in a coming salvatory figure), and millennial expectations recorded in the Bible. A desire that God will overcome evil, vindicate his people, and manifest his power to make all things new has influenced the history of Jewish and Christian life in a variety of ways.

History. *Roots of modern Adventism.* Adventism, in its broader sense, has been an important aspect of Jewish expectations about the future. Jews have looked for the coming of a messiah, and one of the ways in which they have reminded themselves of this hope is by leaving an empty seat at the table during the celebration of Passover, which commemorates their Exodus from Egypt in the 13th century BC, in the hope of an awaited return of Elijah, a 9th-century BC prophet, who is believed to have been taken up into heaven. At various times, persons have appeared claiming to be the promised Messiah. Shabbetai Tzevi (1626–76), for example, made such a claim and announced to his followers that 1648 was the year of redemption. Jews in the Diaspora (those living beyond Palestine), harbouring messianic hopes, have longed for their return to Israel and the restoration of the Temple in Jerusalem. Christians maintain that Jesus is the Messiah, and in the early church they anticipated that he would return again soon after his death, interpreting these expectations in terms of prophecies in the Book of Daniel and in the book of Revelation. Because Jews do not accept Jesus as the Messiah, they may be said—in contrast to Christians—to be awaiting a first advent. In Asia Minor the prophet Montanus (mid-2nd century), assisted by Priscilla and Maximilla, two prophetesses given to ecstatic experiences, stimulated an interest in Adventism that lasted until the 9th century. This emphasis passed into two heretical sects: the Priscillianists, followers of Priscillian, a 4th-century Spanish mystic, and the medieval Cathari (Purifiers). During and after the Reformation, Adventism emerged among the radical reformers known as Anabaptists (*c.* 1534), the Fifth Monarchy Men in England (those who in about 1640 awaited the fifth kingdom; *i.e.*, of Christ), and the German Pietists (Lutheran) in the 17th and 18th centuries. There has always been a concern among Christians about the meaning of eschatology (the last times) and the Second Coming of Christ, and many have written about this subject: Augustine in *De civitate Dei* (*The City of God*); Isaac Newton in his *Observations upon the Prophecies of Daniel and the Apocalypse of St. John* (1733); and Jonathan Edwards in *A History of the Work of Redemption* (preached 1739, published 1774). Adventism in the 19th century drew upon the chronology of a book by Archbishop James Ussher (1581–1656) of Ireland, *Annales Veteris Testamenti . . .* (1650) . . . *Novi Testamenti . . .* (1654), which placed the creation of the world and of man in 4004 BC, the date by which millennialists calculated the Second Coming of Christ (see also MILLENNIALISM).

Expectations of the Messiah

Seventh-day Adventists. During the latter part of the 18th and the early part of the 19th centuries, great anticipation arose among Christians concerning the coming of Christ's millennial Kingdom. Millennialists searched the Bible, assured themselves about what prophecies had been fulfilled, and then figured out those yet to be realized. The Kingdom, they believed, was near at hand. Some theologians saw in the American and French revolutions the beginning of the downfall of papal power, which Protestants equated with the beast of Revelation, and the inauguration of a new age of justice, in which the whole world would be gradually converted to a belief in Christ.

Such Christians have been known as postmillennialists in that they expected Christ's return after 1,000 years. This expectancy was manifested in communitarian experiments of the Shakers and the Mormons, who believed that in Ann Lee (1736–84), the founder of the Shakers, and Joseph Smith (1805–44), the founder of the Mormons, God had manifested himself in special ways (see also MORMONISM).

It was in this atmosphere that William Miller (1782–1849), a founder of the Adventists, began to preach. Miller, while an officer in the U.S. Army in the War of 1812, had become a skeptic. After a conversion, he began to study the books of Daniel and Revelation and to preach as a Baptist, the faith of his father. He concluded that Christ would come on March 21, 1844, and was encouraged in his views by a number of clergymen (such as the reform-minded Joshua Himes of Boston) and numerous followers—about 100,000 according to some estimates. Miller was accused of fanaticism because of his millenarianism, and he stressed the coming of Christ in conjunction with a fiery conflagration, which distressed

The prophecies of William Miller

his hearers. Christ, however, did not return on the first appointed date, so Miller set a second date, October 22, 1844. The quiet passing of this day led to what is called the "Great Disappointment" among Adventists and the convening of a Mutual Conference of Adventists in 1845 to sort out problems. Those who met, however, found it difficult to shape a confession and form a permanent organization.

Among those who persisted after the failure of Miller's prophecy were Joseph Bates (1792–1872), James White (1821–81), and his wife, Ellen Harmon White (1827–1915). These Adventists, called Millerites in the press, continued to believe that Miller had set the right date, but they believed that they had interpreted what had happened in the wrong way. Reading Daniel, chapters 8 and 9, they concluded that God had begun the "cleansing of the heavenly sanctuary,"—*i.e.*, an investigative judgment that would be followed by the pronouncing and then the execution of the sentence of judgment. What actually began in 1844, then, in their view, was an examination of all of the names in the Book of Life. Only after this was completed would Christ appear and begin his millennial reign. Though they did not set a new date, they insisted that Christ's advent was "personal, visible, audible, bodily, glorious and premillennial" and imminent. They also believed that observance of the seventh day, rather than Sunday, would help to bring about the Second Coming. These Millerites founded a magazine in 1850—based on the earlier *Present Truth*—called the *Advent Review and Sabbath Herald* (now the *Review and Herald*), established a headquarters in Battle Creek, Michigan, in 1855, and became an official denomination (Seventh-day Adventists) in 1863. Adventists believe that Mrs. White, a former Methodist, had received the gift of prophecy. She had visions, travelled and lectured widely in America, Europe, and Australia, and wrote prolifically, including the "Conflict of the Ages" series (5 vol., 1888–1917) and *Testimonies for the Church* (9 vol., 1855–1909).

Other Adventist churches and movements. Other Adventist bodies emerged in the 19th century as a direct or indirect result of the prophecy of William Miller. These include the Evangelical Adventists (1845), Life and Advent Union (1862), Church of God (Seventh Day, 1866), Church of God General Conference (Abrahamic Faith, 1888), and the Advent Christian Church. The latter is the largest among these small bodies. It was organized in 1860 by Jonathan Cummings (fl. *c.* 1852), a follower of Miller. Cummings stressed the teaching of conditional immortality. Because of sin, man forfeited immortality. At death he remains in a state of unconsciousness until Christ's Second Coming, when the faithful will enter into life and the unrighteous will suffer extinction. The Advent Christians do not accept the teachings of Mrs. White as in any way prophetic. They also reject the teachings of the Seventh-day Adventists about sabbath observance and dietary laws. They are congregational in polity and coordinate work in the United States and throughout the world through the Advent Christian General Conference of America. In 1964 the Advent Christian Church united with the Life and Advent Union.

Adventist emphases in various religious groups

Adventist emphases may also be found among pentecostal and holiness groups that interpret the outpourings of the Holy Spirit, the manifestation of his gifts, and growth in sanctification as signs of the coming of Christ and his Kingdom. Charles Taze Russell (1852–1916), the founder of the Jehovah's Witnesses, came under the influence of Adventist ideas. His followers became known as Millennial Dawnists, who believed that this present world is under the reign of Satan and in conflict with Jehovah's (God's) rule. Millennial Dawnists considered it their responsibility to announce the Second Coming of Christ, which would be marked by the battle of Armageddon, in which Satan would be defeated, the faithful rewarded, and the wicked annihilated (see also JEHOVAH'S WITNESSES).

Beliefs, practices, and institutions. *Beliefs and practices.* Seventh-day Adventists share many basic beliefs held by most Christians. They accept the authority of the Old and New Testaments. They are Arminian (emphasizing man's choice and God's election) rather than Calvinist (emphasizing God's sovereignty) in their interpretation of Christ's atonement, and they argue that his death was "*provisionally* and *potentially* for all men," yet efficacious only for those who avail themselves of his benefits. In addition to the emphasis upon the Second Advent of Christ, two other matters set them apart from other Christians. (1) They observe the seventh day, rather than the first day of the week, as the sabbath. This day, according to the Bible, was instituted by God since the creation, and the commandment concerning sabbath rest is a part of God's eternal law. The keeping of this law by Adventists is both a testimony to God's sovereign freedom to set aside such a day and an eschatological testing for them. (2) They also avoid eating meat and taking all narcotics and stimulants, which they consider to be harmful. Though they appeal to the Bible for the justification of these dietary practices, they maintain that these are based upon the broad theological consideration that the body is the temple of the Holy Spirit and should be protected. To criticism that they have made Mrs. White's writings canonical, Adventists respond that all of their teachings are based upon the Bible; and to criticism that they have turned sabbath-keeping and dietary practices into a new works righteousness (*i.e.*, salvation through good works), they respond that these works are a result of salvation, not the means.

Adventists stress tithing and therefore have a high annual giving per capita that allows them to carry on worldwide welfare programs. Because of their educational program, Adventists have created an educational elite that—although there have been some tensions—has remained in the denomination. Adventists have never found it easy to keep Saturday as the sabbath in a society that, by law or by custom, has set aside Sunday for a day of rest. Because of their unwillingness to work on Saturday, they have suffered some job discrimination. This has made them deeply concerned about religious liberty. In recent years, despite some of the teaching against ecumenism in the writings of Mrs. White, consultations have been held between representatives of the Seventh-day Adventists and the World Council of Churches to discuss the relation of Adventists to the ecumenical movement.

Organization of the Seventh-day Adventists

Institutions. Sending out its first missionary, John Nevins Andrews (1829–83), in 1874, Seventh-day Adventism expanded into a worldwide movement. The General Conference, the main governing body, has its headquarters in Takoma Park, Washington, D.C., where it was moved in 1903 from Battle Creek. The meeting of the General Conference is held quadrennially and is made up of delegates from divisions, union and local, that are also called conferences. These local conferences provide pastoral oversight for the local congregations, which are governed by elected lay elders and deacons. The business of conferences between meetings is administered by a president and an executive committee. As of 1968 the General Conference governs some 76 union conferences, 374 local conferences, and 15,744 churches involving 1,845,183 persons all over the world. In the United States (1969), Adventists number 407,766 in 3,216 congregations. The conference also supervises the work of evangelism in over 1,000 languages, a large parochial school system, and numerous medical institutions called sanitariums, of which the first was established in Battle Creek in 1866. Publishing houses are operated in many countries, and literature is distributed door to door by volunteers.

Conclusion. The establishment of the State of Israel (1948) has stirred an interesting discussion among Adventists, including those who are members of Adventist denominations and those from other denominations who share with them Adventist beliefs about the Second Coming. Jews have interpreted the restoration of the State of Israel as an end to the Diaspora and as a fulfillment of the promises of God to give them a homeland. Some Fundamentalist Christians have tended to interpret this event as a fulfillment of the promises of God about the Jews and as a sign of the imminent coming of Christ. The Jerusalem Conference on Biblical Prophecy was held in

Jerusalem in 1971, under the leadership of Carl F.H. Henry (1913–), an American conservative religious leader, to study the implications of this event for Christians and Jews.

Though all of the longings and yearnings of Jews and Christians for the advent and the Second Advent of the Messiah, and for the establishment of his Kingdom, seem to remain unfulfilled, in Adventism there is always the element of hope.

BIBLIOGRAPHY. SYLVESTER BLISS, *Memoirs of William Miller* (1853); LEROY EDWIN FROOM, *The Prophetic Faith of Our Fathers*, 4 vol. (1950–54), a very rich description of adventist literature by an Adventist, helpful for research; JESSE O. GIBSON (ed.), *Seventh-Day Adventist Yearbook* (1970); DON F. NEUFELD *et al.* (eds.), *Seventh-Day Adventist Encyclopedia* (1966); FRANCIS D. NICHOL *et al.* (eds.), *Seventh-Day Adventist Bible Commentary*, 7 vol. (1953–57); ERNEST SANDEEN, *The Roots of Fundamentalism: British and American Millenarianism, 1800–1930* (1970); GERSHOM SCHOLEM, *The Messianic Idea in Judaism* (1971), excellent for advent ideas among Jews, especially for the treatment of Sabbatians, and for references; "The Seventh-Day Adventist Church," in *The Ecumenical Review*, 19:17–28 (1967), considered by Adventists as a very fair treatment of the subject; ELLEN HARMON WHITE, *The Great Controversy Between Christ and Satan* (1888 and 1911); ERNEST LEE TUVESON, *Redeemer Nation* (1968).

(J.H.Sm.)

Advertising

Advertising is a form of communication intended to promote the sale of a product or service, to influence public opinion, to gain political support, to advance a particular cause, or to elicit some other response desired by the advertiser. The advertising message, or advertisement, is delivered to its intended audience through the various media including newspapers, magazines, television, radio, billboards, and direct mail. Advertising is distinguished from other forms of communication in that the advertiser pays the medium to deliver the message. For this payment the advertiser receives the opportunity to control the message. Within legal constraints designed to prevent deception and assure fair competition, and within standards of practice enforced by the media, the advertiser is free to say what he wants to say the way he wants to say it. Within limits set by availability, he can select the particular issue of a newspaper or magazine in which the message will run; in the case of television and radio, he can select the hour and even the minute when the message will be broadcast. The style of presentation usually identifies the message as an advertisement. In some countries if the message is not readily distinguishable from a medium's editorial or program content, the notation "This is an advertisement" and the name of the advertiser or sponsoring organization are included.

Users of advertising

Different kinds of businesses use advertising to motivate different kinds of markets toward different kinds of responses. A retailer such as a department store advertises to build store traffic and patronage and to keep his regular customers informed about available merchandise. The manufacturer of a branded item such as a brand of soap advertises to build a brand preference—that is, to motivate a large number of consumers to buy his brand regularly at any of the many stores where it is carried. The manufacturer of industrial goods, such as steel, machinery, lubricants, and office equipment, advertises to other business firms to get them to use his product in their operations. The producer of ethical drugs, building materials, or textbooks advertises to doctors, architects, or educators to induce these professional people to prescribe, specify, or recommend his product to patients, clients, or students. An association of firms in a particular industry, such as sugar refiners, orange growers, or dairies advertise to stimulate consumption of a generic product such as sugar, orange juice, or milk. All business firms, whether they be retailers, manufacturers, or service institutions such as banks, insurance companies, and airlines, advertise to build a respected corporate identity, to make their corporate name well known and highly regarded.

Advertising is also used to awaken, enlighten, and activate the public at large concerning matters that affect society generally. Private, nonprofit organizations in some countries create and distribute various public service advertisements including those for forest fire prevention, rehabilitation of the handicapped, traffic safety, drug abuse information, crime prevention, and continuing education.

Volume of advertising

Of the 64 non-Communist countries surveyed by the International Advertising Association, the United States led in advertising expenditures with more than $18,000,000,000 in 1968, which was 61 percent of the world's total expenditure. West Germany was second with expenditures close to $4,000,000,000 (16 billion deutsche marks); Japan was third with $1,500,000,000 (540 billion yen); Great Britain was fourth with $1,200,000,000 (451 million pounds). In France and Canada expenditures were approaching the level of $1,000,000,000. As a percentage of gross national product, on the other hand, advertising expenditures in 1968 were highest in West Germany at 2.87%. Next was Switzerland at 2.42%, then Jamaica at 2.11%, and the United States at 2.06%.

Per capita advertising expenditures in the United States in 1968 were placed at $89.20. In three other countries, Switzerland, West Germany, and Sweden, expenditures were over $50 per inhabitant. In several countries advertising expenditures were less than 20 cents per capita (Nepal 5 cents, Pakistan 14 cents, Ceylon [now Sri Lanka] 16 cents, India 16 cents). This wide range of per capita expenditures demonstrates that advertising is more prevalent in highly industrialized countries where consumer purchasing power is broadly distributed, where the economic system is competitive, and where mass communications media are more generally available. (Although consumer advertising exists in some Communist countries, its volume is relatively insignificant.)

ESTABLISHMENT OF MODERN ADVERTISING

The first advertising was by public criers in ancient times who circulated through the streets calling attention to the sale of such items as slaves, cattle, and imports. An ancient written advertisement, perhaps 3,000 years old, was discovered by an archaeologist delving in the ruins of Thebes. It offered a "whole gold coin" as reward for the return of a runaway slave named Shem. During the Middle Ages the spoken word in Europe was the principal means of communication, and there was little of the activity that is now called advertising. The invention of movable type about 1450 by Johannes Gutenberg made it possible to produce many copies of books and periodicals cheaply and quickly and ushered in the modern era of advertising.

Printing made possible the transition from simple announcement to the system of argument and suggestion that constitutes modern advertising, and the medium of this development was the newspaper. The weekly papers sometimes carried a few advertisements, including the first offerings of coffee (1652), chocolate (1657), and tea (1658) in England. In June 1666 the *London Gazette*, No. 62, announced the first advertisement supplement as follows: An Advertisement—Being daily prest to the publication of Books, Medicines, and other things not properly the business of a Paper of Intelligence. This is to notifie once for all, that we will not charge the *Gazette* with Advertisements, unless they be matter of state; but that a Paper of Advertisements will be forthwith printed apart, and recommended to the Publick by another hand.

Advertising was prospering by the following century, and in 1758 Samuel Johnson could write in "The Idler": "Advertisements are now so numerous that they are very negligently perused, and it is therefore become necessary to gain attention by magnificence of promise and by eloquence sometimes sublime and sometimes pathetick."

The 19th century was a period of expansion in advertising as well as in business generally. The Industrial Revolution expanded the output of factories, and advertising helped market this output. The growth of the penny press and of magazines, combined with an increase in the educational level in the Western nations, brought forth an in-

crease in audiences. Improvements in transportation made it possible to distribute magazines geographically on a wide scale, thus permitting vast increases in circulation and reductions in price. Slogans and jingles became favourite advertising techniques. Some of the trademarks famous today were popularized during this period.

During the first two decades of the 20th century, advertising was re-examined. Certain excesses in its claims had developed and caused widespread resentment. Regulation came both from within industries and from governments.

After World War I came the era of salesmanship; advertising became "salesmanship in print." Advertising was accepted as an essential tool in selling the booming output of all the nations' factories. "It pays to advertise" became a standard slogan. Soap companies and automobile companies were among the most important advertisers. It was discovered that testimonials by movie stars were effective in selling such items as perfumes and cigarettes.

A major new advertising medium, radio, was added during the 1920s. At first, most radio commercials were indirect and there was doubt as to whether advertising would be allowed on the air. In such countries as Great Britain and France, it was banned from public-owned radio; in the United States, however, commercial radio thrived, and by 1928 it accounted for $10,500,000 in advertising.

Another new medium, television, was added to the advertising scene immediately after World War II. It had a spectacular growth in the U.S.; it became the second largest advertising medium (after newspapers) in terms of total dollars spent and the largest in the top national advertisers' budgets. In the postwar era, too, American advertising agencies began popularizing the concept of the "brand image." According to one agency president, "Every advertisement should be thought of as a contribution to the complex symbol which is the brand image. . . . Every advertisement is not a one-time shot, but a long-term investment in the total personality of the brand." The concept of a brand image was extended to building favourable images for corporations, retail stores, service institutions, and political candidates.

ADVERTISING MEDIA

Advertising media can be classified as follows: newspapers, magazines (consumer magazines, business publications, farm publications, professional journals), television, radio, direct mail, outdoor media (signs, posters, painted bulletins, electric displays), transit media (car cards, outside displays, station posters), and miscellaneous (dealer displays, theatre-screen advertising, specialties, directories).

There is no single best medium for all advertising situations. Each has its own character, and each advertising situation presents a unique set of circumstances. A medium that is best for one advertiser may not be best for another, and each medium must therefore be considered in terms of how well it meets the individual advertiser's requirements.

Newspapers

Newspapers because of their big circulations are the most widely used advertising medium. They offer something for everybody: news of what is going on in the world as well as in the local community, editorial comment on current issues, sports coverage, financial pages, homemaker sections, and entertainment features. With often 60 to 70 percent of their space devoted to advertising newspapers keep their readers informed of the myriad of goods and services available in the marketplace, including the apartment for rent, the house for sale, and the employment opportunity appearing in the classified section.

The newspaper is an ideal medium for most retailers. Its circulation area generally conforms to the retail trading area. Retailers located in suburban shopping centres of major cities usually have access to a local community newspaper, and many major metropolitan papers offer zoned editions, whereby the circulation is split into geographic areas or zones, making it possible for a retailer to buy only that part of the circulation going into his trading area. A newspaper's daily publication readily fits the day-to-day selling and quick action promotions that characterize most retailing. Consumers regularly turn to their newspapers for shopping information including pictures of the merchandise, product specifications, and prices. The wide range of space sizes and publication dates gives small as well as large advertisers an opportunity to advertise profitably.

Big advertisers can achieve wide coverage by using many newspapers, and they use newspapers for a special purpose: to supplement other media such as television and magazines that provide inadequate coverage in particular cities, to concentrate effort where there are exceptionally promising sales opportunities, to gain the support of local dealers, to counter strong local competition, or to test the introduction of a new product.

Magazines

Magazines offer a wide range of advertising opportunities. General magazines such as *Reader's Digest* (U.S.), *Paris-Match* (France), *Der Spiegel* (West Germany), and *Asahi* (Japan) are edited to serve the interests of many different population segments, thereby offering advertisers the opportunity to reach a broad cross section of people in all parts of a nation. Women's magazines such as *Elle* (France), *Good Housekeeping* (U.S.), and *Vogue* (both U.S. and France) appeal to the various special interests of women, thereby offering advertisers the opportunity to reach these important consumers in a variety of editorial environments. For men there are magazines concentrating on sports, or on business or science. There is a magazine for just about every interest group imaginable. Each type of magazine offers the advertiser an opportunity to reach his best prospects when they are reading subject matter relevant to the advertiser's product. A manufacturer of milking machines, for example, can reach dairymen in a receptive frame of mind when he advertises in a dairyman's journal. The greatest number of magazines exist in the fields of business, in which there are one or more publications for every industry and occupational grouping. Advertising in business magazines paves the way for a company's salesmen and at various levels reaches management people who might influence a buying decision. Finally, there are the many professional journals that are read by doctors, dentists, engineers, architects, and educators who are in a position to prescribe, recommend, or specify what others will buy.

The geographic selectivity of consumer magazines has been increased through regional editions that enable advertisers to place an advertisement in a part of the circulation that goes into a particular market area. The advertiser must buy a certain minimum circulation, but this is likely to be sufficiently low to permit him to advertise in only one of the major cities. Regional editions let local and regional advertisers benefit from the prestige of a national publication.

The ideal measure of a newspaper's or magazine's value to a given advertiser would be the number of his prospective customers who read it. It would be extremely difficult, however, to isolate and count prospects for each and every advertiser. Therefore, research organizations that measure advertising for advertisers report the composition of magazine audiences in rather standard demographic terms such as age, sex, occupation, income, and family size. Using these breakdowns the advertiser selects newspapers and magazines whose audiences best match the demographic characteristics of his market.

Television

In the affluent nations today, more people spend more time with television than any other medium. In the United States, for example, during the peak viewing hours of an average winter evening over 50 percent of the households are tuned in. Viewing per household averages six hours daily. Viewing declines about 20 percent during the summer months. Changes in the hour-by-hour activities of various members of the family influence the composition and size of the television audience. Men do most of their viewing at night and on weekends. Preschool age children watch more daytime television than

school-age children. Women usually do more viewing than men in all time categories. The older age groups, 50 and over, do more viewing than younger groups. The audience composition varies by program types. Adults over 50 prefer western, drama, and variety shows. Adults under 50 prefer feature films, whereas nonadults watch more situation comedies.

Television advertisers buy units of time: a minute, thirty seconds, twenty seconds, ten seconds, a quarter hour, a half hour, an hour or longer. In such countries as West Germany all commercials are placed between programs; there are no program interruptions. In the United States they are placed both between programs and within programs. The advertiser may specify the time his commercial is to be broadcast or for a lower rate he may leave its placement to the station's discretion. Program time units may be bought for exclusive sponsorship or multiple sponsorship. Through exclusive sponsorship a single advertiser gains the full benefit of being identified as the program's sponsor, but he also must bear the full cost. Multiple sponsorship gives two or more advertisers the opportunity to be identified with a particular program while sharing the cost.

Rating systems for television

The United States employs by far the greatest variety of methods for measuring television audiences. The "Audimeter," for instance, developed by the A.C. Nielsen Company, is an electronic device that is connected to all television receiving sets in a representative sample of households. The Audimeter makes a minute-by-minute record on photographic film of the time periods when each set is turned on, the specific stations to which the set is tuned, the dialing from one station to another, and the length of time spent with each station. The film cartridges are sent in every two weeks to the Nielsen organization where the data are tabulated and matched against the program schedules of the television stations involved. Although the Audimeter provides a precise record of set tuning, it does not show how many people were viewing nor which members of the family were doing so. The "coincidental telephone" method consists of telephoning a sample of households and asking if a television set is turned on and if so, to what station or program. This method is relatively inexpensive and therefore is used extensively in local market areas. The "roster-recall" method involves personal interviews in which a roster of all stations and programs readily available in the area are presented to a sample of people. The respondents are asked to check those programs or stations they recall having watched. This method is useful for surveying homes without telephones and various ethnic groups. The "diary" method requires that individual members of the family use a prepared diary form and keep a record of their viewing as it occurs. This method makes it possible to estimate the composition as well as the size of the audience.

Commonly used measurements of television audiences are: "program rating," which reveals the percentage of all households in the sample turned to the particular program; "share-of-audience," which gives the percentage of households using television tuned to each program; "projected-audience," which gives the estimated number of households reached by a program (a projection of the program rating to total population); and "audience-composition," which offers a breakdown by sex, age, family size, income, and so forth.

Radio is the most ubiquitous medium; but in most countries the stations are government owned, and advertising is absent. In the United States radio-advertising time is sold much the same as television time. Also, the same methods of audience measurement are used. Radio audiences, however, are not measured as frequently or as extensively with regular surveys. American radio is also primarily a local medium; retailers account for about two-thirds of all radio advertising expenditures.

Direct mail

By mail the advertiser can send his message directly to individual prospects. The message can take one of many forms: letter, folder, booklet, brochure, post card, catalog, or coupons. In letter form the message can be personalized by addressing the recipient by name, by appealing to his personal concerns, and by using the language and style of a personal letter. The advertiser also can single out special groups for special appeals. He can be fairly certain that the message will be at least partly read by most recipients and that it will be read on the date it is received.

The mailing list is a critical factor in direct-mail advertising. Frequently used sources of names and addresses are telephone and city directories, trade directories, professional directories, commercial rating books, tax lists, club membership lists and the records of local governmental bodies. Some firms make a business of compiling lists and selling them to advertisers.

Organized outdoor media include painted signs, electric displays, and printed posters, the latter being the most important. The standard poster or billboard is located mainly on heavily travelled routes in and around urban markets. The fleeting exposure to motorists and passers-by demands a brief message, one that can be seen or read in a glance. A single, simple, visual idea with a caption usually is most appropriate. This brevity of message, however, limits the use of posters to reminder advertising.

Other media

Transit advertising including car cards, outside displays, and station posters reaches the millions of riders of public transportation in urban areas. Like outdoor advertising, transit advertising offers frequent exposure to large numbers of people at low cost. Its limitations are brevity of message and an environment low in prestige.

Dealer displays reach consumers at the point of purchase and therefore serve as a reminder or as an invitation to immediate buying action. Display space, however, in crowded, self-service supermarkets is at a minimum.

Theatre-screen advertising offers the opportunity to reach a captive audience with the impact of sound and motion pictures on a big screen.

Advertising "specialties" include calendars and other useful items that are imprinted with the advertiser's name. They include ash trays, lighters, thermometers, key chains, ballpoint pens, memo pads, paperweights, and rulers. Being items that are used frequently, they serve as frequent reminders and engender some goodwill.

ADVERTISING STRATEGY

Advertising strategy is based on a series of answers to these key questions: (1) To whom should the advertising be addressed? (2) What communication objectives should be sought? (3) What should the message contain? (4) What media should be employed? (5) How should expenditures be budgeted? Advertising often is characterized as a means of selling to the masses. This leads to the misconception that advertising should be addressed to everyone. Any population, however, is composed of different individuals with different needs and wants, different lifestyles, and different tastes. The most efficient strategy is thus to single out those segments of the population who are the best prospects and address the advertising to them. Communication objectives include the central idea to be learned, the image to be perceived, and the action to be taken by the target audience. Strategic decisions on message content concern the visual, verbal, and aural devices, the tone or style, and the technique of presentation to be used. Media selection includes the classes of media to be used—for example, daytime television, billboards, or women's magazines. Budgeting involves deciding how much to spend and how expenditures are to be allocated to markets, to media, and to dates. Thus, advertising is essentially a decision-making process. To aid the decision maker, research is used in a variety of ways.

Research

Researching the consumer gives the advertiser a better understanding of the people whom he wants to influence. Motivations that underlie their buying behaviour, their attitudes toward particular product categories and brands, their perceptions of competitive brands, their satisfactions and dissatisfactions with the products they use, their reactions to new product concepts: these are

some of the kinds of information that help the advertiser to identify his best prospects and to select the strongest appeal, thereby increasing the effectiveness of his advertising. Complicating the process is the fact that motivation is multiple. There is seldom one reason for any response. Also, there is no universally strongest motive. The strength of a particular motive depends on the value systems of the particular persons and particular cultures. An effective strategy is therefore to use a different appeal for different market segments.

Motivation research utilizes depth interviewing and projective techniques such as word associations, sentence completions, and various other tests drawn from clinical psychiatry. Use of these techniques to educe subconscious motives led to the fear in the 1950s that advertisers were gaining the power to manipulate consumer behaviour. The fear was not well founded, for it overlooked the power of the individual to resist persuasion against his own perceived interest. To the degree that techniques of persuasion have become more sophisticated and effective, people have developed resistance to manipulation.

Market research seeks to determine the sales potential of various segments and areas. For many product categories a small proportion of the population consumes a large proportion of the product. Twenty percent of total households in the U.S., for example, buy 99 percent of the prepared dog food. Once the heavy user group is isolated it can be further described. Traditionally, market segments have been described in demographic terms such as sex, age, income, and occupation. Recently there have been more attempts to use psychological measurements such as personality traits and motivational variables. Geographically, the market for most products is concentrated in urban areas, where most of industrial populations now live. Regional differences in climate, ethnic groups, and life-style produce different sales potentials for different products. Because markets are dynamic, however, continuing research is necessary to keep up with the changes.

A company's advertisement is best viewed as one of a series building a consistent identity for the product or the company. Creative strategy, therefore, should produce a central idea or unifying theme for a continuing campaign. The theme may be expressed verbally in the form of a brief proposition, a headline, or slogan. It may also be expressed visually in pictorial content, layout style, and graphic treatment. A copywriter for a television commercial, for instance, develops the message idea and draws on the myriad of elements of photocinematography—the visual, aural, and optical elements. He functions much as a motion-picture scriptwriter does. He develops the plot; visualizes the scenes; plans the action; writes instructions to the producers, performers, and technicians. Often a writer and an artist work together in the creative stages. A story board—an illustrated script or diagram of the commercial—is used to present ideas, to gain approval prior to incurring heavy costs, and to facilitate production. Every advertisement thus exploits the unique communication opportunities offered by the medium.

Media strategy

Creative requirements have important implications in media selection. If superior full-colour reproduction in print is required, magazines have an advantage over newspapers. New product introductions benefit from the "news" format of newspapers, radio, and television. If the copy approach seeks to inspire confidence or dispel doubt, an authoritative medium such as a professional journal is appropriate. The environment in which the message appears has an appreciable influence on the effectiveness of the message itself.

The basic variables of media strategy are "reach" and "frequency." Reach refers to the number of households or individuals exposed to the advertising message in a given period of time. Frequency is the average number of times that they are exposed. These variables are interdependent. With a given budget it is possible to reach more people less often or fewer people more often. The optimum combination depends on the requirements of the particular advertising situation. When introducing a new product whose market is unknown, for example, one may try to reach as many people as the budget permits. When the market is well defined, however, one may try to "expose" the message as frequently as possible. If the marketing strategy requires getting new users, reach is more important. If marketing strategy requires concentrating on particular persons and sustaining brand loyalty, frequency is more important.

The scheduling of advertising, the seasons, months, days, and hours when the ads will run, is another strategic decision. Advertising often should run immediately prior to, or in the early part of, seasonal buying or gift-buying periods. The day of the week and the hour of the day can be selected to reach certain groups when they are available and most receptive.

The contribution of advertising to sales and profits cannot be readily isolated from all the other variables in the marketing mix. Determining the size of the advertising budget is, therefore, a rather arbitrary procedure. Assuming that a relation exists between advertising and sales (even though it cannot be precisely measured), a common practice is to express the advertising budget as a percentage of sales. The percent is high for some products—about 20 percent for cosmetics—and low for others—about one percent for insurance. The procedure involves forecasting sales and establishing a percent of sales for advertising. The size of the advertising budget tends to be greater when most of the selling burden is placed on advertising instead of personal selling, when a new product is introduced, when there is little or no discernible difference among competing brands, or when competitive practice requires heavy advertising.

Testing effectiveness

Various techniques have been developed to test advertising ideas prior to publication or broadcast. Opinion and attitude tests involve showing alternative advertisements to a sample of persons, or "prospects," and measuring their expressed preferences and scaled attitudes (degree of liking or disliking). Recall tests use memory as a criterion of effectiveness: prospects are exposed to the advertisements in a folder or in a simulated publication or in a laboratory setting where the length of exposure can be controlled; then they are asked to recall whatever they remember—the best remembered being presumably the most effective. All these tests involve some degree of artificiality. Whenever people are forced to examine advertisements in an interview situation or in a laboratory setting and whenever they are forced to respond to questions bearing on some aspect of communication, the conditions are different from those under which advertising works in real life. Results, therefore, should be treated as tentative findings rather than precise measurements.

To approximate reality, controlled experiments are conducted in the field. The advertising actually is placed in the media so that it reaches people during their normal reading, viewing, or listening. Exposure is voluntary. Responses are voluntary and objectively observed. One form of experiment is the inquiry test. Using a split run, two or more advertisements are placed in the same position in the same issue of the same publication so that matched samples of readers will be exposed to each advertisement. Their responses in the form of an inquiry or a mailed coupon to a coded address indicate which of the advertisements is most effective. For mail-order selling this is an ideal test. For other kinds of selling, the inquiry merely indicates that the advertisement was seen and succeeded in motivating a certain response. The most sophisticated experimental design is one that runs the advertising in selected local test markets and observes ensuing fluctuations in sales. For control purposes similar markets are selected in which no advertising is run in order to take account of sales fluctuations due to all other causes. The difference between test and control markets should indicate the contribution of advertising. Various multivariable experimental designs test several variables simultaneously—for example, various combinations of media.

An advertising agency, Reynell and Son, was founded in London as early as 1812, but advertising agencies have

Advertising agencies

achieved their greatest development in the United States. In 1841 Volney Palmer, who had been an editorial writer and later a member of the advertising staff of the *Pottsville* (Pa.) *Miner's Journal*, opened in Philadelphia what is generally considered to have been the first advertising agency in the U.S. Unlike modern agents, he was primarily an agent for the publishers, who paid him a commission of 25 percent for selling space, although he did not assume liability for the credit extended to advertisers. Shortly after he began operations he had several competitors.

George P. Rowell, who opened an agency in 1865, added two features—wholesaling and the "list system." He contracted with a large number of weekly newspapers to sell him a column of space for a certain period of time; he then offered this space to advertisers in smaller units at retail rates that were below rates charged by competitors. He was, like Palmer, primarily a seller of space, but he offered exclusive lists of newspapers and magazines, and any advertiser who bought space in these had to buy through the exclusive agent.

In 1875 N.W. Ayer & Son of Philadelphia developed marketing policies that gave advertisers access to the true rates charged by the publishers. Ayer thus put the emphasis on the agency's service to the advertiser rather than on selling space for the media.

The next major step was the development of a full line of services for the advertiser. In the late 19th and early 20th century the emphasis was on the creation of advertising copy. After copy came art services, advertising production, and the selection of media. By 1920 most agencies could plan complete campaigns for their clients, including research and budgeting, and could also execute these campaigns by preparing the advertisements and arranging for their insertion in the various media, which now included outdoor advertising and direct-mail. By 1970 there were more than 4,000 bona fide advertising agencies in the United States—as contrasted, for instance, to 500 in Great Britain. Thirty-three U.S. agencies handled over $100,000,000 worth of advertising in 1970; the only agency outside the United States to have billings of over $100,000,000 was in Japan.

The typical advertising agency today is staffed to perform the following services for its clients: (1) planning marketing and advertising strategies; (2) writing copy; (3) preparing artwork, layouts, and television story boards; (4) buying artwork; (5) arranging for mechanical reproduction of printed advertisements; (6) producing television commercials; (7) budgeting expenditures; (8) selecting media; (9) contracting for space and time; (10) checking, billing, and paying media; (11) researching markets, media, copy, and results; (12) merchandising promotional programs to the trade; (13) counselling and engaging in public relations activities. Agencies performing all these functions are referred to as full-service agencies and are likely to be the larger firms. Many smaller agencies concentrate on planning, preparing, and placing advertising.

ECONOMIC AND SOCIAL ASPECTS

Advertising has been most highly developed in countries having an economic system that is basically a free market system, in which consumer choices largely determine what goods and services will be produced. The many suppliers competing for consumer favour use advertising to influence these choices by informing consumers that a given product exists, that it offers certain want-satisfying qualities, that it is priced at a certain amount, and is available at certain retail outlets. Advertising performs these communication functions quickly and efficiently. Although personal selling can be appropriate for reaching relatively small, well-defined, highly concentrated markets such as industrial buyers, advertising is the only feasible means for communicating with millions of ultimate consumers.

Manufacturers are willing to invest in research and development of new products because they know they have the means to introduce a new or improved product quickly to a large market, thereby increasing their chances of building a profitable volume of sales and recovering their initial investment in a reasonably short period of time. Having the means to tell the consuming public about a new product is, in itself, an incentive to invention and innovation.

The fact that large sums of money are spent on advertising and the knowledge that consumers ultimately pay this cost in the prices they pay has led some people to believe prices would be lower without advertising. Defenders of advertising argue that this overlooks the fact that advertising facilitates mass production with its concomitant lower unit costs and that a free market system counts on competition to pass along the benefit of such cost reductions in the form of lower prices.

The prices consumers pay are an indicator of the value they place on the product, and value is something more than the sum total of all the production and marketing costs. Value to the consumer is the sum total of his perceptions of the product and the gratifications he attaches to it. Anything that enhances his perceptions adds value. Advertising adds value by endowing a product with subjective qualities. Consumers who pay a 20 percent higher price for a well-known advertised brand over an unadvertised brand are indicating the greater confidence, and therefore higher value that they place on the advertised brand.

The use of advertising to build a brand preference among consumers and the fact that few firms can afford costly national advertising programs giving them the power to dominate particular product categories leads to the assumption that advertising tends to build monopolies. When a manufacturer registers a trademark or brand name, he is granted an exclusive right to use that mark or name. In effect he is granted a monopoly to sell the product bearing that brand name. Whatever reputation he is able to build for the brand accrues solely to that brand. Through advertising he seeks to differentiate his brand from competing brands, to embellish it with attractions so that a large number of consumers will prefer it over all others. He thus gains a degree of monopoly power. He can charge a premium price. He can count on some stability in sales. Whatever control he gains, however, is based on the loyalty of his customers to the brand. If the price differential becomes too great, he risks losing customers to competing brands. Brand monopolies, therefore, are subject to the whims of consumers and the machinations of competitors.

The high cost of using national media such as television on a large scale restricts the use of such media to the few advertisers who can afford them. Their ability to bear the high cost of introducing new products nationally and to outspend small competitors is thus assumed to discourage would-be entrants and to make it difficult for others to survive. It is further assumed that large firms gain a larger share of market because they can afford to gain a larger share of the public mind and that they can afford to hire the most sophisticated marketing managers thereby widening their competitive advantage. Undoubtedly, large firms with great financial resources can do many things—including large-scale advertising—that small firms cannot do. The opportunity to advertise, however, gives the small businessman a chance to get started. If he has a good product or service fairly priced he can advertise and build a market. He obviously cannot match the giant firms on a national scale, but he might be able to outspend them in his local market area. As his business grows he can use advertising to enter new markets and expand his own competitive capabilities.

The fact that the press in most of the non-Communist world is usually supported financially by advertising raises some questions about the freedom of the press from advertiser control. Many newspapers and magazines get 60 to 70 percent of their revenue from advertising. Commercial television and radio are financed by advertising. The notion that "he who pays the fiddler calls the tunes" underlies the suspicion that advertisers influence news coverage by suppressing and distorting the

news to their own advantage. There are some studies, however, that suggest if advertisers do dictate policy it seems to occur among the weaker or smaller newspapers or radio stations. A financially stable newspaper is generally able to resist pressure from advertisers. Moreover the advertiser's dependence on the press may be a deterrent to his intrusion on press freedom. He needs their circulation and audiences to advertise effectively. It is at least arguable that if he influenced news coverage, the press would risk losing public confidence, which in turn would reduce circulation and jeopardize the effectiveness of the press as an advertising medium.

BIBLIOGRAPHY. M. MAYER, *Madison Avenue U.S.A.* (1958), a perceptive report on the advertising agency business and national brand advertising strategies; H. NORINS, *The Compleat Copywriter* (1966), a contemporary view of creative work by a successful copywriter; S. BAKER, *Advertising Layout and Art Direction* (1959), on visual content, design, and graphics; C.H. SANDAGE and V.R. FRYBURGER, *Advertising Theory and Practice* (1971), a comprehensive analysis of the entire field including institutions, strategies, and procedures; L.O. BROWN, R.S. LESSLER, and W. WEILBACHER, *Advertising Media* (1957), on the interrelationship of media and marketing variables; R. BARTON, *Advertising Agency Operations and Management* (1955), a description of agency organizations and practices; W. TAPLIN, *Advertising: A New Approach* (1963), a nontechnical description of how advertising affects and is affected by consumer behaviour; N.H. BORDEN, *The Economic Effects of Advertising* (1942), the first and only systematic study of the relationship of advertising to classical economic concepts of supply, demand, competition, and prices; P. MARTINEAU, *Motivation in Advertising* (1957), an articulate exposition of motivation research and the symbolic content of advertising; H.W. BOYD and S.J. LEVY, *Promotion: A Behavioral View* (1967), a popularized translation of behavioural science concepts to marketing communications; F.P. BISHOP, *The Ethics of Advertising* (1949), a standard, nonlegal description of social responsibility in theory and practice; D.B. LUCAS and S.H. BRITT, *Measuring Advertising Effectiveness* (1963), an analytical review of popular copy-testing techniques and media research methods; F. WAKEMAN, *The Hucksters* (1946), V.O. PACKARD, *The Hidden Persuaders* (1957), and S.S. BAKER, *The Permissible Lie: The Inside Truth about Advertising* (1969), notable works in criticism of the advertising industry.

(V.R.F.)

Advertising Design

The purpose of advertising is to influence the minds of the people to whom it is addressed in a way conducive to the success of a campaign or promotional scheme. Among the means used to achieve this end is advertising design, which represents the aspect of "visual communication" in advertising media and advertising aids. Advertising media consist of sales letters, insertions in newspapers and magazines, enclosures, prospectuses, brochures, catalogs, leaflets, books, posters, films and slides, television tapes, aerial advertising, billboards, shop window displays, exhibits, and so on. Advertising aids may be of a direct or indirect kind. The direct aids include packaging, wrapping paper, book jackets, phonograph record sleeves, and the like, which serve not only to protect and identify the contents but also to advertise. Indirect aids include any number of items that tend to inspire confidence in a firm and help to shape its image favourably, such as calendars, greeting cards, and various gifts.

Varieties of advertising media and aids

Advertising requires that its ideas be given a visual form by artistic and graphic design. The means used to this end are typography, drawing, photography, and colour. The secret of good and effective design is to conceive the composition as a whole in which form and content, and representation and expression, are interdependent and matched to each other.

GENERAL CONSIDERATIONS

The relation of advertising design to other arts. The function of advertising design is primarily a practical one directed toward a particular end. It must communicate quickly and clearly and, in so doing, (1) attract attention; (2) lodge the advertising message in the conscious or unconscious mind; (3) induce an active, positive attitude to the message instead of a passive one; and (4) impress a favourable memory of the message on the mind. Because advertising design is aimed at producing an effect desired by an advertiser, it is obviously not free to choose its theme as the fine arts are. The function of the fine arts is primarily aesthetic. Even so, both advertising design and art use the same resources for the realization of their ideas: form, colour, and composition. Advertising design may use these effects successfully without necessarily producing art; it has served its purpose if image and text create the desired effect. A free work of art, on the other hand, becomes true art only when all the components of the design have genuine qualities of form and colour and together make up a vital composition which is complete in itself. Advertising design, however, can also rise to art if the designer has a strong artistic personality and an original imagination and if the commission allows the artist a certain freedom in the way he presents his theme and selects his artistic means—for example, the famous posters of Toulouse-Lautrec.

The relation of advertising design to technology. Both the nature of advertising design and the tasks that it has to perform have been profoundly influenced by advancing technology. New sophisticated techniques of reproducing and representing acoustic and visual material have given advertising new scope. To relief, intaglio, and lithography has now been added silk-screen printing. Hand composition, monotype, and linotype have been supplemented by phototypesetting techniques. Cinema slides and advertising films have been joined by television commercials. Stationary, two- or three-dimensional signs have been supplemented by moving displays that enable objects to be seen in all their perspectives. Many such displays are programmed by computers.

Techniques used in design

The technical and artistic demands on the advertising designer have become more exacting. The variety of tasks that he has to accomplish calls for detailed knowledge of printing techniques and composition, of cinema and television techniques, of the properties and working qualities of a great diversity of materials used in exhibition design, of the technique of audiovisual presentation and multiprojection, of the potentialities of electronic sign advertising, and of the aids that can be obtained from computers. The advertising designer must familiarize himself with these new techniques and their effects, thereby acquiring more scope for his artistic imagination, a greater field of application for his ideas, and a more comprehensive impact for his advertisements.

The role of psychology in advertising design. The success of advertising design depends on the topicality of what is depicted, on the quality of the design, and, to no less extent, on psychological factors. These largely determine the reactions of the person to whom the advertisement is addressed. Consciously perceived information is capable of inducing moods and feelings that are in harmony with the intention of the advertisement or at variance with it. The static use of the designer's resources in a composition, for instance, may express rest, solidity, conservatism, security, and trust, whereas a dynamic use may evoke love of experiment, flexibility, a forward-looking attitude, aggressiveness, and youthfulness. The correct choice of colours is especially important. Colours with a symbolic significance in politics or religion are taboo and must be used with caution, depending on the country and religion involved.

The role of the advertising designer or commercial artist. With the increasing sophistication of advertising after World War II, the importance of the free-lance advertising designer naturally diminished. Advertising campaigns, some of which have become very extensive and complex, can be planned and carried out only by specialists working as a team. A full service advertising agency will handle all the problems arising in connection with an advertising assignment, including budgeting, and thus has a far greater capacity than the designer working on his own. The free-lance advertising designer, generally speaking, can only execute commissions for smaller firms or take on jobs assigned to him by the agencies in connection with advertising campaigns.

The designer as an employee of an agency

The upshot of this development is that the advertising designer has virtually become a staff employee and team colleague of the agency. His freedom as an independent creative designer is thus restricted to the extent that he must subordinate his gifts to the advertising conception elaborated by the marketing experts. Depending on his talents, he will be employed in a variety of capacities or as a specialist in one particular line—for example, as an illustrator, letterer, draftsman, filmmaker, or display designer. In most cases his work remains outwardly anonymous and is integrated with the overall output of the agency. Actually, his work is identified by the layout man, who sees to the rough draft of the advertisement and prepares sketches showing how the separate units are to be integrated into a whole. The designer also works in close collaboration with the copywriter and the contact man who is in touch with the agency's client.

KINDS OF MEDIA

Graphic or printed, drawn, and painted media. These include advertisements in periodicals, posters, brochures, leaflets, pamphlets, buttons, and the like. Frequently they contain a "device," "trademark," or "trade name" used to denote the maker or supplier of the product or service. The device or trademark represents the most concise graphic form that a particular subject can be given. It should be simple, unmistakable, applicable, and timeless. The trade word has the advantage over the device in that it can be read at once, whereas the meaning of the device must first be made known. Its disadvantages, of course, are that it contains more formal elements (depending on the length of the name) and that the designer's freedom to shape it at will is limited by the demands of legibility; it thus usually has a weaker impact.

The special characteristics of each medium

Each medium has its own demands. Whereas a poster designed to produce an effect at a distance, for instance, must be drawn and coloured on a generous scale, a newspaper advertisement has to be worked out with greater attention to details. In any case, the idea is to achieve the maximum of clarity and impact with the minimum resources of colour and form. The illustrative poster, for example, presents the theme or object with an atmospheric appeal; the realistic poster supplies objective information through its copy and the truthful depiction of the object; the constructivist poster arranges its pictorial elements in accordance with their thematic importance. The experimental poster surprises the viewer by the novel way in which it presents a problem or by the new techniques it employs.

Kinetic or moving media. Television commercials and other kinds of films afford the advertising designer a variety of interesting opportunities. If the film or tape consists of live action, then titles, subtitles, possibly functional diagrams, statistics, movie sets, and so on have to be designed. If the film consists of an animated cartoon, then the designer's imagination is given almost unlimited scope; every type of drawing can be used, from fairy-tale illustration to surrealistic montage, and every kind of style can be employed—naturalism, simplified and stylized presentation, abstract symbolism, and so on. But, in any event, it is important that the designer depict his pictorial ideas clearly and simply, that they do not change too quickly, and that he not present more than one or two ideas in the picture at the same time. He must consider colours and their sequences with the same care. In all types of films, lettering, image, word, and sound must complement one another, and thus the designer must collaborate closely with the producer, the copywriter, and the cameraman to make sure that the various components work together to enhance the total impact. The eye and mind of the viewer must be captured amid the multitudinous visual impressions.

In parades and demonstrations there are posters, transparencies, banners, and boards, but, unlike most ordinary posters, they are continually in motion and must be visible from various angles. Lettering and images must be recognizable and legible at a distance.

Identifying and directional signs. One of the most interesting jobs in advertising design is the creation of a uniform "image" for a business or organization. All advertising aids from visiting cards through commercial forms and newspaper advertisements to outdoor signboards should be given an unmistakable character with regard to graphic form, typography, and colour, so that each element is seen as part of an overall conception. All these different impressions must have a family resemblance, so that their effect is cumulative and produces a clear image in the eye of the beholder. Uniform advertising is a powerful aid in the creation of a firm's image.

Creating the "image" of the firm

In comparing lettered signs on doors, windows, buildings, factories, and so on, the advertising designer is confronted by three specific problems: the lettering must be visible at a certain distance; it must make an impact by reason of its formal originality; and it must be integrated with the architecture in size, form, colour, and material. If the lettering is to be visible at night, the designer can make his choice between neon lettering and spotlighted lettering. If he chooses neon lettering, he must adapt the shape of his letters to technical or engineering requirements.

In composing lettered signs on delivery vans, trucks, and so on, the designer must take the factor of movement into account: the letters must be somewhat spaced out so that they will not run together when read on the moving vehicle. Similar care is also needed in selecting colours; as a rule dark colours are more suitable because they produce more contrast.

Although not always thought of as part of advertising, directional signs and other communication systems, if well designed, help to promote a firm's good public relations. These are the signs that figure importantly at airports, railway stations, administrative buildings, and other places where large numbers of people have to be directed to their destination by the shortest route in the shortest possible time. The signs and lettering must be uniform in design, clear and factual, and somehow correspond to the organization's advertising design or image. Large airlines, for instance, have recognized that uniform liveries for their employees and uniform lettering for their check-in counters, passageways, and so on are important for advertisement as well as for directional signals, and they have engaged graphic artists to design them.

Promotional items. There is an enormous variety of items that an advertiser can distribute in order to attract the attention of the public it addresses. These include calendars, matchbooks, glassware, ashtrays, notebooks, writing instruments, playthings, and decorative items of all sorts. Such items may be distributed year-round or be used as holiday gifts.

How long an object will continue to attract attention depends on the quality of the material, its workmanship, and the timelessness of its form or colour arrangement. The designer must be guided by aesthetic considerations in affixing the name of the firm, which may be stamped or printed on the object, attached to it on a label, or cast in its interior if the material is transparent. A decorative version of the name may be a factor in determining the form of the object, but often it is placed discretely on the underside.

HISTORICAL DEVELOPMENT

Origins. Advertising is as old as humanity itself. Advertising tries to do what every man wants to do: to influence others, to win their suffrage, and to convince them. First of all there was the word, which expressed man's thoughts and feelings; then followed the picture, which gave his ideas lasting form; and finally came writing, which perpetuated the spoken word beyond the moment of its utterance.

Ancient designs and posters

In the ancient civilizations of China, Mesopotamia, Egypt, Greece, and Rome, tradesman and merchant became recognized callings. They used books of account, business deeds, contracts, written powers of attorney, and—most important here—stamps for making their products, marks to denote the ownership of their livestock, and seals that were recognized and protected by the law as pictorial "signatures." These were the prede-

cessors of today's trademarks. In Rome there were a number of marketplaces and streets reserved for particular craftsmen and tradesmen or their goods and bearing their names. Professional "poster painters" inscribed election manifestos and advertisements on walls (*dipinti* and *graffiti*), and some 1,600 of these announcements have been found in Pompeii alone. Frescoes painted on the houses of tradesmen publicized their products or their work.

The next most crucial development was the invention of printing in the 15th century. Printing with movable letters allowed the distribution of an unlimited number of copies of an item. In the following three centuries, from the 15th to the 18th, there appeared an ever expanding number of books, calendars, leaflets, travel guides, exhibition guides, book catalogs, playing cards, chronicles, newspapers, advertising sheets, weeklies, letterheads, certificates, labels, business cards, maps, and posters. The prosperous condition of the printing trade gave rise to printer's marks, watermarks, bookplates, monograms, trademarks, and devices, and encouraged the development of various type fonts. Printers and type designers cut letters that bear their names and are still in use today, being frequently employed by the advertising designer. Claude Garamond cut his type in 1544, William Caslon in 1722, John Baskerville in 1757, Firmin Didot in 1783, and Giambattista Bodoni in 1788. François-Ambroise Didot invented the typographical point system, which has become indispensable for establishing type sizes. The printed picture also received attention. At first the primary function of the illustration was to support the text, but later it became an attractive means of visual communication in its own right, reaching its peak in the 19th-century poster.

19th and early 20th centuries. The industrialization of the West in the 19th century and the explosive development of the economy gave impetus to the graphic arts. Advertising began to make a more deliberate use of means of visual communication. Government authorities, publishers, business firms, industrialists, and politicians harnessed newspapers, magazines, posters, and pamphlets to their purposes. Interest was shown in the design of letterheads, packages, wrapping paper, and shop signs.

Development of lithography and photography

The invention of lithography, or chemical printing from stone, in 1795–98 greatly widened the scope of reproduction and printing. It allowed the finest lines to be printed and at the same time was simpler and cheaper than letterpress printing. In 1827 it was followed by colour lithography. The negative aspect of this invention was that a large number of new fonts appeared which were suitable for the lithographic technique but which were generally inferior aesthetically. Among the exceptions were "Egyptian," designed by Vincent Figgins, and the "Display Sans Serif," developed by the Berthold type foundry in Berlin and Vienna. Since World War II this latter face has been enjoying a renaissance in Europe.

In 1826, Joseph-Nicéphore Niepce succeeded in producing an image on a metal plate coated with light-sensitive asphaltum and thus invented photography. Louis-Jacques Daguerre improved the process with his daguerrotype, with which it was possible to make a usable negative. In 1903, the newspaper *Le Matin* printed the first half-tone block from a photograph. In the same year Louis Lumière invented colour photography. Concurrently, "moving" photography, or motion pictures, was being developed. Science, industry, process engraving, and advertising soon found that photography and films had become indispensable mediums of information. In the latter part of the 19th century, Jules Chéret, Toulouse-Lautrec, T.A. Steinlen, Pierre Bonnard, and others, influenced by the coloured woodcut of Japan, brought the lithographic poster to an artistic peak, which at the same time marked the birth of the modern poster. In the early 20th century, Max Klinger, Ludwig Hohlwein, and other artists working in Germany produced the realistic picture poster, which displays the subject or theme starkly and vividly. About the same time there appeared the first Constructivist posters by Peter Behrens (in Berlin) and J. Hoffmann and Joseph Olbrich (in Vienna), in which the picture is built up architecturally and geometrically.

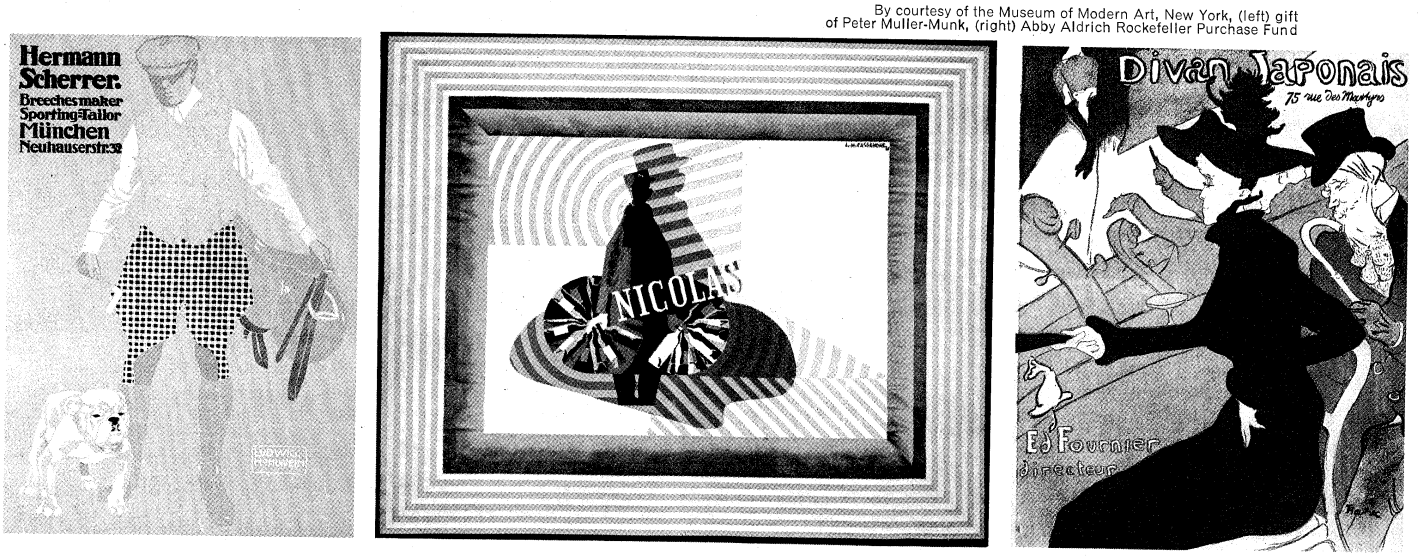
In the mid-19th century, advertising agencies appeared in Germany and thereafter engaged in the design and placing of advertisements in newspapers, periodicals, and catalogs. Well into the 20th century, however, the artistic quality of their output often was poor.

From the 1920s to the present. After World War I there was an efflorescence of intellectual and creative activity in all fields of design. Soviet Constructivism offered Cubist and geometric themes; Italian Futurism presented the idea of the machine with an almost photographic vision; the de Stijl movement in Holland espoused rectangular forms and primary colours; Dadaism and Surrealism in Switzerland, Germany, France, and the U.S., offered fantasy, humour, and near animation. Among the most renowned advertising designers of this period between two wars was A.M. Cassandre, whose witty and formally sophisticated posters dominated the billboards in France.

Bauhaus influence

The Bauhaus, the art school founded by Walter Gropius in 1919, became the centre of modern design. Fundamentally, it sought to end the old schism between the artist and the technically expert craftsman by training

By courtesy of the Museum of Modern Art, New York, (left) gift of Peter Muller-Munk, (right) Abby Aldrich Rockefeller Purchase Fund



POSTER DESIGN FROM 1890–1940.
(Left) "Hermann Scherrer. Breechesmaker Sporting-Tailor," lithograph poster by Ludwig Hohlwein, 1911. (Centre) "Nicolas," offset lithograph poster by A.M. Cassandre, 1935. (Right) "Divan Japonais," lithograph poster by Henri de Toulouse-Lautrec, 1892.

students equally in both fields; in effect it aimed at preparing the artist-designer for the industrial problems of mass production.

The Bauhaus revolutionized graphic design as well as other areas of art and architecture. The aims of the Bauhaus school of artists were pursued by many other more independent artists. Jan Tschichold, for instance, laid down the principles of "basic typography," which were to be functional, utilitarian, and designed to convey a message that was short, simple, forceful, and free from ornamentation.

When the Nazis closed the Bauhaus in 1933, some of the professors emigrated to the United States and Switzerland. In 1937 László Moholy-Nagy opened the Institute of Design (known as the New Bauhaus) in Chicago, where he continued as director and teacher to propagate the Bauhaus ideas. Using a variety of fonts, the American advertising designer gained acceptance for functional typography, and eventually the characteristic American typography developed out of it. Photomontage, the technique of combining photographs that Moholy-Nagy, Herbert Bayer, and Herbert Matter had experimented with in Europe, was taken up and widely used for commercial purposes. In the 1960s Art Nouveau, a romantic use of asymmetrical, undulating lines, and Pop art, a direct depiction of comic and commonplace things, found their way into graphic design. After World War II Switzerland became a refuge for Bauhaus ideas; Johannes Itten went to Zürich and became director of the Kunstgewerbeschule (School of Arts and Crafts) there. The so-called Swiss school (including Richard P. Lohse, Josef Müller-Brockmann, Hans Neuberg, and Karl Gerstner) acquired an international reputation through its systematic and functional advertising.

By courtesy of J. Muller-Brockmann



"Weniger Lärm" ("Less Noise"), lithograph poster by Josef Müller-Brockmann, 1960.

Prospects for use of information theory

A newer development that is still in the experimental stages is the adoption of information theory to assist advertising design. Information theory is concerned primarily with the discovery of mathematical principles governing systems designed to communicate and manipulate information. It attempts to measure the capacity of various systems, including the human brain, to transmit, store, and process various pieces of information. Measurements of the effects of colours, lights, sound levels and pitches, frequency of signals, and so on have obvious usefulness for the designer.

Other notable developments since World War II include the formation of national and international bodies representing the interests of the profession, the appearance of professional journals and annuals, and also books on design problems, and finally the organization of international design conferences and international poster and graphic design biennials.

BIBLIOGRAPHY

General works (all theoretical): RUDOLF ARNHEIM, *Art and Visual Perception* (1965) and *Visual Thinking* (1969); MAX BENSE, *Aesthetica* (1965); GYORGY KEPES, *Vision and Value Series* (1965); MARSHALL MCLUHAN, *The Gutenberg Galaxy: The Making of Typographic Man* (1962) and *Understanding Media* (1964); ABRAHAM MOLES, *Théorie de l'Information et perception esthétique* (1958; Eng. trans., *Information Theory and Esthetic Perception*, 1968) and *L'Affiche dans la société urbaine* (1969); HANS RONGE, *Kunst und Kybernetik* (1968).

Kinds of Media: KARL GERSTNER, *Programme entwerfen* (1964; Eng. trans., *Designing Programmes*, 1964); FREDERIC HENRY KAY HENRION and ALAN PARKIN, *Design Coordination and Corporate Image* (1967); RICHARD P. LOHSE, *Neue Ausstellungsgestaltung* (1953; Eng. trans., *New Design Exhibitions*, 1954); JOSEF MULLER-BROCKMANN, *Gestaltungsprobleme des Grafikers* (1961, Eng. trans., *The Graphic Artist and His Design Problems*, 1961).

Historical and stylistic development: HANNS BUCHLI, *6000 Jahre Werbung: Geschichte der Wirtschaftswerbung und der Propaganda* (1962); KARL GERSTNER and MARKUS KUTTER, *Die neue Graphik* (1959; Eng. trans., *The New Graphic Art*, 1959); GYORGY KEPES, *Language of Vision* (1945); LASZLO MOHOLY-NAGY, *Visions in Motion* (1947); JOSEF MULLER-BROCKMANN, *Geschichte der visuellen Kommunikation* (1971; Eng. trans., *A History of Visual Communication*, 1971); and with SHIZUKO, *Geschichte des Plakates* (1971; Eng. trans., *History of the Poster*, 1971); JAN TSCHICHOLD, *Die neue Typografie* (1928); HANS MARIA WINGLER, *Bauhaus, 1919–1933* (1968).

Periodicals: Annual of Advertising and Editorial Art; Graphis Annual; Prisme International Annual; Gebrauchsgraphik (monthly); *Graphic Design* (quarterly); *Graphis* (bi-monthly); *Print* (bi-monthly).

(J.M.-Br.)

Aegean Civilizations

The adjective that is traditionally used to name those Bronze Age civilizations that arose and flourished in the area of the Aegean Sea between *c.* 3000 and 1000 BC is Aegean.

The area consists of Crete, the Cyclades and some other islands, and the Greek mainland, including the Peloponnese, central Greece, and Thessaly. The first high civilization on European soil, with stately palaces, fine craftsmanship, and writing, developed in the island of Crete. Later, the peoples of the mainland adapted the Cretan civilization to form their own, much as the Romans adapted the civilization of later Greece. The Bronze Age civilization of Crete has been called Minoan, after the legendary King Minos of Knossos, the chief city of the island throughout early times. The Bronze Age of the Cyclades is known as Cycladic, that of the mainland as Helladic, from Hellas, the Greek name for Greece. Early, middle, and late stages have been defined in each of these, with further subdivisions according to recognizable changes in the style of pottery and other products associated with each. The civilization that arose on the mainland under Cretan influence in the 16th century BC is called Mycenaean after Mycenae, which appears to have been its most important centre. The term Mycenaean is also sometimes used for the civilizations of the Aegean area as a whole from *c.* 1400 BC onward.

Dating of the Aegean Bronze Age. The dates suggested here are approximate and conventional. In a general way, they are based on correlations with Egypt, where, from the beginning of the Dynastic Period (*c.* 3100 BC onward), a historical chronology can be established with a leeway of a few centuries and is fixed within reasonably narrow limits after *c.* 2000 BC. Bronze Age pottery from the Aegean has been found in datable contexts in Egypt, and many Egyptian objects have been recovered in the island of Crete.

Two important landmarks are fragments of Cretan pottery from the town at Kahun in the Fayyūm, built for



Principal sites associated with Aegean civilizations.

workers who were engaged in the construction of a pyramid for the pharaoh Sesostris II (1897–78), and a large quantity of Mycenaean pottery from the mainland found at Tell el-Amarna, site of Akhenaton's capital, and imported during his reign (1379–62). Radiocarbon dates appear consistent with those based on correlations with Egypt.

History of exploration. The poems of Homer, although they reflect Greek society in the 8th century BC, describe warriors using obsolete bronze weapons and objects such as helmets plated with tusks of wild boar that went out of use before the end of the Aegean Bronze Age. Massive Bronze Age defense walls survived at Mycenae and elsewhere on the mainland and were called Cyclopean because, according to Greek tradition, the Cyclopes had built them. Apart from these Cyclopean walls, virtually nothing was known about the Aegean Bronze Age before the middle of the 19th century, when, in 1876, a German archaeologist, Heinrich Schliemann, discovered unplundered royal shaft graves at Mycenae and thought that the men buried in them were the Greek heroes of Homer's siege of Troy. Actually, the graves spanned the 16th century BC and proved to be older than any likely date for a Trojan war. Schliemann's discoveries, however, led to intensive exploration of Bronze Age and earlier sites on the Greek mainland. Meanwhile, in the 1860s, Bronze Age settlements of the period of the Mycenae shaft graves had come to light under the debris of an immense eruption on the volcanic island of Thera (Thíra). Later in the 19th century, Christos Tsountas, a Greek archaeologist, dug cemeteries of earlier phases of the Bronze Age on other Cycladic islands and continued the work begun by Schliemann at Mycenae. At the end of the century, a British expedition excavated an important Bronze Age town at Phylakopi on Melos (Mílos). When Crete eventually became independent of Turkish rule in 1898, attention was turned to Bronze Age sites there. In 1900 Sir Arthur Evans, an English archaeologist, began to uncover the palace at Knossos, the largest Bronze Age centre of the island, discovering clay tablets with the first positive evidence for Bronze Age writing in the Aegean. Greek, U.S., French, and Italian excavators added further knowledge of the Cretan Bronze Age during the years that followed, and U.S. and German expeditions opened new sites on the mainland. Inscribed clay tablets such as those found at Knossos in Crete at the turn of the century were recovered in Messenia; others have since come to light at Mycenae and elsewhere on the mainland. The belief that the language of these tablets might be Greek was established by an English archaeologist, Michael Ventris, in 1952, though acceptance of this is not yet universal. In 1962 a large palace, destroyed by fire *c.* 1450 at Zákros (Zakro) in eastern Crete, was found. In 1967 the clearing of a major settlement that was buried by the eruption of Akrotíri, on Thera, began.

The work of Schliemann

This article is divided into the following sections:

I. Early Aegean civilizations
Early cultures
Paleolithic (Old Stone Age)

Neolithic (New Stone Age)
The Bronze Age
The Early Bronze Age (*c.* 3000–2200)
End of the Early Bronze Age on the mainland (*c.* 2200–2000)
The Middle Bronze Age on the mainland (*c.* 2000–1550)
The Shaft Grave period on the mainland (*c.* 1600–1450)
Period of the Early Palaces in Crete (*c.* 2200–1700)
Period of the Later Palaces in Crete (*c.* 1700–1450)
II. The decline of the early Aegean civilizations
The eruption of Thera (*c.* 1500) and the conquest of Crete (*c.* 1450)
A new social order
Events on the mainland
Conquest of the Cyclades
The fall of Knossos and the Mycenaean empire
The end of the Bronze Age in the Aegean (*c.* 1200–1000)
Invasion from the north
The coming of the Greeks
The people of the Bronze Age Aegean
Physical appearance
Dress
Society
Trade
Transport
Warfare
Religion

I. Early Aegean civilizations

EARLY CULTURES

Paleolithic (Old Stone Age). Chipped stone tools made by Paleolithic hunters have been found in many parts of mainland Greece, but none are yet recorded from Crete or the other islands. As elsewhere in Europe, the latest Paleolithic industries evolved into Mesolithic (Middle Stone Age) ones with diminutive stonework. During this Mesolithic stage, obsidian, a volcanic glass, was brought from the island of Melos for making implements, indicating that boats capable of sea voyages were already in use in the Aegean area by then.

Neolithic (New Stone Age). If radiocarbon dates are to be trusted, agriculture was being practiced in some parts of the Aegean area as early as the 7th millennium. The first agriculturalists in the Aegean, like those of Anatolia and Palestine, may have been ignorant of the art of making fired clay vases—traces of agricultural settlements without pottery have been identified at several places in Thessaly and at Knossos in Crete. The island of Crete appears to have been uninhabited before this time, and the first agriculturalists must have reached it by sea from western Anatolia or from somewhere more distant, such as Cilicia. Other immigrants from the east may have brought agricultural techniques and ways of life to the mainland, where they mingled with the Mesolithic hunting peoples whom they found there. The Aegean is one of the most favoured regions of the Mediterranean basin. Immigrants from the coastal areas of Anatolia, Syria, or Palestine would have found the climate and ecology similar to what they had known in their homelands. The olive and vine, sources of oil and wine, the staples of Mediterranean diet, grow in most parts of the Aegean area and may have been native there. Water, which is a problem in the present century, was probably more abundant in early times when forests were more extensive than they are now.

Ecology of the Aegean

Agricultural communities were eventually established in every part of Greece, making pottery by hand and grinding stones to shape edged tools, axes, adzes, and chisels. Wheat and barley were among crops grown. Hunting and fishing still played their part in obtaining food alongside raising of cattle, sheep, goats, and pigs. Arrowheads of chipped stone were used on the mainland and in the Cyclades, but none is recorded from Crete, where bone points may have served to tip arrows. Another long-range weapon was the sling, and clay sling pellets were made in Thessaly where suitable beach pebbles were not to hand. In Crete, clubs were armed with stone heads as in Egypt and elsewhere in the Near East in early times. Houses with rectangular rooms are attested at Knossos in Crete, at Saliagos in the Cyclades, and in Thessaly. But some Aegean communities may have lived in circular huts of the kind found in Predynastic Egypt and in early Syria and Cyprus. Caves appear to have been used as cemeteries in Crete. Some cremations are assigned to the end of the Neolithic on the mainland. By that time, to judge from a house, larger than the rest, occupying a central position at Dhimíni (Dhimínion) in Thessaly, some communities may have been organized under chiefs. Several Thessalian settlements were surrounded by defense walls or ditches. Copper tools—simple, flat axes and knives—were in use before the end of the Neolithic both in Crete and on the mainland. Indeed, some of the pottery shapes characteristic of later phases of the Neolithic in Thessaly appear to be copied from those of metal vases. But this Thessalian late Neolithic may actually overlap in time with the beginning of the Bronze Age farther south.

THE BRONZE AGE

The Early Bronze Age (c. 3000–2200). The transition from Neolithic to Bronze Age in the Aegean was marked by changes in pottery and other aspects of material culture. These may reflect the arrival in Crete and the Cyclades of new people from more civilized lands farther east bringing knowledge of metalworking with them. In Crete and the islands the changes that inaugurated the Bronze Age were more or less contemporary with the beginning of Dynastic times in Egypt. What is defined as the Bronze Age in the Peloponnese appears to have begun later under the influence of settlers from the islands. The Bronze Age in central Greece and Thessaly may have begun later still. Evolved types of metal tools appear to have been current considerably before the end of the Neolithic there.

Flourishing metal-using cultures were established by the middle of the 3rd millennium in Crete, the Cycladic islands, and the southern part of the mainland. Each of these three cultures had its own distinctive characteristics; but they shared much in common, and their peoples may have spoken the same or similar non-Greek languages. Many place-names throughout the Aegean, notably ones ending in *-nt-* and *-ss-*, such as Corinth and Knossos, seem to reflect a time when the same language (or a group of related languages) with Anatolian affinities was spoken there before the introduction of Greek. A large number of words came to be adopted into Greek from this earlier language group.

Early metal tools

These Early Bronze Age peoples of the Aegean seem to have employed similar types of metal tools, including axes, adzes, and short daggers, but double axes may have been special to Crete. Tweezers were used for plucking facial hairs and rectangular stone palettes for grinding face paints with small pestles made of attractive veined stones or Spondylus shell.

Architecture. Lerna and other settlements on the mainland were eventually surrounded by massive walls with projecting towers, but whether similar defenses existed in the Cyclades and in Crete is uncertain. Houses with several rooms were being constructed in most parts of the Aegean by this time, and buildings at Knossos and at Vasilikí in Crete have been identified as the residences of local rulers. The so-called House of Tiles at Lerna, destroyed by fire toward the end of the period, appears to have been another of these "palaces." It was a large rectangular building with two stories taking its name from the baked clay tiles found in its ruins. These small, flat tiles are thought to have come from a sloping roof and may be the earliest roof tiles known. Similar tiles were recovered from a huge circular structure of the same period at neighbouring Tiryns, of which only a section has been excavated, as it lies deep below the level of the later Mycenaean palace there. It was evidently a public building of some kind, a meeting place, perhaps, or possibly even a temple. Some shrines of this period in Crete may have been circular, as were many of the early Cretan tombs.

But circular buildings at Orchomenus in central Greece are interpreted as granaries or huts. A remarkable stone

vase of this period, from Melos, is in the shape of a granary complex with round silos, such as the granary of the later palace at Mallia (Mália) in Crete. The vase shows an entrance porch with a gabled roof, but it would appear that the roofs of houses in the Cyclades and in Crete were mostly flat at this time and later.

Minoan burials. Cretans in the Early Bronze Age buried their dead in communal tombs. These belonged to clans or extended families and might remain in use for many generations. Traces of hundreds of burials have been noted in some of them. Caves and rock-shelters, as well as buildings of various kinds, were used as tombs. Circular tombs were characteristic of the Mesara (Mesarás) region of southern Crete. They were built above ground, and none has survived complete; but the massive walls that are left seem designed to support the weight of vaults, which may have been constructed of rough fieldstones held together with mud, as were the walls. Some of the largest tombs, however, with a diameter of 40 feet (12 metres) or more inside, may have been vaulted in mud brick. Annexes with cult rooms were built in front of the entrances of some tombs, and others had chambers for offerings around the sides. When a tomb became full, a new floor was laid above the earlier burials, or parts of the tomb's annex were brought into use as burial chambers. Sometimes the remains of earlier burials were removed to separate buildings or enclosures nearby. These circular tombs may have been introduced to Crete by people whose ancestors had lived in round huts. The inhabitants of the Neolithic settlement at Khirokitia in Cyprus, for instance, buried their dead under the floors of the round huts in which they lived. Communal tombs in some parts of Crete had rectangular compartments or rooms and flat roofs, such as those in contemporary houses. At Knossos, where the local rock was soft, artificial caves were dug as tombs. Everywhere in Crete the dead were normally trussed into a tight contracted position, knees to chin. Sometimes the bodies were then squeezed into large storage jars or small clay chests or coffins. There was evidently much ceremonial in connection with burial, and, apart from objects of personal use, such as seals, jewelry, and weapons, left with the dead, vases with offerings were regularly placed inside or outside the tombs.

Cycladic burials. In contrast to the Cretans, the people of the Cyclades during the earlier part of the Bronze Age buried their dead in small graves that held a single body, sometimes a pair, and never more than six or seven. The graves were often grouped in family cemeteries, which might be surrounded by a wall. The bodies were placed in them lying on one side in a loosely flexed position. Some Cycladic graves were little, stone-built chambers with an entrance, although the standard type consisted of a box (cist) made with large slabs set on edge and roofed with slabs.

Helladic burials. Less is known about contemporary burials on the mainland, but the graves there normally contained several bodies, which suggests that they belonged to families but not to large units, such as the clans that existed in Crete. Various types of mainland graves of this period are known, including chambers cut in the rock and stone-built tombs, such as those in the Cyclades. Circular cairns, each covering several burials, on the island of Levcas in western Greece appear to go back to this time.

Early pottery

Aegean arts. Pottery was still made by hand throughout the Aegean area. A useful type of vase first attested there at the beginning of the Early Bronze Age was a handled jug with a spout for pouring. Some of the earliest jugs from Crete have round bottoms and yellowish surfaces, as if they were copies of vessels made from gourds. Distinctive spouted bowls of oval shape that may have been drinking cups but are nicknamed sauce boats were quite typical of the Early Bronze Age on the mainland and usually have a fine reddish or dark overall wash. Pottery with a similar wash and with the surface often deliberately mottled is found in Crete and is known as Vasilikí ware, after a site with a little "palace" where large amounts of it were recovered. The pottery of the Cyclades during this period shares features in common with that of Crete and the mainland. The art of making stone vases flourished in the Cyclades from the beginning of the Bronze Age and may have spread from there to Crete, but it was hardly practiced on the mainland after the end of the Neolithic. The techniques used were simple and included boring with a hollow reed, which twirled an abrasive, either emery from Naxos or sand. The people of the Cyclades used their fine white marble not only for vases but for remarkable figurines, mostly female but including men, some playing double pipes or seated on chairs with harps. While the majority of these figures are only a few inches high, some females are larger and a few are nearly life-size. Some have traces of painted decoration. These marble figures were often placed in graves, and groups of them have been found in sanctuaries, though whether they represented gods and goddesses is uncertain. They were exported to the mainland and Crete and may have been imitated there. Vessels of gold and silver were current in the Aegean by now, and a few have survived, including two gold sauce boats of mainland type and gold and silver bowls from the islands. Gold and silver jewelry of this period, mostly from Crete, includes bracelets, necklaces, earrings, headbands, and hair ornaments of various kinds. Some of the finest of this early jewelry was found in communal tombs at Mochlos (Konídha) on the northern coast of Crete. The inspiration for it no doubt came from the east, and much of that from Mochlos, notably hairpins with flower heads, is reminiscent of jewelry from the royal tombs at Ur in Mesopotamia.

Seals began to come into use in the Aegean about the middle of the 3rd millennium. In an age before the invention of locks and keys, seals were employed to stamp wet clay, which was used to secure doors or affix lids to storage jars or other containers. The design impressed by the seal might add the threat of magic to that of detection if the sealing was broken. Many seals resembling those current in Egypt and Syria have been recovered from early tombs in Crete. Some of these early Cretan seals were made of ivory, possibly obtained from Syria, where elephants still abounded. Others were made of bone or soft, easily cut stones, such as serpentine and steatite. They were of various shapes, some of animals or birds or their heads, others cylindrical, adapted from Syrian versions of early Mesopotamian cylinder seals. They were engraved with a variety of designs, including abstract patterns and pictures of animals, notably dangerous lions and scorpions or poisonous spiders (rogalidhas) of a species native to Crete. Seals appear to have been in use in the Cyclades and on the mainland during this period, but very few have been recovered there. A stone cylinder seal from Amorgos resembles early Syrian ones. Most Aegean seals of this period, however, even in Crete, may have been made of perishable wood. Clay-seal impressions, preserved by fires that destroyed buildings at Lerna, including the palatial House of Tiles, look as if they had been stamped by wooden seals. The designs on these Lerna impressions resemble those found on seals in Crete.

Protective seals

Pictures of boats with many oars or paddles were scratched on clay vases of this period in the Cyclades and on the mainland. The boats have a high prow often surmounted by a fish ensign, the stern being low with the keel apparently projecting beyond it. Similar vessels, though with a single mast for a square sail in addition to oars, are represented on early Cretan seals. Ships of this kind would have been capable of voyages to Syria and Egypt, whence skills and fashions were reaching Crete along with imports such as Egyptian stone vases and Syrian daggers.

End of the Early Bronze Age on the mainland (c. 2200–2000). The comparative unity of incipient civilization in the Aegean area was eventually shattered by new movements of people into the Cyclades and the southern part of the mainland. Toward the end of the 3rd millennium, many of the settlements on the mainland, such as that at Lerna, were destroyed by fire, and the houses built afterward were of a different type and more primitive. These

new houses were long and narrow, only one story high, and apparently gable roofed. The entrance was at one end, and there was often a small compartment, which might be semicircular (apsidal), at the other. The new houses were evidently built by foreign invaders settling in the places they had destroyed. But some of the previous inhabitants may have survived as hewers of wood and drawers of water. There was continuity in the pottery, although new shapes and fabrics appeared alongside the old and sauce boats were no longer made. The site of the House of Tiles appears to have been reserved as sacred or unlucky ground, with a ring of large stones above its burnt ruins.

Either now or, possibly, somewhat later, there is evidence of a complete revolution in burial customs on the mainland. Family tombs were replaced by individual graves, usually lined and covered with slabs, such as the cist graves of the Early Bronze Age in the Cyclades. But these later graves on the mainland are often found inside the settlements, whereas before this burial outside the settlement area appears to have been the rule everywhere in the Aegean. The custom of "intramural" burial had lingered, however, in the hinterland of Anatolia, where houses were built in the style of those introduced to the mainland by the invaders who destroyed the mainland settlements *c.* 2200. Perhaps they came from there, but this remains uncertain.

About the same time as this invasion of the mainland, other people from Anatolia settled in some of the Cycladic islands, such as Syrus (Síros) and Ceos (Kéa) and in Euboea. On a rocky hilltop at Khalandriani in Syrus, they built an impressive fortified settlement that may have replaced an earlier native one that lay across a gorge to the south and had extensive cemeteries around it. It is clear from their pottery that these people who settled in the islands came from the coastal regions of western Anatolia, which had been comparatively civilized by then by sea contacts with Cilicia and Syria. Some of their pottery was made on a fast wheel, a useful invention that had originated many centuries earlier in Mesopotamia and began to be adopted on the mainland and in Crete at about this time.

The Middle Bronze Age on the mainland (c. 2000–1550). The mainland seems to have undergone further disturbances at the beginning of the 2nd millennium. Settlements such as those at Lerna and Eutresis were destroyed again at this time. There is some slight evidence to suggest that the destroyers on this occasion came from the north. Small handmade pots of primitive appearance with possible relatives in the Balkans were found at Lerna in houses built immediately after this second destruction. Scraps of pottery of a different kind, with string-impressed decoration, at Eutresis may come from vases brought by the people who destroyed the settlement there.

The Kurgans

Corded ware such as this was made by the Kurgan people, who had spread from their original homeland in south Russia into the Balkans. Kurgan is the Russian word for a mound and refers to their custom of burying their dead in artificial mounds or barrows. Burial mounds similar to theirs are found in many parts of the mainland, notably in the western Peloponnese but also in Attica from this time onward. The Kurgan people were pastoralists and breeders of horses, and horse bones are first attested in mainland settlements at this time. Since the Kurgan people have a good claim to be regarded as Indo-Europeans, the first people of Indo-European speech, if not, as is widely believed, the first Greeks, may have begun to settle in southern Greece at this time. As often happened in ancient times, when pastoralists overran settled agricultural communities, these invaders appear to have adopted the material culture of those they had conquered. Longhouses of the kind introduced by the invaders a couple of centuries earlier remained in use, and the custom of burial inside the settlement, which is first attested for certain now, although it may have been practiced by the earlier invaders, continued alongside burial in mounds. There is some evidence for the existence of a new class of rulers who may have been the ancestors of the princes buried in the later shaft graves and tholos tombs of the mainland. An isolated mound at Elátia (Dhrakhmáni) in central Greece, covering the richly furnished grave of a princess, appears to be an early forerunner of these. A complex of buildings around a spacious yard dominated by a longhouse in the level immediately after the destruction at Lerna may have been the "palace" of the new ruler there.

Middle Helladic culture

The development of civilization on the mainland continued to lag behind the Cyclades and Crete, and many centuries were to elapse before the art of writing spread from Crete to the mainland. But gold jewelry is found with some burials, and, by the time of the Mycenae shaft graves in the 16th century, a luxuriant style of native goldwork had been created. The people of the mainland remained in touch with the more civilized world of the Cyclades and Crete. Cretan pottery found its way to Lerna, and a fashion for decorating vases with designs in dark lustreless (mat) paint, which began on the mainland soon after the invasion, appears to have been adopted from the Cyclades. Much of the pottery on the mainland continued to be made on the fast wheel, and a class of gray burnished ware, already fashionable before the invasion, was now usually wheel made, with sharp angular shapes copied from those of metal vases. The polished gray surfaces of this "Minyan" ware (as it was named by Schliemann after the legendary inhabitants of Orchomenus in central Greece, where he first came upon it) look as if meant to imitate silver.

The Shaft Grave period on the mainland (c. 1600–1450). *The spread of Cretan influence.* On the island of Cythera (Kíthira) between western Crete and the southern tip of the Peloponnese, a colony of Cretans appears to have replaced a settlement of people from the mainland toward the end of the 3rd millennium. In the 17th or 16th century, Cretan colonies were established at Triánda in Rhodes and at Miletus on the western coast of Anatolia. Later Greek legends seem to refer to colonies from Crete, if not from Knossos, in some of the Aegean islands. Much Cretan pottery found its way to the Cyclades and was imitated there; but, although the Cycladic people adopted fashions and ideas from Crete, they retained their own distinctive traditions. Cycladic imitations of the designs on pottery from Crete tend to be less refined but more lively than their Cretan models. Wall paintings found at Phylakopi in Melos may be the work of Cretan artists, but others of about the same date from the Akrotíri settlement on Thera, in spite of conventions and techniques learned from Crete, are very different in style and spirit from Cretan ones. Settlers from Crete, however, may have made their homes in Cycladic towns such as Phylakopi, and many of the islands might have been tributary to Knossos for a time.

The mainland, like the islands of the Aegean, came increasingly under the influence of Cretan civilization, which seems to have extended to every aspect of life, including religion and dress. Its spread is vividly illustrated by the contents of the unplundered royal shaft graves at Mycenae.

The Mycenaean shaft graves. These royal shaft graves belonged to leading, if not royal, families of the place that would eventually supplant Knossos as the chief centre of the Aegean. The burials in them seem to range over a period of 100 years or more, from shortly before *c.* 1600 to the beginning of the 15th century. The princes buried here were not Cretans, although it is possible that they paid tribute to Crete. The story of Theseus suggests that Athens was tributary to Knossos at one time, but there is no hint in legend of any such payment of tribute by the rulers of Mycenae. Two groups of shaft graves have been discovered at Mycenae in different parts of a large cemetery area. Each group was eventually surrounded by a circular enclosure wall. The circle designated B with the earliest burials lay outside the limits of the later Bronze Age defenses, but the other circle A, enclosing the richest burials in six large graves, was deliberately incorporated within them. Schliemann excavated the graves of circle A in 1876, but it was not until 1951 that circle B was noted. The graves consisted of deep, rectangular shafts with stone-walled burial chambers at the

bottom. After the bodies were placed in them, the chambers were roofed with timbers, and the shafts above filled again. Both men and women were buried in the graves, many of which contained several bodies. Sometimes the remains of earlier burials seem to have been pushed aside to make room for later ones; but, if so, the shafts must have been laboriously reopened to admit new burials. Large stone slabs with carvings in flat relief had been set above some of the graves. The carvings included spiral designs and pictures of the dead riding in their chariots to war or to the hunt. According to another interpretation, they are taking part in their own funeral games with chariot races as described in Homer. These tombstones provide the earliest evidence for chariots on the mainland. With the dead and, more especially, with those in the graves of circle A had been placed a fantastic array of gold and silver cups, jewelry, and dress ornaments. Golden diadems and elaborate hairpins decked the heads of women. Beads in necklaces were of amethyst, probably from Egypt, and amber from the Baltic. The men were buried with supplies of bronze weapons, including great slashing knives and spearheads and rapier-like swords of types in use in Crete at the time. Several of the swords had blades ornamented with reliefs; others had gold-plated hilts and pommels of polished stone or ivory. Over the faces of some of the men were gold portrait masks, which show them with beards and moustaches, unlike Bronze Age Cretans but curiously like Celts and Germans of Classical times.

Earliest evidence of chariots on mainland

Some bronze dagger blades were inlaid with remarkable pictures or designs in other metals, chiefly gold and silver and electrum (a mixture of gold and silver) in various shades. Black niello was used as a background for these pictures or to heighten the incised detail. The most famous of the shaft-grave daggers shows men armed with bows, spears, and great body shields, hunting a pride of lions; another has catlike animals chasing wild fowl among papyrus flowers beside a silver stream. This technique of "painting in metals" appears to have originated in Syria, although the workmanship and style of the pictures on the Mycenae daggers look Cretan. Whereas other daggers, however, and some metal cups with inlaid designs of this kind have been found on the mainland, none has yet been recovered in Crete. But many of the treasures from the shaft graves are imports from Crete, among them a faience ritual cup such as some from the palace at Knossos; others may have been made by Cretan craftsmen working at Mycenae or by native craftsmen trained by Cretan masters. There were comparatively few clay vases in the rich graves of circle A but large numbers in those of the earlier circle B. While some of these clay vases were Cretan imports, most of them were locally made in traditional mat-painted or Minyan ware. A few, however, with designs in dark shiny paint on a light burnished surface, represent a new style that had developed on the mainland under Cretan influence. This Mycenaean pottery, very different from the mat-painted and Minyan wares it came to replace, seems to have originated in Messenia in the southwestern tip of the Peloponnese, whence it spread to other parts of the mainland and eventually as far north as Thessaly.

Mycenaean pottery

The use of Mycenaean pottery is the hallmark of the Mycenaean civilization of the Late Bronze Age. Other distinctive features of Mycenaean civilization that may have developed in Messenia under Cretan influence during the 16th century are the stone-built tholos tomb and the rock-cut chamber tomb. Cretan influence was so strong in Messenia at that time as to suggest the presence of communities of Cretan settlers there. But no positive evidence for the existence of Cretan settlements has been discovered. The tholos tomb (*tholos* is the Greek word for a round building), with high corbelled vault the shape of an old-fashioned beehive, is found in every part of the mainland to which Mycenaean civilization spread. But tholos tombs are most thickly concentrated in Messenia, and the oldest of them have been identified there. Some of the oldest of the Messenian tholos tombs resembled the early circular tombs of Crete in that they were built above ground and were used by families or clans for successive burials. But, in one important respect, they differed from the Cretan tombs. For, like all later tholos tombs, they were invariably covered with a mound. Burial mounds with many graves in them were common in Messenia throughout the Middle Bronze Age, continuing in use alongside tholos tombs in backward areas. The tholos tomb, which is in effect a circular tomb of the earlier Cretan type covered by a mound, appears to have been created therefore in Messenia either by settlers from Crete or more probably by natives of the region who had learned how to build round vaulted tombs in the Cretan manner.

Most tholos tombs on the mainland were built in a deep pit that was usually cut into the slope of a hillside. The top of the corbelled vault projected above the surface and was covered with a mound. The entrance to the circular chamber was roofed with huge slabs, and it was approached down a long passage (dromos) that was left open to the air. The most splendid of these tombs evidently belonged to kings or princes and may have been intended for only one occasion of burial. The few unplundered princely tholos tombs that have been excavated, notably one at Dhéndra (Dendra) near Mycenae, held several burials. Unless the tombs were reopened for successive burials, this would imply that wives and retainers were killed to accompany their lords and masters in death. The earliest of a group of three princely tholos tombs at Peristéria in Messenia was in use during the time of the Mycenae shaft graves. But the fashion for majestic stone-built tombs of this kind rapidly spread from Messenia to other parts of the Peloponnese, and at Mycenae they appear to have replaced shaft graves as the resting places of the local rulers about the middle of the 15th century. Nine tholos tombs have been discovered at Mycenae, and the largest and most splendid, the so-called Treasury of Atreus and Tomb of Clytemnestra, appear to be among the latest. These great tombs, built of squared masonry and richly decorated, may have been constructed early in the 14th century, although arguments for a date as late as the 13th can be advanced. The mound that covered the Tomb of Clytemnestra overlay part of the earlier shaft grave circle B, and this might seem to imply a change of dynasty at Mycenae from rulers who buried in shaft graves to ones who used tholos tombs. But the later circle of graves A was respected and eventually incorporated within the defense walls when they were built in the 13th century.

The great tombs of Mycenae

While stone-built tholos tombs became the standard resting places for kings and princes in all parts of the mainland to which the Mycenaean civilization penetrated, the mass of the population changed from a custom of burial in single graves, whether in mounds or cemeteries or inside settlements, to the use of family vaults. In some regions, such as Messenia and the frontier area of Thessaly, families built small tholos tombs for themselves. But the most common type of Mycenaean family tomb was a rock-cut chamber with a dromos leading down to the entrance. The entrance was blocked with stones and the passage filled with earth after each burial. The rock-cut tomb may have been developed in Messenia during the 16th century under Cretan influence, like the tholos tomb. In the Knossos region of Crete, rock-cut tombs had been in use for communal burials for many centuries before this. Whatever its origin, the idea of family burial in rock-cut tombs soon spread to Mycenae and other parts of the mainland. Some rock-cut tombs in Messenia and elsewhere were carved in the shape of the beehive vaults of tholos tombs. A few large rock-cut tombs, including some of this shape, were used for royal or princely burials.

Period of the Early Palaces in Crete (c. 2200–1700). Crete does not seem to have been affected by the movements of people into the Cyclades and the mainland at the end of the 3rd millennium. But important changes were taking place there. Great palaces of a distinctive type built around large rectangular open courts seem to have been constructed within a comparatively short time at the leading centres of Knossos, Phaestus, and Mallia. The art of writing is first attested for certain in Crete at

Earliest writing in Crete

the beginning of this Palatial period. These developments in Crete appear to be the result of local evolution. It has been suggested that the models for the great palaces were Anatolian, but the Anatolian palaces, like that at Beycesultan, which have been compared with them, are different in plan and may be later in date. It is always possible, however, that immigrants from some other part of the civilized world of the Near East gave an impetus to these developments, if they did not impose their rule on Crete. Another suggestion is that refugees might have reached Crete from the mainland when it was invaded *c.* 2200. Seal impressions from the House of Tiles at Lerna, which was destroyed at that time, show designs reminiscent of those on Cretan seals of the period immediately before the rise of the great palaces. If the fashion for seals with designs of this kind became established in Crete only after the destruction of Lerna, mainland refugees might have been responsible for their adoption. But the dating of events on the mainland in relation to what was happening in Crete is not yet established with certainty, and the seals that made the impressions found at Lerna may have been Cretan ones or local imitations of them.

Middle Minoan culture. At all events, Crete advanced rapidly along the path of civilization during the period of the Early Palaces, while the mainland relapsed into comparative barbarism. The art of the seal engraver made great strides in Crete at this time. Hard stones, such as jasper and rock crystal, began to be employed for some of the finer seals. A new and much favoured shape, which may have been adopted from Anatolia, was the signet with a stalk. Anatolian seals found their way to Crete and impressions of them have been identified in a great deposit of clay sealings from the early palace at Phaestus. Cretan seal designs now included elegant abstract patterns incorporating spirals and concentric circles neatly made with the drill, as well as lifelike pictures of animals, birds, and insects, together with mythical beasts, such as sphinxes and griffins adapted from Egyptian or Oriental models. Attractive hard stones, such as gabbro, were attacked by the Cretan vase makers, although they still made use of the softer chlorites and serpentines. Some of the fine stone vases from communal tombs in the Mesara region and at Mochlos may date from this period, rather than earlier, in the light of discoveries since 1950 in the early palace at Phaestus.

The fast potter's wheel began to come into use in Crete about the same time as in the Cyclades and on the mainland. Meanwhile, a revolution in the style of Cretan pottery was taking place. During the Early Bronze Age most of the finer vases everywhere in the Aegean area had been decorated with designs in dark rather shiny paint—shades of red, brown, and black—on a light surface. But, toward the end of that period in Crete, there was a change to a "light-on-dark" style of decoration in which the vases were given an overall wash of the shiny paint previously used for decoration and designs were applied to this dark surface in white. This new "light-on-dark" fashion was also adopted, to some extent, in the Cyclades and on the mainland; but in Crete it was developed much further, and, from the beginning of the Palatial period, decoration in white was regularly supplemented with red to create a striking polychrome effect. This kind of pottery, which flourished in Crete throughout the time of the first palaces and later (*c.* 2200 to 1600), is known as Kamáres ware from a sacred cave of that name on Mt. Ida, where vases with fine polychrome decoration were recovered at the end of the 19th century. Most of the smaller vases in Crete, notably the drinking cups, now copy metal ones in their shapes and often in their molded or impressed decoration, and the exquisite "eggshell" ware, made in the workshops of the great palaces, with walls as thin as those of metal vases and shiny black surfaces adorned with abstract flowerlike designs in a combination of white, red, and orange, is among the finest pottery ever produced in Greek lands. The imitations in clay suggest that vessels of precious metal—gold and silver—were in general use in the palaces of Crete by this time. A silver, two-handled goblet of this period was recovered from a tomb at Gourniá in eastern Crete, and the treasure of silver cups, more than 150 in all, buried along with foreign lapis lazuli in temple foundations at Tod in Egypt in the time of Amenemhet II (1929–1895), may well be Cretan. Silver was more valuable than gold in Egypt at this time. It occurs in the Cyclades, and it was being mined before the end of the Bronze Age near Laurium in Attica on the mainland.

"Eggshell" ware

Commerce. There were many contacts between Crete and the rest of the Levant during this period. Scarabs and stone vessels from Egypt reached Crete and were imitated there. Cretan Kamáres ware was exported to Cyprus, Syria, and Egypt, where it has been found in tombs and on town sites. Letters recovered from the ruins of the city of Mari on the Euphrates, destroyed by Hammurabi *c.* 1760, refer to objects of Cretan workmanship. It seems that Cretan metalworkers were already pre-eminent in the civilized world of the time. The daggers they made were of types ultimately derived from Syria, but they were exported to Cyprus in exchange, perhaps, for copper, although supplies existed in Crete. Westward, they may have reached Italy, where native copper daggers are of Cretan shapes and flint imitations of them seem to have been made. It was during this period that tin-bronze began to come into more general use in the Aegean, replacing copper or bronze made by adding arsenic, an effective process but dangerous for the craftsman who undertook it. Tin may have reached the Aegean first from Iran through Syria, although Etruria on the western coast of Italy was another possible source.

Burial in Crete was still normally in communal tombs, and many of the Early Bronze Age ones continued in use; but cemeteries of burials in storage jars are also attested at this time. No royal tombs of this period have been identified, however, and kings and queens may have been laid to rest, like their subjects, in the tombs of their clans. A large rectangular building with many rooms or compartments in the cemetery area just outside the city at Mallia might have been the tomb of the royal clan there. The local inhabitants plundered it during the 19th century, and its modern name—Chrysolakkos ("Gold Hole")—suggests what they found. A gold cup and jewelry, including elaborate earrings and pendants, acquired by the British Museum in 1892 and allegedly from a Mycenaean tomb on the island of Aegina near Athens, appears to be plunder from Chrysolakkos. French excavations there in the 1920s led to the recovery of similar jewelry, notably a gold dress-pin with flower head and a pendant in the form of a pair of bees (or wasps) facing each other over a disc, which may be meant for a honey cake. This pendant shows that the Cretan jewellers were masters of the art of hard soldering and could use it to fix wire (filigree) or minute globules of gold (granulation) to a background. But the jewelry from Chrysolakkos is difficult to date, and it may belong to a slightly later period, after the destruction of the early palaces at Knossos and Phaestus.

The Chrysolakkos treasure

Middle Cycladic culture. Life in the Cyclades seems to have continued much as it had in the Early Bronze Age. But, apart from signs scratched or painted on pottery from Phylakopi in Melos, there is little evidence of acquaintance with writing or the use of seals. Some time after the beginning of the period of the Early Palaces in Crete, Phylakopi was defended by a massive wall. Cretan Kamáres ware and Minyan ware from the mainland were imported to Phylakopi and doubtless reflect the trade in Melian obsidian, which may have been still in demand for cheap knives and razors, although metal ones were already in use in the Aegean area from the Early Bronze Age onward. Chamber tombs cut in the rock at Phylakopi appear to go back to this period, but burial in slab-lined cists continued elsewhere in the islands. At some point the fortified settlement at Khalandriani on Syrus was destroyed by fire and abandoned.

Period of the Later Palaces in Crete (c. 1700–1450). Various disasters occurred in Crete about the turn of the 18th and 17th centuries BC. The palaces at Knossos and Mallia were damaged, while that at Phaestus and a building that may have been the residence of a local ruler in a large settlement at Monastiráki west of Mt. Ida were de-

Destruction of the palaces

stroyed by fire. What caused these destructions is uncertain. Accident, internal warfare, or foreign invasion are among possible agents. The damage at Knossos might have been caused by one of the earthquakes that afflict that area. It has been suggested that Crete was first conquered by Greeks at this period or by people from Anatolia speaking another Indo-European language called Luwian and related to Hittite. But there is no strong evidence for an invasion of Crete at this time.

The flowering of Minoan culture. The two or three centuries following these disasters were indeed the most flourishing of the Aegean Bronze Age, during which Cretan civilization reached its zenith. The palaces at Knossos, Phaestus, and Mallia were restored with greater splendour than before. From the dimensions of the new and entirely rebuilt palace at Phaestus, it has been possible to calculate the unit of length used by the Cretan architects: a foot only a fraction shorter than the standard English foot. In plan, the later palaces were basically the same as the earlier ones, with agglomerations of rooms clustering around long, rectangular central courts oriented roughly from north to south either for ritual or for catching the best of the winter sun. Many parts of these palaces were two or three stories high. A section on the eastern side of that at Knossos, built into a cutting in a steep slope below the level of the central court and housing the royal living quarters, may have had five stories. Large areas of the palaces, especially at Knossos, were reserved for cult, while other parts were occupied by storerooms for the produce brought as taxes in an age before the existence of money. Wide, paved squares flanked the palaces, and around them spread extensive towns, which, by this time, if not earlier, seem to have been unwalled. Cobbled streets with raised central paths of smooth squared blocks for the convenience of pedestrians ran through the towns. Surface water was carried away by covered drains, and skillfully jointed clay waterpipes were found in the palace at Knossos. The only settlement of this period that has been entirely excavated is a small town at Gourniá in eastern Crete. This was built on the slopes of a ridge overlooking the sea on top of which stood a little palace with a small open court in the centre and a public square beside it on the sheltered landward side. Down the ridge from the palace toward the sea was a small shrine facing the end of a path that led to it from the main street. Even in a small town such as that at Gourniá, many of the houses were evidently two stories high, and houses with three stories are depicted on faience inlays from the palace at Knossos assignable to the 17th century BC.

In palaces as well as houses, the lower parts of walls were still normally built of rough fieldstones held together with mud, the upper stories being continued in mud brick. But carefully squared and fitted blocks of limestone were employed for some important facades. Now, as earlier, walls were often laced together with a framework of timbers set vertically and horizontally in their faces and joined by crossbeams running through them. There was also an extensive use of timber for columns and pillars and for the rafters supporting upper floors and roofs, which, it seems, were usually flat. Pictures of wooden columns show them with a characteristic downward taper, which may reflect an original custom of placing tree trunks upside down. The lower parts of the walls inside the palaces and great houses were often clothed with large slabs of attractively veined gypsum, a soft crystalline stone that outcrops in the region of Knossos and Phaestus. Gypsum was also much employed for pavements, but a hard lime plaster was more commonly used for coating walls and floors. Plastered walls were decorated with brightly coloured pictures, which may be an innovation of this period, since they are not yet attested for certain earlier in Crete. These pictures are described as frescoes because they were normally painted while the plaster was still damp. Lines impressed with string in the wet plaster helped to guide the artists. White, red brown, or blue were usually chosen as a background, while yellow and black were among other basic colours used. Many of these pictures, especially those from the palace at Knossos, were concerned with religion and show elaborately dressed goddesses, together with sacred dances and ceremonies, such as bull leaping, which may have had a religious or magical basis. But scenes such as a frieze of partridges and hoopoes adorning a room in what appears to have been an inn for strangers opposite the palace at Knossos look entirely secular. Monkeys, imported from Egypt, are depicted more than once, along with native wild goats and extraordinarily lifelike flowers, sometimes of identifiable species but often imaginary. The pictures ranged in scale from those with life-size figures, which might occupy most of the wall surface, to panels and friezes, including a class of miniatures with little figures of men and women two or three inches high. Parts of some wall pictures at Knossos were in relief, and plaster reliefs of this kind are occasionally found elsewhere in Crete. Floors and ceilings might also carry painted decoration.

Minoan frescoes

Their wall paintings were probably the finest achievements of the Cretan artists, but only battered or fire-discoloured fragments of these have survived. The minor arts are better represented in the archaeological record. Now, if not earlier, hard rock crystal began to be used for making vases and seals, together with the volcanic glass, obsidian. A variety of this, flecked with spots of white pumice from Yialí (Glass Island), near Cos (Kos), was favoured for vases. Other fine stones imported for vase manufacture were Egyptian alabaster (calcite) and green and red marbles (antico rosso and lapis lacedaemonius) from the southern Peloponnese. Antique stone vases from Egypt might be adapted to local tastes by the addition of spouts and handles. Vessels with narrow necks were carved in two pieces that were afterward joined together, an example being a crystal libation vase from Zákros (Zakro) with the handle formed of crystal beads threaded on copper wire. A number of cult vases are carved with pictures in relief, including an octopus, a mountain shrine, a ritual dance of some kind, and games, such as bull leaping, wrestling, and boxing, which appear to have had magical or religious connotations. Soft stones, such as chlorite or serpentine, were used for making these vases, the surfaces of which were often coated with gold leaf to judge from scraps of it that have survived. This economical system of gilding was sometimes applied to seal stones, although solid gold and silver seals also occur. A class of gold signet rings has oval bezels engraved with ritual scenes that may include episodes from the story of a goddess and her consort, with his death and resurrection. Seals of other shapes, in a wide range of attractive stones, display a variety of designs, including pictures of animals, such as lions, bulls, and wild boars. Sometimes a bull is being attacked by a lion, or a wild goat is escaping or standing at bay before a hound. Birds, fish, and butterflies also figure on these seals, and most of the designs on them appear to be entirely secular in character. But a class of gems crudely engraved with pictures of jars and leafy branches may have been rain-charms. Their skill in carving hard stones for seals and vases and their ability to model large plaster reliefs suggest that the Cretans might have sculptured stone reliefs and statues had they wished. But there is little evidence for Bronze Age sculpture in Crete, apart from a few small stone heads that may have come from statues with wooden bodies. Some bronze curls from the palace at Knossos appear to have adorned the head of a more than life-size wooden statue of a goddess. Figurines cast in solid bronze and mostly representing worshippers, both men and women, placed as votives in sanctuaries are often of great beauty. Statuettes of bull leapers and perhaps of gods and goddesses were made of imported ivory in several pieces cunningly joined together by means of pins and dowels. The technique of faience manufacture was presumably learned from Egypt. Exquisite faience plaques of animals, along with statuettes of goddesses or priestesses and small vases of the same material, appear to be products of the palace workshops at Knossos.

Rings and seals

The art of the potter, however, was in eclipse in Crete for a time, which may reflect the widespread use of gold and silver plate for tableware. Even cooking vessels were

now being made of copper or bronze, including huge cauldrons in which a sheep or goat could be boiled whole. Among a variety of serviceable bronze tools were axes, adzes, and double-bladed axes such as those of earlier times. The sockets of these were improved toward the end of the period from a circular to an oval shape, which prevented twisting of the haft. New tools current by then included long bronze chisels and immense saws capable of slicing the gypsum required for paving and wall veneer, as well as for cutting timber. During this period, if not earlier, little trowel-like blades were replaced by leaf-shaped ones as razors. Helmets of copper or bronze are depicted on faience inlays from Knossos and stone relief vases, but plate armour is attested only from the end of the 15th century onward. For defense the Cretans of this time appear to have relied on huge rectangular or eight-shaped shields of bull's hide. (Homer's reference to the shield of Ajax "like a tower" preserves a memory of body shields of this kind.) Weapons included spears and daggers, as well as rapiers with long slender blades and short tangs for affixing wooden hilts. Massive pommels of attractive stones, such as rock crystal, or of gold-plated wood or ivory helped to balance the blades. Toward the end of the period, swords are found with strong, flanged hilts and short blades adapted for cutting as well as thrusting strokes.

Evidence of writing. Signs scratched or painted on clay vases, not only in Crete but on the mainland and in the islands, from about the middle of the 3rd millennium onward may reflect some acquaintance with writing among the peoples of the Aegean area then. But the first positive evidence for the use of writing in the Aegean is found in Crete at the beginning of the Palatial period—*c.* 2200 or somewhat later. This earliest Cretan writing is known as pictographic or hieroglyphic because its signs are pictures of animals or things in the manner of Egyptian hieroglyphs. But only a few of the signs resemble Egyptian ones, and the system appears to be of Cretan origin, even if the idea of writing was inspired by Egypt or Syria. During the period of the Early Palaces and while the Cretan hieroglyphic script was still in use, a simplified linear script was being scratched on clay tablets at Phaestus. A more evolved script with linear signs of this kind is attested in various parts of Crete during the period of the Later Palaces and is known as Linear A, to distinguish it from the variety of script (Linear B) current both in Crete and on the mainland from the end of the 15th century onward (see below). Most of what has survived of Aegean Bronze Age writing is on clay tablets of the kind used in Syria and Mesopotamia in early times. But ink was used to write Linear A inscriptions around the insides of two clay cups from Knossos, and the bulk of what was written in the Aegean during the Bronze Age may have been in ink on some kind of paper, made from papyrus as in Egypt or from palm leaves, as later Greek tradition hints. Palm trees grew in Crete during the Bronze Age and some of the later (Linear B) clay tablets there and on the mainland seem to have been made in the shape of palm leaves.

Linear A and Linear B

II. The decline of the early Aegean civilizations

THE ERUPTION OF THERA (C. 1500) AND THE CONQUEST OF CRETE (C. 1450)

Cretan civilization reached its highest peak in the 16th century. An important change of fashion that began in Crete then was the abandonment of the "light-on-dark" style of vase decoration of Kamáres tradition in favour of a return to "dark on light." The new-style Cretan pottery, with attractive designs of plants and flowers in shiny black or brown paint, was soon to inspire the development of Mycenaean pottery on the mainland. But this flourishing period in Crete ended in a series of disasters. About 1500 the volcano on the island of Thera, long, it seems, quiescent, erupted to bury the settlements there under many feet of pumice ash. The story of Atlantis, if Plato did not invent it, may reflect some Egyptian record of this eruption, one of the most stupendous of historical times. Knossos was shattered by a succession of earthquakes that preceded or accompanied the eruption, while great waves resulting from it appear to have damaged settlements along the northern coast of Crete. Later Greek traditions, such as the story of Deucalion's flood, may enshrine a memory of similar waves that swept the coasts of the mainland at this time. Some Cretan settlements might have been wrecked by blast from the eruption, although Thera lies about 70 miles (110 kilometres) away from Crete. But, whatever the damage caused, it appears to have been soon repaired. Similarly, in the Cyclades there are few signs of any gap in occupation as a result of the eruption. The so-called Second City at Phylakopi in Melos was totally destroyed, but the surviving inhabitants appear to have set to work at once to build the Third City on its ruins. The settlements on Thera, however, lay buried deep in pumice, and the largest of these, at Akrotíri, opened by excavations since 1967, offers a unique picture of a Bronze Age town, the walls of the houses standing in places two stories high, with paintings miraculously preserved on them, and the floors with storage jars and other objects resting on them as they were left when the inhabitants escaped from the eruption or from the earthquake that is thought to have preceded it.

After the eruption, Crete appears to have enjoyed comparative prosperity for a time, while the influence of Cretan civilization continued to spread on the mainland. Alongside vases with plant and flower designs, the Cretan potters began to make ones decorated in an attractive Marine Style, with octopuses and other sea creatures. The Marine Style may have originated at Knossos, but vases with this type of decoration, many of them of a ritual character, were exported all over Crete, as well as to the Cyclades and the mainland. About the middle of the 15th century, however, a generation or so after the eruption of Thera, most of the important sites in central and southern Crete were destroyed by fire. Whether this horizon of fire destruction can be traced in western Crete is uncertain because of the lack of excavation there. Destruction was not confined to palaces and towns but extended to country houses, farms, and rural shrines. Many settlements were never inhabited again, such as that at Mochlos, where the excavators found the remains of numbers of people who had perished in the destruction. The site of the destroyed town at Gourniá was eventually occupied by a scatter of houses, but the palace there was not rebuilt. The large palaces at Mallia and Zákros were also destroyed by fire and afterward abandoned.

The Marine Style

A new social order. The fact that palaces and country houses, centres of landed estates, were not rebuilt suggests a total overthrow of the existing social order. A number of magnificent stone ritual vases and bronze tools have been recovered from the ruins of the palace at Zákros in excavations since 1962, but virtually no gold or silver objects were found. Indeed, it looks as if the palaces and houses everywhere in Crete had been ransacked before they were destroyed. Alone of the four great palaces, that at Knossos may have escaped serious damage at this time, but parts of the city there were wasted by fire. In the early days of Cretan exploration, it was taken for granted that these destructions were the result of war. Since the 1930s, however, it has been suggested that they were, in some way, caused by the eruption of Thera. But the eruption began and ended a generation or more before this horizon of destruction, while conquerors are found in Crete immediately after it. The destruction appears to have been their work.

They evidently came from the mainland and made their capital at Knossos, but they seem to have established another centre of power at Phaestus; and legend hints at a third based on Kydonia (modern Khaniá) in western Crete. Vases of shapes already popular on the mainland, such as drinking cups with high feet, became fashionable at Knossos after the conquest and eventually spread to other parts of the island. A rather stiff, formal "Palace Style" of vase decoration, using motifs derived from the earlier Plant and Marine styles, may reflect an adaptation of Cretan fashions to cruder mainland tastes. The old clan tombs went out of use in the Knossos region and were replaced by rock-cut tombs of the type evolved un-

der Cretan influence on the mainland. Some of these contain the burials of warriors and their families, accompanied by rich assortments of weapons and jewelry, resembling the Mycenae shaft graves and the mainland tholos tombs. There was a cemetery of similar rock-cut tombs, with richly furnished burials (the Tombe dei Nobili), at Phaestus. Tholos tombs sunk in the ground and covered by mounds appear to have been introduced to Crete from the mainland now. One, on the Kefála ridge north of the palace at Knossos, may have been built soon after the conquest.

A decline in taste and artistic imagination becomes marked during the period after it throughout the Aegean area. But skills and techniques remain at a high level, and the fabric and firing of pottery improve both in Crete and on the mainland. Great advances were made in military equipment, to judge from the formidable and often elegant and richly ornamented bronze weapons from the warrior graves of the time. Short, strong swords had blades adapted for cutting as well as thrusting. Their hilts were sometimes sheathed in gold, and pommels were made of fine stones or ivory, as in the past. Warriors were armed with large thrusting spears as well as throwing spears or javelins. Bronze arrowheads, virtually unknown in Crete before the conquest, are of types introduced from the east or copied from the flint and obsidian arrowheads long current on the mainland. Conical helmets of felt or leather might be cased in bronze, such as one from a warrior grave of *c.* 1400 at Knossos, which was surmounted by a knob for holding a plume and had separate cheekpieces. An old mainland custom of plating helmets with the tusks of wild boar is first attested in Crete during this period. Ivory plaques from the mainland show men wearing such helmets, and Homer describes one in detail. One placed in a tomb at Dhéndra, near Mycenae, *c.* 1400 had separate bronze cheekpieces, and the warrior to whom it belonged also had a pair of bronze greaves (leg guards) and a corslet of bronze plates cunningly fitted together. Similar corslets are depicted on Linear B clay tablets from Knossos as part of the issue to chariot warriors. Plate armour of this kind may have been developed in connection with chariot warfare for which the old body shield was impractical. Small, round shields also began to come into use in the Aegean area at about this time. Some Aegean ships by now had railed platforms at prow and stern joined by a gangway down the length above the heads of the rowers. Before the end of the Bronze Age, ships of this type had been developed as warships with rams.

Advances in weaponry

Events on the mainland. While there are many signs of mainland influence in Crete in the period after it, the conquest may have helped to spread Cretan fashions and techniques on the mainland through the medium of captive artisans sold as slaves. The earliest wall paintings on the mainland appear to date from this time and are thoroughly Cretan in style. But the Mycenaean civilization of the mainland remained very different from that of Crete. Mycenaean pottery is distinguishable from Cretan, and religious customs, such as worship in caves or hilltop sanctuaries, which continued in Crete, do not appear to have taken root on the mainland. The standard mainland palace of this period, although built with Cretan techniques and adorned with wall paintings by Cretan artists or in the Cretan manner, differed from the traditional Cretan palace centred around a large, rectangular court. The focal point of a mainland palace, such as that at Pylos (Pílos) in Messenia, was a great hall with the roof supported on four pillars and a vast circular central hearth. The hall was entered through an anteroom with a columned porch beyond it. This complex appears to be an adaptation of the type of longhouse found on the mainland since the end of the 3rd millennium. The descriptions of palaces in Homer are evidently based on memories of palaces such as these, and what Homer calls the megaron corresponds to the great hall.

Mainland palaces

Conquest of the Cyclades. People from the mainland overran the Cyclades when Crete was conquered or a few years later. Settlements such as that at Ayía Iríni on Ceos (Kéa) were destroyed by fire, and the old Cretan colony on Cythera was abandoned. The so-called Third City of Phylakopi on Melos suffered damage, and, immediately afterward, a small palace of mainland type was built there. Similar palaces may have been built in Crete after the conquest. The largest house of this period at Gourniá is very similar in plan to the palace at Phylakopi. Massive foundations at Ayía Triádha near Phaestus may have belonged to the hall of a palace of orthodox mainland type. But the great palace at Knossos retained its traditional Cretan form, in spite of many alterations that were made to it. The paintings on its walls in the period after the conquest were still largely of a religious nature and included scenes of bull leaping and friezes of people taking part in processions or ritual. These friezes were inspired by Egyptian models, although their monumental character may have appealed to mainland tastes. The rulers of the mainland also adorned their palaces with religious pictures of this kind, but they appear to have indulged a passion for scenes of warfare and hunting to a degree that is not attested at Knossos.

Linear A was the script in use in Crete at the time of the conquest. But this had been replaced by Linear B by the end of the 14th century. Linear B clay tablets have been recovered from Knossos and from several of the palaces on the mainland. The people of the Cyclades may have been acquainted with the Cretan Linear A script, to judge from signs of it on local vases from settlements on Thera buried by the eruption of *c.* 1500, but it is doubtful if a version of this script was used on the mainland. If it was, the people of the mainland might have adapted it to form the Linear B script for writing their own language before they conquered Crete. Alternatively, Linear B may have been evolved afterward at Knossos. The Linear B tablets consist mostly of short lists or inventories, and picture signs (ideograms) often give an indication of what is listed on them. Thus, some tablets from Knossos, with a chariot, horse's head, and corslet, evidently record equipment issued to warriors. Ever since the first Linear B tablets were discovered at Knossos in 1900, many people have thought that their language might be Greek, and it was deciphered as Greek by Michael Ventris in 1952. His decipherment has been widely accepted, but, with short inscriptions of the kind available, it is difficult to verify. Those who doubt his decipherment include some who agree that the language of the Linear B tablets is Greek, whereas others think it is a non-Greek language spoken by the conquerors of Crete or a form of the old Cretan language that they adopted for official purposes at Knossos and on the mainland.

The fall of Knossos and the Mycenaean empire. The new rulers of Crete evidently thrived for several generations after the conquest. Great tombs were built and richly furnished. The unplundered burial of a princess of this period has been found at Arkhanes near Knossos. Then, suddenly, this comparative prosperity was brought to an end. The palace at Knossos was destroyed by fire and never rebuilt. Since 1960 the date of this final destruction of the Knossian palace has been disputed with a range of opinion from *c.* 1400 to 1150, but it probably took place in the 14th century, when there is evidence for destruction at other Cretan sites, such as Mallia, and when burial in the warrior graves of the Knossos region and at Phaestus ceases. The elimination of the warrior aristocracy that had flourished in Crete since *c.* 1450 was surely the work of enemies, possibly from the mainland. But the mainland also appears to have undergone changes for the worse about this time. The era of great princely tholos tombs on the mainland seems to come to an end now or not much later. From the beginning of the 14th century, the arts of wall painting, seal engraving, and vase decoration everywhere show symptoms of stagnation and decline. The pottery of the 13th century is remarkably standardized and homogeneous throughout the Aegean area. After the fall of Knossos, the chief centre of power in the Aegean appears to be Mycenae, which may have exercised control over an empire including Crete and the islands. From *c.* 1400 or earlier, the records of the Hittite Empire in Anatolia mention a state called Ahhiyawa, which may be some Mycenaean dependency,

Final destruction of the palace at Knossos

if not an empire ruled by Mycenae. Ahhiyawa could be the same word as Achaioi (Achaeans), the name given to the Greeks in Homer.

THE END OF THE BRONZE AGE IN THE AEGEAN (C. 1200–1000)

There is no evidence that any palaces existed in Crete after the final destruction of that at Knossos; but, on the mainland, palaces continued to flourish at centres such as Mycenae, Tiryns, and Pylos. In the 13th century massive defense walls were built at many of these mainland centres, including Mycenae and Tiryns and Athens, where the acropolis was fortified. At Athens and Mycenae underground passages led to springs concealed beyond the walls. The Lion Gate at Mycenae dates from this time, although the limestone relief of confronting lions above it may have been carved earlier. A wall appears to have been constructed during this period across the Isthmus of Corinth against enemies from the north.

Toward the end of the 13th century, the chief centres of the Peloponnese were destroyed by fire, and their palaces were never rebuilt. Centres farther north, such as Thebes and Gla, may have been destroyed some time before this. The site of Pylos was afterward deserted, but occupation of a kind continued at Mycenae. This destruction of Mycenaean centres on the mainland seems to herald a collapse of civilized life throughout the Aegean area. Wall painting, gem engraving, and writing vanish from the scene; but a more efficient type of slashing sword appears in the Aegean about this time, and the skill to make complicated military equipment, such as chariots and warships with rams, was not lost, while eventually iron replaced bronze for weapons and tools, as elsewhere in the Near East. After the destruction of the Mycenaean centres in the Peloponnese, there is evidence for depopulation there and for a great exodus of the former inhabitants, some westward and southward to Cephallenia and Crete, others eastward to the Cyclades and as far as Cyprus and Cilicia, where their settlements have been noted. An important cemetery of this period has been excavated at Peráti on the eastern coast of Attica.

Introduction of iron

There has been much speculation as to the causes of these disasters and upheavals. Revolutions or civil wars and vast piratical raids by land or sea have been blamed for the destruction of the Mycenaean centres. The dispersal of refugees from the Peloponnese has been explained in terms of climatic changes leading to drought and famine there. But the evidence for such changes is doubtful, and the refugee movements look very much like flight before the onslaught of invaders.

Invasion from the north. There are indeed signs that barbarous peoples from the north had begun to enter the Peloponnese by the 12th century. Traces of them are slight and controversial, because they adopted elements of the Mycenaean civilization, and most of their pottery was made on the fast wheel in a version of Mycenaean style and, at first, presumably by Mycenaean craftsmen whom they had enslaved. Burial in single graves, current on the mainland before the rise of the Mycenaean civilization, had lingered in regions beyond its frontiers, such as Epirus and northern Thessaly. Now, single graves, which may be those of invaders, are found in the Peloponnese and elsewhere on the mainland alongside family tombs of Mycenaean tradition. Pairs of long dress pins, with burials of women, are among northern features that appear in the Mycenaean area at this time. These pins, worn one on each shoulder, suggest the introduction of a new type of garment, different from anything Mycenaean and akin to the simple blanket-like peplos still retained by the womenfolk of conservative Dorian Greeks in Classical times.

The coming of the Greeks. The Classical Greek dialects fall into an eastern and a western group, which might reflect two successive movements of people of Greek speech into Greece. According to tradition, the Dorians, or West Greeks, only entered the Peloponnese in the 11th century. It is therefore difficult, although not impossible, to regard them as the destroyers of the Mycenaean centres there *c.* 1200. As to when the bearers of the East Greek dialects first appeared in Greece, there is a wide range of opinion. At one extreme is the idea that the Neolithic inhabitants of the Aegean already spoke a form of Greek. According to another view, both groups of Greek dialects were actually formed on Greek soil early in the Bronze Age, the East Greeks being the creators of the Mycenaean civilization, while the West Greeks remained in primitive conditions beyond its frontiers until the 12th century or later. One widely held view brings the ancestors of those who spoke the East Greek dialects from Anatolia or the north in the centuries around 2000, when there is evidence of successive invasions on the mainland. At the other extreme is the belief that Greek was not spoken in the Mycenaean area until *c.* 1200, when the bearers of the East Greek dialects came as the destroyers of the Mycenaean centres in the Peloponnese and adopted a barbarized version of the Mycenaean civilization as their own. This view, however, is naturally difficult to reconcile with a belief that the language of the Linear B tablets is Greek.

Theories of the origin of Greek dialects

THE PEOPLE OF THE BRONZE AGE AEGEAN

Physical appearance. The inhabitants of the Aegean area in the Bronze Age may have been much like many people in the Mediterranean basin today, short and slight of build with dark hair and eyes and sallow complexions. Skeletons show that the population of the Aegean was already mixed by Neolithic times, and various facial types, some with delicate features and pointed noses, others pug-nosed, almost negroid, are depicted in wall paintings from the 16th century. But men and women are always represented with black hair, and the presence of fair-haired people is not attested in the Aegean until later Greek times. Some very tall men buried in the Mycenae shaft graves may be descendants of invaders who entered the mainland at the end of the 3rd millennium. A few skeletons from the single graves that appear on the mainland at the very end of the Bronze Age suggest the presence of new people from the north.

Dress. Clay figurines of *c.* 2000 from Crete show men wearing a narrow codpiece with a belt or loincloth and bare above the waist. This was to remain the basic fashion for Cretan men throughout the Bronze Age. Cretan women wore short-sleeved jackets, which left the breasts bare, and ankle-length flounced skirts, although shorter skirts to just below the knees are also attested. Marble figurines of men from the Cyclades assigned to the Early Bronze Age have belts and narrow codpieces such as the Cretans. There is little evidence for dress on the mainland until the time of the Mycenae shaft graves in the 16th century. A considerable variety of dress is represented from that time onward throughout the Aegean area, but it is difficult to recognize fashions peculiar to the mainland. Tasselled shorts worn by men on the Mycenae lion-hunt dagger are also attested in Crete at the time. Aegean envoys painted on the walls of the tomb of Rekhmire, vizier of the Egyptian pharaoh Thutmose III (ruled 1504–1450), at first wore large codpieces of a type fashionable in Crete in the Shaft Grave period, but these were later changed into kilts without codpieces as worn by men in paintings at Knossos after the mainland conquest of Crete, *c.* 1450. This might reflect mainland envoys coming to Egypt after the conquest and wearing a different type of dress from the Cretan. But kilts of this kind appear to be represented in Crete both before and after the conquest. Curious scaly cloaks and long single-piece robes are among a variety of ritual garments in Crete. Linen was known in Crete by the beginning of the Bronze Age and fragments of it were recovered from the Mycenae shaft graves. But in Crete, at any rate, clothes were mostly, it seems, made of wool, and wall paintings show them woven with colourful and intricate designs, including pictures of animals and birds and even musical instruments. One of the dyes used was purple crushed from murex shells. Cretan men wore knee boots and sandals with upturned toes. Men with leggings or greaves are represented in wall paintings on the mainland. Caps of various kinds appear on the heads of men and high, pointed hats and tiaras on those of women and of gods and god-

desses or their priests. Clay figurines of the Early Palace period show Cretan women with elaborate hair arrangements. Women put jewelry in their hair, including strings of beads. Necklaces, earrings, bracelets, and armlets were displayed by men as well as women. Seal stones were carried on strings around the neck or on the wrist. Cretan men normally left their hair long but were clean shaven. Beards and moustaches are attested on the mainland in the Shaft Grave period and later.

Society. In the 3rd millennium the whole Aegean area appears to have been divided into small units ruled by chiefs with "palaces" not much larger than houses. But the Early Palaces in Crete suggest larger kingdoms there by *c.* 2000, with a stratified society including nobles and probably classes of scribes and free artisans, as well as slaves. In the period of the Later Palaces, if not earlier, every sizable settlement in Crete was dominated by a large house, or palace, evidently belonging to the chief family of the place. These families may have formed a nobility owing allegiance to the kings of the great palaces at Knossos, Phaestus, Mallia, and Zákros. The kings may have shared power with councils of nobles, if not with popular assemblies, but this is speculative and based on doubtful analogies with early Sumeria. Primitive customs of the kind loosely called matriarchal, such as inheritance through the female line, may have survived in Crete, and women evidently played an important part in religious cult there. The Mycenaean society of the mainland appears to have had an aristocracy of conquest. Toward the end of the Bronze Age, the evidence suggests the formation of larger states extracting tribute from local rulers and eventually perhaps suppressing them to exercise direct control.

Social organization

Trade. Foreign manufactures reaching the Aegean and especially Crete during the Bronze Age included Cypriot pottery, Mesopotamian and other Oriental cylinder seals, and Egyptian stone vases, ivories, and scarabs, while Cretan and eventually Mycenaean pottery is found in Egypt and elsewhere in the Levant. But the bulk of what was traded evidently consisted of perishable fabrics and foodstuffs and, above all, raw materials. Obsidian from Melos reached other parts of the Aegean from Mesolithic times onward. Imported tin became a necessity during the Bronze Age, and copper was brought from Cyprus to supplement native sources. Massive copper or bronze ingots of an international type are found in Crete and elsewhere in the Aegean from the 15th century onward. Ostrich eggs and stone for making vases were among imports to Crete from Egypt, and ivory came from there or from Syria. Amber from the Baltic reached the mainland in some quantity during the Shaft Grave period and later but is rarely found in Crete. Exports from the Aegean may have included woollen goods, olive oil, and timber, as well as silver. In Crete, at any rate, foreign trade may have been largely under palace control, but a class of private merchants engaged in overseas commerce no doubt existed in the Aegean.

Transport. Ships with a mast and square sail in addition to oars or paddles are attested in the Aegean from the Early Bronze Age. On land, goods were no doubt transported by pack animals or on poles slung between bearers, a principle also adopted for passenger chairs, of which there are clay models. A model of a four-wheeled cart from Crete is datable to *c.* 2000 or earlier. The wheels of such carts were evidently solid, and they were no doubt drawn by oxen, although horses may have been ridden in Crete by then, as they seem to be depicted on early Cretan seals. These horses could have come from the east, a different breed being introduced into the mainland from the north at the beginning of the Middle Bronze Age *c.* 2000. The light spoke-wheeled chariot drawn by horses appears to have developed in Syria or northern Mesopotamia early in the 2nd millennium but spread rapidly throughout the Near East, owing to its usefulness in war. Chariots are depicted on tombstones of the Mycenae shaft graves and on Cretan seals before the time of the mainland conquest *c.* 1450. Apart from warfare, they were used in the Aegean for hunting and probably for travel. During the later part of the Bronze Age, terraces were built to support roads wide enough for wheeled vehicles both in Crete and on the mainland. Such roads were carried across streams on bridges, examples of which have survived in the region of Mycenae.

Warfare. Short daggers of types derived from Syria were in use in the Aegean during the Bronze Age. Long rapiers, evolved from those in Crete, are found on the mainland by the time of the Mycenae shaft graves in the 16th century BC. Short swords adapted for cutting as well as thrusting begin to appear in the following century and may have been developed in connection with chariot warfare. Bronze armour and small, round shields more serviceable in chariots replace the old Cretan body shields at approximately the same time. Bows and slings were probably used everywhere in the Aegean area, but, whereas arrowheads of flint and obsidian are found on the mainland, they are virtually unknown in Crete, where arrows may have been tipped with bone or wood until the appearance of bronze arrowheads in the 15th century. Settlements on the mainland and in the Cyclades were defended by walls from the Early Bronze Age onward, and the town at Mallia in Crete appears to have been protected by a wall during the period of the Early Palaces; but, by the time of the Later Palaces, Cretan towns may have been unwalled. Faience inlays of the 17th century from Knossos, however, seem to show an attack on a walled town such as that depicted on a silver-relief vase from the Mycenae shaft graves. The town, in each case, appears to be in the Aegean but not necessarily in Crete. Methods of warfare had become highly developed by the end of the Bronze Age, with improved weapons, complex and well-designed fortifications, extensive use of chariots, and warships with rams.

Development of fortified towns

Religion. Little is known about religion in the Cyclades and on the mainland before the period when they came under strong Cretan influence. An open-air sanctuary filled with marble figurines on the island of Kéros (Káros) is assignable to the Early Bronze Age. In Crete during the Early Palace period, there were many open-air sanctuaries on the tops of hills and mountains. Some of these had small shrines in them, and shrines with one or more rooms and benches for offerings and cult statues are found in the countryside and in the towns in Crete. Parts of the palaces and of large houses there were also set apart for cult. Shrines not unlike Cretan ones existed in settlements in the Cyclades and on the mainland in the Late Bronze Age; but hilltop sanctuaries are not well attested there, and most of those in Crete appear to go out of use after the mainland conquest, *c.* 1450. Caves were also used as sanctuaries in Crete, and cults in some of these continued until the end of the Bronze Age and later.

The chief deity everywhere in the Aegean in the Bronze Age was evidently a goddess. Perhaps there were several goddesses with different names and attributes. Clay statues of goddesses have been found in shrines in Crete dating from the 15th century. Shrines with clay goddess figures have also come to light at Ayía Iríni in Ceos and at Mycenae. But most cult statues were probably made of wood everywhere in the Aegean throughout the Bronze Age. There is evidence for the worship of a male god, who may have been a counterpart of gods such as the resurrected Osiris in Egypt. Much of Aegean Bronze Age religion survived into later Greek times, and the idea of an Elysion, or Isles of the Blessed, very different from the gloomy Hades of Homer, appears to be a legacy from Bronze Age Crete.

BIBLIOGRAPHY. For recent work see the accounts that appear each year in *The American Journal of Archaeology, Archaeological Reports* (published by the Society for the Promotion of Hellenic Studies and the British School of Archaeology at Athens), *Bulletin de Correspondance Hellénique*, and Greek periodicals like *Arkhaiologikon Deltion.*

General works: Cambridge Ancient History, 3rd ed., vol. 1–2 (1970–72); REYNOLD A. HIGGINS, *Minoan and Mycenaean Art* (1967); SINCLAIR HOOD, *The Home of the Heroes: The Aegean Before the Greeks* (1967); SPYRIDON MARINATOS and MAX HIRMER, *Kríti kai Mikinaiki Hellás* (1959; Eng. trans., *Crete and Mycenae*, 1960); FRIEDRICH MATZ, *Kreta und frühes Griechenland* (1962; Eng. trans., *Crete and Early Greece*, 1963). For the Aegean Bronze Age as a background to

Homer, see the sections in *Archaeologia Homerica* by various authors since 1967. H.L. LORIMER, *Homer and the Monuments* (1950), is still valuable. For the end of the Bronze Age, see VINCENT DESBOROUGH, *The Last Mycenaeans and Their Successors* (1964).

Special works: (*Crete*): A.J. EVANS, *The Palace of Minos . . . at Knossos*, 4 vol. (1921–35, reprinted 1964), is still basic; SINCLAIR HOOD, *The Minoans: Crete in the Bronze Age* (1971); CHRISTIAN ZERVOS, *L'Art de la Crète néolithique et minoenne* (1956), for pictures. (*The Cyclades*): CHRISTIAN ZERVOS, *L'Art des Cyclades* (1957), for pictures and references. (*The mainland*): G.E. MYLONAS, *Mycenae and the Mycenaean Age* (1966), for the late Bronze Age and Mycenaean shaft grave circle B; EMILY T. VERMEULE, *Greece in the Bronze Age* (1964), is the standard work. (*Architecture*): J.W. GRAHAM, *The Palaces of Crete* (1962). (*Religion*): M.P. NILSSON, *The Minoan-Mycenaean Religion and Its Survival in Greek Religion*, 2nd ed. (1950), remains the standard work. (*Seals*): JOHN BOARDMAN, *Greek Gems and Finger Rings* (1970). (*Writing and the decipherment of linear B*): JOHN CHADWICK, *The Decipherment of Linear B* (1958); MICHAEL VENTRIS and JOHN CHADWICK, *Documents in Mycenaean Greek* (1956), with a good survey of Aegean writing. References for authorities against the decipherment are listed by Hood in *The Minoans*, p. 167 (*op. cit.*).

(M.S.F.H.)

Aegean Sea

The Aegean Sea was the cradle of two of the great early civilizations, those of Crete (Kríti) and Greece, from which much of modern Western culture is derived. During the classical period it also served as a naval school, providing knowledge to all the people of the known world. In addition, it is an important natural feature of the Mediterranean region, possessing several unique characteristics that make it of considerable scientific interest. (For additional information see AEGEAN CIVILIZATIONS.)

It constitutes an arm of the Mediterranean Sea, located between the Greek peninsula on the west and Asia Minor on the east. About 380 miles long and 186 miles wide, it has a total area of some 82,625 square miles (214,000 square kilometres). The Aegean is connected through the straits of the Dardanelles, the Sea of Marmara, and the Bosporus with the Black Sea, while the island of Crete can be taken as marking its boundary on the south. The sea has an intricate configuration and could well be considered as a bay within the eastern Mediterranean Basin, to which it is connected by the straits to the west and east of Crete. It also has a good connection with the Ionian Sea to the west, through the strait lying between the Peloponnese peninsula of Greece and Crete. Virtually throughout the Aegean area, numerous islands large and small emerge from the clear blue waters. These are the mountain peaks of Aegeis, the name given to a now-submerged land mass. At the dawn of European history, these islands facilitated contacts between the people in this area and of three continents. Throughout the entire Aegean shoreline—that is, both the continental shores surrounding the Aegean Sea and those of the islands—bays, ports, shelter creeks, and so on are also abundant. These also facilitated the task of seamen travelling in the Aegean Sea, making longer and longer voyages possible at a time when shipbuilding was in its infancy. No other maritime area of the Mediterranean has such a great shoreline development as compared with its size.

Recent scientific investigation of the Aegean area has been carried out by various research vessels belonging both to Greek and foreign institutes. The latter include modern, well-equipped vessels from the United States, the U.S.S.R., France, and Sweden. Greek scientific bodies, however, continue to play a leading role in research.

Physical characteristics. The maximum depth of the Aegean is to be found in the east of Crete, where it reaches 11,627 feet (3,543 metres), although the average depth is 1,188 feet. The rocks making up the floor of the Aegean are mainly limestone, though often greatly altered by volcanic activity that has convulsed the region in relatively recent geological times. The richly coloured sediments in the region of the islands of Santoríni (Thíra) and Mílos, in the south Aegean, are particularly interesting. By the 1970s, Santoríni, in particular, had become a topic of major international scientific importance, analysis of its surrounding sediments having been linked with a possible explanation of the ancient legend of the "lost" island of Atlantis.

Winds and tides

North winds prevail in the Aegean Sea, although from the end of September to the end of May, during the mild winter season, these winds alternate with milder southwesterlies. The tides of the Aegean Basin seem to follow the movements of those in the eastern Mediterranean generally and have a limited tidal range of about 12–16 inches (30–40 centimetres). The tide of Euripus (Porthmós Evrípou)—a strait lying between continental Greece and the island of Euboea (Évvoia), in the heart of the Aegean—is, however, extremely important, as it displays a tidal phenomenon of international significance, to which it has, in fact, lent its name. The euripus phenomenon—characterized by violent and uncertain currents—has been studied since the time of Aristotle, who first provided an interpretation of the term. Aegean currents generally are not smooth, whether considered from the viewpoint of either speed or direction. They are chiefly influenced by blowing winds.

Maximum temperature values on the surface levels of the sea amount to 75°–77° F (24°–25° C) during the month of August but generally only reach 50°–52° F (10°–11° C) in February. In depths of up to about 1,600 feet, temperature fluctuates around 57°–64° F (14°–18° C), while it is practically steady around 57° F at sea level. Aegean Sea temperatures are influenced by the cold water masses of low temperature that flow in from the Black Sea to the northeast.

Salinity fluctuates around 36–39 parts per thousand, according to the season of the year, with the low level of Black Sea salinity (15 parts per thousand) influencing the salinity rate of the northern Aegean.

Marine life. The Aegean Sea, like the Mediterranean in general, is the most impoverished large body of water known to science. The nutrient content, as indicated by the amount of phosphates and nitrates in the water, is on the whole poor: the less saline waters coming from the Black Sea have a distinct ameliorative influence, but the role of their fertility in the Mediterranean in general has been little studied. Generally, marine life in the Aegean Sea is very similar to that of the northern area of the western basin of the Mediterranean. In view of its limpidity and as a result of its hot waters, it is not surprising that the Aegean Sea accommodates large quantities of fish at the time of their procreating maturity. Such fish enter the Aegean from other areas, notably from the Black Sea.

The resources of the Aegean other than its fish have been little utilized, although recent research has revealed the possibility of the existence of oil deposits beneath the seabed, to say nothing of mineral and chemical materials on the sea floor.

BIBLIOGRAPHY. M. BERNARD, "Recent Advances in Research on the Zooplankton of the Mediterranean Sea," in H. BARNES (ed.), *Oceanography and Marine Biology*, vol. 5, pp. 231–255 (1967), various investigations considered from the points of view of the particular disciplines or methods employed; J.J. BLANC, "Sedimentary Geology of the Mediterranean Sea," *ibid.*, vol. 6, pp. 377–454 (1968), an examination of general data on the sedimentation in the Mediterranean Sea and its relation to oceanographical and biological factors; A.D. HATZIKAKIDIS, "The Relations Between Air and Sea Temperatures," *Practica of the Hellenic Hydrobiological Institute of the Athens Academy*, pp. 1–112 (1963), a study of the relationship between air and sea temperature in the sea area; "Prognosis for the Development of Oceanic Resources in the Greek Seas," *Proceedings of* OECON, *the First International Offshore Exploration Conference*, pp. 455–468 (1966), a report on the probabilities of the existence of petroleum and chemicals in the Greek seas.

(A.D.H.)

Aerial Sports

Aerial sports have a fundamental position in aviation and encompass the true origins of aviation. Individuals in

the 1800s and early 1900s imagined, contrived, and developed contraptions that would enable them to experience the unique sensation of being airborne. The development of aircraft for commercial and military purposes came much later. Flight was founded in sport, and sport flying still thrives.

The Fédération Aéronautique Internationale was founded in Paris in 1905 to encourage the progress of world aeronautics. It has long exercised general supervision over most phases of aerial sport, certifying world records and sanctioning international competition. It includes national federations from more than 50 countries in its membership.

This article is intended for the general reader who may have no knowledge of the activities described. It deals chiefly with the four most popular varieties of aerial sport—power-plane sports, soaring (or gliding), sport parachuting (skydiving), and sport ballooning—sketching the history and present status of each. For more detailed information on specific activities, the reader should consult the works listed in the bibliography.

POWER-PLANE SPORTS

Power-plane activities involve flight with an airplane that is self-propelled. They began with the Wright brothers' first successful powered, sustained, and controlled airplane flights in 1903 and have expanded and diversified greatly ever since. Aside from simple recreational flying of stock aircraft, most current activity falls into one of five categories: air racing, aerobatics, and flying of homebuilt craft, antiques, and rotorcraft.

Air racing. *History.* Men began competing with each other in airplanes even before they knew how to fly them well. The first international races were held at Reims, France, in 1909, and three-fourths of the 38 entries crashed at least once during the week-long contest.

Between 1909 and World War II, air racing captivated the enthusiasm and attention of designers, manufacturers, governments, and the public. The major national and international events such as the Bennett Cup, Schneider, Thompson, Bendix, and Pulitzer races drew huge crowds: up to 500,000 persons for annual races and an estimated 1,000,000 for one Schneider race at London. Many of these were closed-circuit races: the Pulitzer was run over a three-cornered, 29-mile (47-kilometre) course.

Technological improvements to planes came very fast and several countries held leading positions at some point during this period. An American, Glen Curtiss, won the 1909 Bennett Cup contest at Reims. The French took over leadership for the next few years with their Nieuport and Deperdussin racers. From 1921 to 1926 the American Curtiss biplanes set records and won the major international contests. From 1927 to 1930 Great Britain excelled with its Supermarine racers, predecessors of the Spitfire, a celebrated World War II fighter. In the early 1930s Italy's design achievements were outstanding, and in 1933 an Italian aircraft set a 440-mile-per-hour (708-kilometre-per-hour) seaplane speed record that still stands. Landplane speed records were dominated in the 1930s by a succession of privately designed U.S. race planes. So rapid was the design progress during this period that one year's Pulitzer or Thompson winner was inevitably outclassed a year later. In 1939 a German Messerschmitt 209V1 flew an official 469.22 miles per hour (about 755 kilomètres per hour), setting a speed record that remained until 1969, when American pilot Darryl Greenamyer established a new standard of 483.041 miles per hour (about 777 kilometres per hour) in a modified Grumman Bearcat.

Speed races

The early races were more than simply sporting events. They were a professional challenge for designers, mechanics, and pilots to develop new ways to increase aircraft speed, range, and dependability. Many of the ideas and methods spawned in racing were utilized by military and commercial aviation. Racing pilots were either military pilots or professionals subsidized by related industries such as oil companies. Racing planes were expensive and required the economic support of government or industry.

Sport racing. Not until after World War II did the fever for design improvement give way to interest in pure sporting competition. The withdrawal of most commercial support and increasing costs brought a decline in racing. Another factor was the rising number of fatal accidents that plagued closed-circuit racing as speeds increased. In 1947, however, a new racing class was established for midget monoplanes with limited power engines and other design restrictions. This was a class that the sport racing pilot could afford. Midget races on closed courses soon became popular with both pilots and spectators, who could now see the entire race course instead of only the finish line.

In 1964 the first U.S. National Air Races since 1949 were held at Reno, Nevada. Competition classes included midgets (later called Formula One), sport biplanes, and unlimiteds (modified World War II fighters). After 1964 air racing improved, as many more cities hosted races, and classes expanded to include stock planes, military trainers, and a promising new category of Volkswagen-engined planes. Many women are active in air racing. An association of woman pilots called the Ninety-Nines hosts the Powder Puff Derby, an annual U.S. transcontinental stock plane race. (For world air speed records see RELATED ENTRIES under SPORTING RECORD in the *Ready Reference and Index.*)

Aerobatics. An increasing number of pilots are learning to enjoy aerobatics—the art of guiding a plane through unusual attitudes of flight. Aerobatics are strenuous exercise for both plane and pilot. The plane must be strong, for it is subjected to great stresses. The pilot is challenged physically and mentally because the precise control of the craft during manoeuvres requires concentration, coordination, timing, and judgment.

History. Until 1964 aerobatics were mostly performed by professional stunt pilots as an airshow attraction and by military student pilots as part of their proficiency training. In the United States, a new enthusiasm for sport and competition aerobatics began in 1964, when the first sanctioned national aerobatic contest was held at Reno, Nevada. At about the same time the Champion Aircraft Corporation introduced the Citabria, the first aerobatic production plane since before World War II. Many amateur-built aerobatic planes began to appear, and numerous flight schools introduced aerobatic courses.

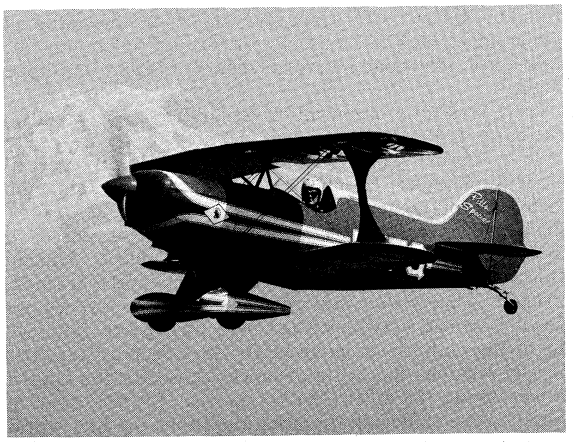
Aerobatic airplanes. Several U.S. aircraft companies manufacture planes stressed for primary aerobatics. Advanced aerobatics require aircraft with special features in addition to extra strength—they need high power and low weight; a wing that will develop lift in either upright or inverted flight; large control surfaces and fuel and oil systems that will allow the engine to operate during inverted flight.

The Yak (U.S.S.R.) and Zlín (Czechoslovakia) are the top European aerobatic aircraft. Both are manufactured as pilot trainers but have proven to be excellent aerobatic craft when equipped for inverted flight. In 1970 Germany introduced a new aerobatic ship called the Akro Star.

U.S. competition aerobatic planes are mostly antiques and trainers that have been rebuilt and modified by their owners. An exception is the Pitts Special, a tiny biplane with a 17-foot (five-metre) wingspan that was conceived as a sport and aerobatic homebuilt plane in 1948 and has been duplicated in dozens of home workshops. Later Pitts Specials with more powerful engines won several aerobatic championships and performed well in international competition.

Competition. Competition is at three levels: primary, advanced, and unlimited. Primary is the most popular category. In it, the competitor flies at between 1,500 and 3,000 feet (460 and 900 metres) and must complete ten predefined manoeuvres within eight minutes. Competitors in the advanced and unlimited categories fly a greater number of more difficult manoeuvres and are allowed a much lower altitude limit. Besides the predetermined group of manoeuvres, these competitors must devise and fly a free group sequence according to certain rules. Aerobatics judging is based on methods introduced by José L.

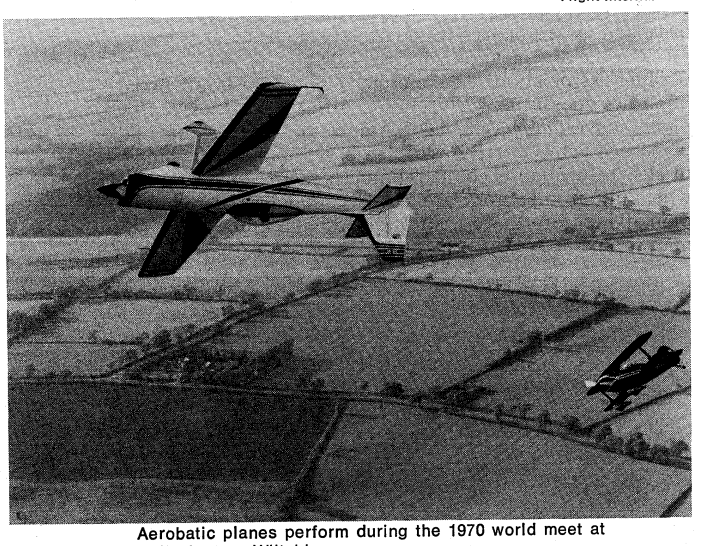
Bases for judging



Pitts Special aerobatic biplane.
Jim Larsen

Aresti, Spanish master of aerobatics, who devised a shorthand method of diagramming and scoring manoeuvres. Each manoeuvre has a "difficulty coefficient," or K factor. In competition judges observe each manoeuvre and score it from one to ten according to how well they think it was performed. This figure is multiplied by the manoeuvre's K factor. Local, national, and international aerobatic meets are held. The Experimental Aircraft Association (U.S.) hosts one of the largest contests with 80 or more participants. A world meet is held every

Flight International



Aerobatic planes perform during the 1970 world meet at Hullavington, Wiltshire.

two years under the auspices of the Fédération Aéronautique Internationale. Competition has been dominated by the U.S.S.R. and Czechoslovakia. With 11 countries competing, the U.S. team won for the first time in 1970.

Homebuilts. A strong motivation to fly in a machine of his own creation has kept many pilot-craftsmen busy gluing, rivetting, welding, bolting, and painting in home workshops throughout the world. The greatest amateur activity is in the United States, where there were about 3,000 "homebuilt" craft registered in 1970.

History. As early as 1912 plans were available to the home craftsman for simple gliders and primitive power planes. Some of these early products were so crudely designed and built, however, that flight in them was akin to suicide. Starting about 1927 government regulations imposed new high standards on all aircraft construction. These new rules provided no category for recreational flight in homebuilt planes except for racers able to meet new quality, testing, and licensing requirements. "Homebuilders" were discouraged, and there was little activity in home shops until 1947, when homebuilt racing was revived. In 1947 amateur pilot George Beaugardas flew a small, homebuilt plane from Portland, Oregon, to Washington, D.C., to demonstrate the dependability of amateur-built aircraft and to petition the government for more liberal regulations. Officials were impressed and a new amateur-built aircraft category was established.

Kinds of planes. Most homebuilts are one- or two-place designs intended strictly for recreational use in good weather. They are usually built on a small budget and four passenger or larger sizes are uncommon. Most of the planes are made from published plans of a proven design. They are easily constructed and may have folding wings for easy storage and transport. Some homebuilders are inspired by notable planes of antiquity, such as the SPAD, S.E.5, Fokker Triplane, and Sopwith (all World War I fighters), and many full-scale replicas have been made from original plans. Even more common are three-quarter scale replicas, which are more practical to build and operate yet still offer great nostalgic rewards.

Antiques. Aircraft technology has progressed so rapidly that a 30-year-old plane is regarded as an antique. Many aviation enthusiasts are fascinated by the planes of yesteryear, and a surprising number of relics have been restored and maintained in flying condition.

Kinds of planes. Antiques are classified into three principle categories. The first covers planes of World War I and older vintage. They have special mechanical and operational problems and are flown mainly for exhibition purposes.

The most popular class covers planes of the 1928 to 1941 era. Most of these are small biplanes, two-seat Aeroncas, Cubs, and a few four-place Stinsons. These planes have an operating cost within reach of the recreational pilot.

The third class of antique is composed of the larger biplanes and cabin monoplanes, which were the commercial airliners and mailplanes of the 1930s. They are not practical for sport flying, but many still function as workhorses. Many Travel Airs, Steermans, Howards, and even a few Ford Trimotors still earn their keep as crop dusters, cargo planes, and jump planes for sport parachutists.

Modified antiques

Many antiques are not pure but have been modified to make them more practical to fly. A more modern and dependable engine is one common modification. Tail skids were the rule on early sod flying fields, but hydraulic brakes and a steerable tailwheel increase the safety of ground operations on modern hard-surfaced airports. These changes do not disqualify the plane as an antique. If the plane is of totally new construction, however, either full-sized or in scaled-down proportions, it becomes a homebuilt replica and not an antique.

Organization. Antique aircraft organizations include informal clubs made up of owners of a particular kind of airplane as well as national groups.

Rotorcraft. By virtue of their slow-speed manoeuvrability the rotary winged aircraft hold great fascination for many sport flying enthusiasts. Helicopters are beyond the economic reach of most recreational pilots, but the autogyro provides almost as much manoeuvrability with only a small fraction of the expense and complexity of the helicopter. Autogyros were developed and used commercially in the 1920s and 1930s, but then were abandoned in favour of the more versatile helicopter. In about 1956 Igor Bensen introduced one-man rotorcraft and his company has marketed plans and kits under the trade name Bensen Gyrocopter. Autogyros licensed in the amateur-built category acquired a large following in the 1960s.

Design. Instead of wings to provide lift the gyrocraft uses a pair of rotor blades about ten feet (three metres) long that rotate about a centre pivot. Suspended from the pivot bolt is a simple tubular framework with three wheels and a seat for the pilot. The rotor gets its rotational power from airflow and not from engine drive. This eliminates the need for a complicated transmission, blade controls, and a tail rotor, which are required on a helicopter. The craft cannot take off vertically or hover as a helicopter does, but it is extremely manoeuvrable

and can fly as slow as 15 miles per hour (24 kilometres per hour). It can be flown without an engine like a kite in the wind or by towing behind a car. In this configuration it is called a gyroglider. The same framework can be used for a self-powered gyrocopter by installing an engine to drive a pusher propeller. The versatile gyrocraft framework can also be attached to plywood floats for operations on the water. The unpowered version is called a hydroglider and must be towed by a power boat. A more powerful engine and longer rotor blades are recommended for the hydrocopter to offset the added weight of the floats. By following the steps shown in available training manuals a pilot can teach himself to fly and obtain a private pilot license.

Organization. Canada, the United States, Great Britain, France, Germany, Czechoslovakia, New Zealand, and Australia all have gyrocraft activity. The Popular Rotorcraft Association was founded in 1962 to help unify the efforts of rotorcraft enthusiasts. Its membership had topped 12,000 by the 1970s with about 40 U.S. and four foreign chapters.

SOARING

Soaring, or flying, in a motorless sailplane, or glider, is a favourite sporting pastime for many who desire the feeling of being airborne without the distractions of engine noise and vibration.

A sailplane, or glider, pilot usually begins his flight with a tow to 2,000 feet (600 metres) from a small power plane. After releasing from the towline the pilot seeks to fly within areas of "lift" or upcurrents of air to gain altitude. If his judgment is good and the proper weather prevails, he may be lofted to 10,000 feet (3,000 metres) or above.

Altitude is wealth to a sailplane pilot, for each foot of height will yield 30 or 40 feet of horizontal travel. Altitude can be spent exploring through miles of sky searching for more updrafts, or it can be spent performing aerobatic stunts, such as loops, rolls, and wingovers.

Sailplanes, or gliders, are popular in almost all countries of the world with most activity occuring in North America, Great Britain, Germany, Poland, France, Australia, New Zealand, and the Soviet Union.

History. Gliding is really the most basic mode of flight and predates power-plane flight by almost half a century. In 1855 a Frenchman, Capt. Jean-Marie Le Bris, actually got off the ground in a glider towed by a horse. Between 1891 and 1896 German aeronautical pioneer Otto Lilienthal made thousands of flights in a primitive glider—running downhill until airborne, with handholds for support, he could maintain the craft in balance by shifting his weight.

Weight shifting was not satisfactory for American inventors Orville and Wilbur Wright, who worked intently around 1900 to develop a more sophisticated method of achieving control and stability. They achieved a 9 minute 45 second glider flight at Kitty Hawk, North Carolina, in 1911, a world record that stood for ten years.

Following World War I the Germans were denied the use of power planes by the terms of the Treaty of Versailles. Their enthusiasm for flight was directed toward gliders, and in 1920 the first of a series of yearly glider meets was held at the Wasserkuppe, a hill in the Rhön Mountains. Wolfgang Klemperer dominated the 1920 and 1921 events. In 1921 Klemperer flew his "Blaue Maus" for 13 minutes utilizing thermal upcurrents and achieved an altitude of 400 feet (120 metres) above his launch point.

Slope soaring in Germany

At the 1922 Wasserkuppe meet a glider called the "Vampyr" was flown with many of the design features of modern sailplanes. It was a high-wing monoplane with unusually long and slender wings and an almost fully enclosed cockpit. It had a sinking speed of only 2.6 feet (0.8 metres) per second. The plane was designed and built by the Technical University at Hannover, Germany. Using techniques of slope soaring (soaring from hills) developed by A.H.G. Fokker, the "Vampyr" was kept aloft for two hours.

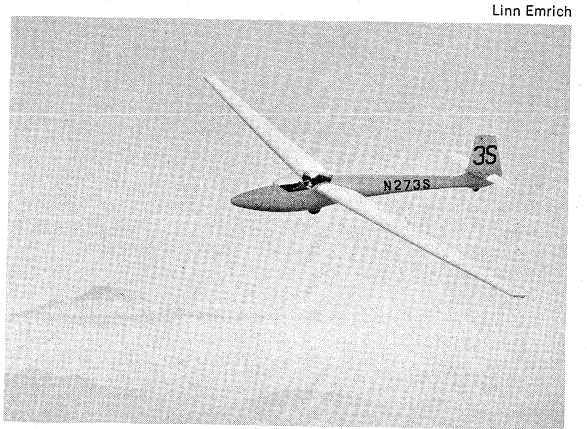
On a wave of enthusiasm generated by this flight the London *Daily Mail* sponsored the first international glider meet in 1922. It was won by a Frenchman, M. Maneyrol, who made a world record 3 hour 21 minute flight.

There was little organization to gliding activity before the early 1930s, and a majority of soaring records were held by Germany and Austria with only limited competition from other countries. Next, Great Britain, the U.S.S.R., the United States, France, and Poland all took an active interest in soaring, and records moved back and forth between countries with great rapidity. During this period national soaring federations were formed, bringing supervision and regulation to a sport whose reputation had been tarnished by a number of fatalities involving enthusiastic novices. In 1934 Richard du Pont briefly regained the world distance record for the United States, flying 158 miles (254 kilometres). By 1937 soaring had become sufficiently international that the first world championship contest was held in Germany on the Wasserkuppe.

Following World War II the Germans were again restricted for a time from operating power planes and again turned to the betterment of sailplane design. Richard Eppler and F.X. Wortmann both achieved recognition for their success in aerodynamic design improvements. Most of the principle glider factories use wing designs originated by either Eppler or Wortmann.

Organization. Current organizations include national groups such as the Soaring Society of America (SSA), formed in 1932, which with 10,500 members is the principle U.S. soaring pilots' organization and is designated by the Fédération Aéronautique Internationale (FAI), the world governing body, to conduct contests, maintain records, and award badges for soaring proficiency. The major soaring organization in Great Britain is the British Gliding Association. The club originated in 1930 and publishes one of the world's best soaring magazines, *Sailplane and Gliding*. The 1-26 Association is a group of owners of Schweizer 1-26 sailplanes, a popular make and model. 1-26 contests compare to class racing in sailboats, where the competitors are equally matched in equipment. There are more than 500 sailplane clubs in Germany alone.

Linn Emrich



High-performance glass fibre sailplane over Mt. Ralnier in Washington.

Glide ratio

Performance. A sailplane's performance or efficiency is measured by its maximum glide ratio (the ratio of distance travelled horizontally to distance dropped vertically) and minimum sink speed (rate at which a glider loses altitude). A medium performance sailplane has a glide ratio of 25:1 and minimum sink of 2.5 feet (0.8 metres) per second. High performance designs will glide at 40:1 or above and sink at about 1.8 feet (0.5 metres) per second. One of the most sophisticated modern designs, made in Germany especially for competition, boasts a specially built wing with a 72-foot (22-metre) span and a glide ratio of 5:1 at a gliding speed of 56 miles per hour (90 kilometres per hour).

Probably the most important single factor in determin-

ing sailplane efficiency is the aspect ratio of the wing, expressed as a number which is the wingspan divided by the average wing width. In the 1920s, it was realized that a long, narrow wing offers less resistance to the air than a short, wide wing of the same area. Early gliders had an aspect ratio of eight or ten. As construction problems were solved and stronger materials utilized, such as aluminum and fibre glass, designers were able to increase aspect ratio to 30 and above. A recent construction development utilizes hard and stiff carbon filaments laminated into fibre-glass wing spars.

Competition. Soaring is a competitive sport. Through application to his national organization, a pilot may be awarded silver, gold, or diamond achievement badges which are issued by the Fédération Aéronautique Internationale and recognized the world over. To qualify for the badges a pilot must make soaring flights that meet certain minimum standards of endurance, distance, and altitude gained. For a silver badge he must remain aloft for five hours, fly a distance of 31.1 miles (50 kilometres), and gain 3,281 feet (1,000 metres) above the release altitude. These "tasks" or "legs" are usually accomplished on separate flights. A gold badge requires a 186.4-mile (300-kilometre) cross-country leg and a 9,842-foot (3,000-metre) altitude gain. A diamond badge requires a 310.7-mile (500-kilometre) cross-country flight and a 16,404-foot (5,000-metre) gain of altitude, plus a flight of 186.4 miles (300 kilometres) to a predeclared goal.

In addition to this personal level of competition, pilots match skills at national and international contests sponsored in nearly every country that has soaring activity. There are two categories of competition: open class, in which planes have no wingspan restrictions, are allowed water ballast, variable geometry wings, and drag chutes; and standard class, in which planes must have a wingspan of 15 metres (49 feet) or less, speed-limiting dive brakes, and a fixed-geometry airfoil. Both classes allow retractable landing gear. Winners in national events are participants in a world meet, held every two years.

Flying the sailplane. *Controls.* The primary controls on a sailplane are the same as those on a power plane. A control stick between the pilot's legs operates the ailerons and elevators (the movable horizontal control surfaces at the trailing edge of the wing and tail). Two pedals on the floor control the rudder, and a spoiler handle actuates movable panels in the wings that are useful during landing approach to increase the glider's descent rate. Instrumentation is simple—normally an altimeter, an airspeed indicator, a compass, and a variometer (an instrument sensitive to air pressure, which tells the pilot of slight altitude changes). A two-way radio and an oxygen system are often carried on cross-country or high-altitude flights.

Although sailplanes are incomparably graceful when airborne, they are cumbersome to handle on the ground. Two or more people are needed for ground manoeuvring. The long wings make them difficult to hangar, so most sailplanes are kept on a trailer with wings removed. The trailer is an important accessory, as it is also used for retrieving the plane from its landing point.

Launching. Sailplanes most commonly begin their flights by being towed to an altitude of 2,000 to 3,000 feet (600 to 900 metres) by a small power plane. A 200-foot (60-metre) length of line is attached to a towhook at the tail of the power plane and to another towhook on the nose of the sailplane. The power plane takes off and climbs with the sailplane pilot working constantly to maintain a precise position behind the towplane. The towhook can be released by either pilot if trouble develops. The aerotow has proven to be the safest method of launching.

The ground-launch method of towing a sailplane aloft with an auto or winch is still used at some glider fields where a long runway is available.

Sustaining flight. After releasing from the towline the pilot of a sailplane is able to bank and turn, exploring the air for favourable air currents. By reference to the variometer the pilot can tell if "lift" is encountered. More experienced pilots learn to judge where upcurrents are most likely to occur, as over the upwind side of a hill, some distance downwind from a hill or ridge, over sun-warmed buildings or fields, along a cold front, or beneath or within cumulus clouds. A sailplane can gain altitude by flying a pattern within these rising currents of air. By the 1970s, world soaring records were over 46,000 feet (14,000 metres) in altitude, 700 miles (1,100 kilometres) in distance, and 95 miles per hour (150 kilometres per hour) in speed over a 100-kilometre (60-mile) course. Sailplanes can make only one landing approach after each flight so they are equipped with efficient air brakes, called spoilers, a wheel brake, and sometimes a drag chute.

Motor gliders. In an effort to gain independence for the soaring pilot, experiments have been made with small auxiliary engines mounted in the sailplane. Without aid of support crew these planes can taxi about on the ground, take off, and climb to altitude. The motor may then be shut down and the pilot can proceed with the pleasures of soaring. By restarting the engine in flight the pilot can regain altitude or fly across country when necessary.

Powered sailplanes are commonly called motor gliders, and factories in Germany, Great Britain, and France have produced several models. Interest in motor gliding is expanding rapidly. The Powered Sailplane Association was founded in the United States in 1970 and publishes the magazine *Motorgliding*. The first German national motor glider contest was held in 1970.

SPORT PARACHUTING

Sport parachuting, or skydiving, is the fastest growing aerial sport. In the 1960s parachuting expanded from an infant group of less than 1,000 participants to an organized body of over 250,000 men and women in 28 countries, including the United States, Canada, Great Britain, France, East and West Germany, Poland, Yugoslavia, Czechoslovakia, the U.S.S.R., Australia, and New Zealand.

History. The parachute was developed before 1900, but for the first 50 years it was used mainly by military and civilian pilots as an emergency method of escape if their aircraft became disabled in flight. It was also used by a few individuals who made a profession of jumping at air shows.

In France in the early 1950s a small group of hardy individuals were making jumps for the fun of it. They discovered that by packing the parachute into a cloth sleeve, the opening shock at high speed was reduced from an abrupt snap to a more comfortable tug. They were also responsible for developing the fundamentals of a stable body position for free fall (before the parachute opens) and the steerable parachute. The French techniques and enthusiasm for sport parachuting spread during the 1950s, and by 1960 there were parachute clubs at small airports in many areas of the world.

The sport. The basic techniques of the sport are virtually the same everywhere. Equipped with two chutes each, main and reserve, a group of three or more jumpers rides aloft over a target area called a drop zone. As the plane climbs, the jumpers note the landing point of a bright yellow sheet of weighted crepe paper, which they dropped over the drop zone. This tells them the average wind speed and direction. At a chosen altitude from 2,000 to 12,000 feet (600 to 3,600 metres) above the ground the pilot makes an upwind pass across the drop zone. One of the jumpers scrutinizes the ground below from the doorway of the plane and signals heading corrections to the pilot. At the chosen exit point, either singly or as a group, the jumpers launch themselves from the plane into the air.

During free fall the jumper savours the unique sensation of falling at a constant speed of 120 miles per hour (190 kilometres per hour), supported by resistance of the air. Free fall may last from a few seconds to a full minute or more, depending upon the exit altitude. Chutes are always opened at 2,000 feet (600 metres) to allow ample time for opening the reserve chute, if necessary.

The opening is an abrupt, but not uncomfortable, cli-

max to the free fall. The next two minutes are spent gliding and guiding the canopy with steering toggles toward the chosen landing point. The landing may be a gentle standing touchdown under ideal wind conditions, or the brisk, tumbling roll common on target accuracy jumps.

There are three principal reasons for the rapid growth of parachuting as a sport. The intense thrill of the jump may be one; another relates to the social enjoyments of close companionship during training, packing, riding in the plane, and returning from the drop zone. Mass hookups can make a social experience out of the free fall itself. The third reason is that parachuting is relatively inexpensive. Older model aircraft and reworked surplus chutes may be used, and costs shared among members of the participating group or club. Initial expenses for training vary with location.

Training. Length of training varies from a commercial four-hour jump course to the full day of preparation before a beginner's first jump allowed by most clubs.

During the training period a student jumper learns how to leave the airplane in a poised and stable position, how to use a reserve chute, how to steer the parachute, and how to fall and roll when landing. At least five jumps are usually required before a student is allowed to pull his own rip cord.

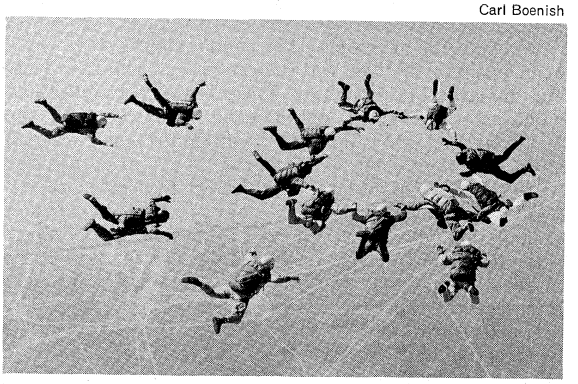
Static line jumping

The static line is a length of webbing used in lieu of a rip cord. One end of this is fastened inside the airplane to a special hook. The jump instructor also grips the static line in his hand. At the other end it is attached to the student's main parachute in such a way as to cause the container of the main parachute to open immediately when the static line is pulled taut and allow the canopy to begin deployment.

Curiously, it appears to take more courage to make the next few subsequent jumps than it does the first. Only about one-third of the first jump students return for further training jumps, and most of these succumb to anxieties and quit somewhere between the third and tenth jumps. If a student can persevere through ten jumps, prejump anxiety begins to subside and enjoyment increases. A jumper needs to experience about 100 jumps before he is proficient.

Advanced jumpers who have mastered the exacting art of free-fall manoeuvring can perform relative work, establishing contact with others. "Stars" or "snowflakes" are formed when eight or more jumpers link together in free fall to form a circle. Hookups of as many as 20 jumpers have been formed.

Carl Boenish



Nine-man free-fall hookup at approximately 4,500 feet above Elsinore, California.

Equipment. There are three principle types of main parachutes in use. The most economical chute is the military surplus chute that has been modified for sport use and is often used for student training jumps. The larger sizes are used for women, men over 35 years of age, or anyone over 180 pounds (80 kilograms). Surplus chutes are reliable and are sufficiently manoeuvrable to learn the fundamentals of descent and steerage. A second type is the commercial parachute, designed especially for sport use and used by both competition and recreational jumpers because of its good forward glide speed (about 12 miles per hour) and quick turning response (360° in about four seconds). A third type originated in 1970, when a way was discovered to inflate a fabric airfoil section with air ramming into openings in the leading edge, producing a parachute with a 3 to 1 glide ratio and a forward speed approaching 30 miles per hour (48 kilometres per hour).

Besides a main parachute the sport chutist wears a reserve chute, to be used in the extremely rare instance that the main chute might malfunction. Special high-topped boots with air cushion soles, goggles, helmet, and coveralls complete the parachutist's equipment. Some jumpers may also wear an altimeter and a stopwatch.

Competition. Parachuting competitions are usually of two kinds: accuracy and style. Accuracy contests usually allow three to six jumps per contestant and require the jumper to guide his canopy so that he can touch his foot upon a six-inch (15-centimetre) target disk. Judging is by measurement of distance from disk centre to the nearest landing point. The final score is simply the accumulated distance in centimetres for all of the contestant's jumps in the meet, and the winner is the one with the lowest score. Good competitors often make several consecutive landings directly upon the target disk.

Judging jumpers for style and speed

In style contests jumpers are judged by the speed with which they can complete a series of free-fall manoeuvres. Commonly two loops in the vertical plane and four 360° turns in the horizontal plane constitute a series, and ten seconds or less to complete this series is considered a competitive score.

Other events, such as team free-fall hookup events, may be arranged at the discretion of the contest directors.

Local clubs usually host meets. In the United States regional meets are sanctioned by the United States Parachute Association to choose the 120 best jumpers to compete nationally. The top ten men and top five women at the national meet make up a national team that participates in international contests.

Sport parachuting came into its own at the first world championship meet, held in Yugoslavia in 1951 under the sponsorship of the Fédération Aéronautique Internationale with five nations participating. World championship meets have been held every two years since 1954, with Czechoslovakia and the U.S.S.R. usually producing the strongest teams, although the United States and France have also performed well. (For lists of world champions and records see RELATED ENTRIES under SPORTING RECORD in the *Ready Reference and Index*.)

SPORT BALLOONING

History. Balloons were the first aerial conveyance by which man succeeded in sustaining himself aloft. Two brothers, Étienne and Joseph Montgolfier, made the initial successful balloon experiments at their paper factory at Annonay, France, about 1782. They proved that a balloon 40 feet (12 metres) in diameter could lift a man if the air inside it was warmed.

Hydrogen soon was discovered to be a good lifting gas if contained by a rubberized cloth balloon. Flights of many hours' duration were possible in gas balloons, but many accidents occurred due to the highly flammable properties of hydrogen. Helium is a safer gas now used for this type of balloon.

Gas ballooning for sport never really became popular because of the high cost of the gas and because of the tremendous output of labour necessary to make an inflation. A crew of perhaps 12 persons would be occupied for several hours tending the envelope and adjusting sandbags for a single inflation. Although there are still a few annual events for gas balloons each year in Europe and the United States, hot-air ballooning has become more popular for sporting purposes.

Hot-air ballooning. In the late 1950s a new type of sport hot-air balloon was marketed in which heat was provided by a propane burner. It proved to be efficient and economical and finally gave the sport balloonist an answer to his need for an economical lifting force, as the amount of propane needed to make a flight is quite inexpensive. In this type, fuel tanks are attached to the

gondola and a burner is mounted above the pilot on a platform. Vertical control of the balloon's flight path is achieved by valves that regulate the amount of propane flowing to the burner. Inflation can be accomplished by a crew of four in just a few minutes.

There are two basic designs. The ripping balloon has an opening at the top that can be pulled open to accomplish deflation. The dumping balloon has no opening at the top and is deflated by inverting the balloon with a line attached to the top after releasing it from the gondola attachment.

Balloons are classified by the Fédération Aéronautique Internationale according to volume with an AX number from 1 to 10. The popular single-place 30,000-cubic-foot (850-cubic-metre) model is an AX-4. The next most popular size of 56,000 cubic feet (1,600 cubic metres) will carry three people and is designated AX-6.

Competition. Most competition is on the local level, and there are many ways for pilot skills to be matched. In a hare-and-hounds race a single balloon, the hare, takes off and flies for a certain period of time. After a short delay the competitors take off, and the winner is the balloonist who navigates skillfully and lands his craft closest to the hare. A similar contest is the spot landing event, in which a predetermined target is chosen downwind, and the competitors attempt to land as close to it as possible. The cross-country distance event is won by the balloonist who travels the farthest in a certain period of time. In this event the balloonist attempts to choose a flight level with the strongest wind flow. The Fédération Aéronautique Internationale maintains world duration, distance, and altitude records in each balloon class for races held under sanctioned conditions. (For world ballooning records see RELATED ENTRIES under SPORTING RECORD in the *Ready Reference and Index*.)

BIBLIOGRAPHY. LINN EMRICH, *The Complete Book of Sky Sports* (1970), an instructional text on the fundamentals of flight in five categories: soaring, parachuting, hot-air ballooning, gyrocopters, and power planes; CARL CONWAY, *Joy of Soaring* (1969), a complete modern training manual on sailplane flying; P.M. BOWERS, *Guide to Homebuilts* (1962), a guide to the history, design, construction, testing, costs, and material sources for homebuilt aircraft, *Antique Plane Guide* (1962), a comprehensive study on the identification, restoration, costs, material and data sources for antique planes; DON VORDERMAN, *The Great Air Races* (1969), a fascinating resume of prominent planes, pilots, and air races since 1909; DUANE COLE, *Roll Around a Point* (1965), an instructional text on how to perform aerobatics by a world authority.

Periodicals: Sailplane and Gliding (bimonthly), official organ of the British Gliding Association; *Soaring* (monthly), a publication of the Soaring Society of America; *Parachutist* (monthly), publication of the U.S. Parachute Association; *National Aeronautics* (quarterly), official publication of the U.S. National Aeronautic Association. See also *Popular Rotorcraft Flying* (bimonthly), *Sport Flying* (monthly), *Motorgliding* (monthly), *Antique Airplane News* (monthly), *Sport Aviation* (monthly), and *Ballooning* (quarterly).

(L.E.)

Aeronautical Engineering

Aeronautical engineering traditionally applies the principles of the physical and mathematical sciences to the development, design, construction, testing, and operation of flight vehicles that operate in the Earth's atmosphere. During the decade of the 1960s aeronautical engineering was broadened to include all vehicles that operate above the surface of the Earth, in space as well as in the atmosphere. The terms aerospace engineering or aeronautical and astronautical engineering are sometimes employed in place of aeronautical engineering to describe the broadened engineering process.

The principal technologies encompassed by aeronautical engineering are those of aerodynamics, propulsion, structures and stability, and control. The aeronautical engineering process begins in academic, industrial, and government laboratories with the evolution of new concepts and devices. Industrial designers, using the latest technological developments, continue the process by laying down an initial vehicle design that is calculated to satisfy the project's requirements. The definition of an initial design is followed by a long process, lasting months and sometimes years, in which the detail design and development of the many vehicle components are carried out by aeronautical engineers. The design process precedes a period of construction and testing of one or more prototype vehicles. Satisfactory completion of flight testing of prototypes is followed by quantity production and operation. Aeronautical engineers participate in all steps of the total process.

Initial vehicle design

Historical survey. Following the first heavier-than-air flight by the Americans Orville and Wilbur Wright on December 17, 1903, the design of aircraft was based on empirical knowledge and crude construction techniques. The resources of science were not brought to bear on aircraft design and development for another decade, when it became apparent that the airplane was a potentially useful military weapon. It was recognized by engineers before World War I that the effective development of the airplane as a military weapon depended upon application of the physical and mathematical sciences to the numerous technical problems involved in such research. Perhaps the first organized attempt to apply science was the establishment in 1909 of an Advisory Committee for Aeronautics at the National Physical Laboratory in Great Britain. In 1915 the United States government followed with the establishment of a similar laboratory organization under the direction of the National Advisory Committee for Aeronautics. During this period, under the stimulus of World War I and with the rapid application of scientific principles, the airplane progressed in a few years from a relatively unimportant machine shown at aerial exhibitions to an effective military vehicle.

Development during the World Wars. World War I brought demands for increased payloads, greater speeds, and higher altitudes. It was recognized that increased engine power was needed to satisfy these demands. One of the most significant power-plant developments during this period was that of the American Liberty, a 400-horsepower engine with a V-12 arrangement. After the United States automobile industry was enlisted to produce Liberty engines, they were turned out in greater quantity than the aircraft designed to use them. As a result, surplus Liberty engines were used well into the 1930s in a variety of civilian and military aircraft.

As early as 1915, the basic elements of modern aircraft construction, the full cantilevered wing with metal monocoque (in which the metal skin carries the stresses of the fuselage), had been incorporated in the German Junkers J.1 monoplane and tested in flight. Another 20 years passed, however, before these elements were universally employed. The welded steel fuselage and wood spar wing, both fabric-covered, with biplane or externally braced monoplane configurations, dominated the 1920s and 1930s. By the late 1930s, hastened by preparation for World War II, the all-metal, stressed-skin monoplane construction was universally adopted.

The development of aeronautical engineering was again accelerated by World War II. The internal combustion engine, although used predominantly during World War II, began to give way to the turbojet before the war ended. The first flight with a turbojet engine took place when a German Heinkel He 178 fighter flew on August 27, 1939. Wartime research and development on jet propulsion in Britain and Germany progressed vigorously. By the summer of 1944 the Royal Air Force had the jet-powered Gloster Meteor in service, and shortly thereafter the German Luftwaffe introduced the Me 262 jet fighter into wartime combat. Although the turbojet engine was quickly adopted by aeronautical engineers for military aircraft, commercial applications developed slowly because of its high fuel consumption. The turboprop engine, in which a propeller is driven by a gas turbine, was used on commercial aircraft for about a decade until the specific fuel consumption of the turbojet engine was reduced to economical levels by research and development.

Turbojet engine introduced

Postwar advances. The availability of rocket and turbojet engines in the post-World War II period permitted new aircraft to be driven at ever increasing speeds. The first supersonic flight in manned aircraft in level or

climbing flight was made by Capt. Charles E. Yeager of the U.S. Air Force at Muroc, California, in a rocket-powered research airplane, the Bell X-1. As aircraft speeds approached and exceeded the speed of sound in air, pronounced changes in external air-flow patterns resulted. Much of the work of aeronautical engineers in the post-World War II period has been directed toward mastery of the aerodynamic, stability, and control problems that result from flight near and above the speed of sound.

The post-World War II period was one of rapid change in aeronautical engineering, as new knowledge in the fields of aerodynamics, structures, propulsion, stability, and control developed at increasing rates. Its application improved the performance of aircraft still further and raised new problems: low-speed aerodynamics gave way to transonic, supersonic, and hypersonic aerodynamics; frame structures were replaced by thin metallic shells that required the development of new aluminum, magnesium, titanium, and steel alloys; internal combustion engines were replaced by rocket and turbojet engines; and manual control developed into automatic control. The result was a new era of aeronautical engineering in which scientific methods were logically applied by teams of engineers in order to develop new high-performance airplanes with a high level of predictability of the final result.

Development of aeronautical engineering as a profession. Aeronautical engineering as a profession evolved along with infusion of the physical and mathematical sciences into the design and construction of aircraft. Aeronautical-engineering professional societies emerged in most advanced industrial countries of the world as the airplane achieved greater importance. The Royal Aeronautical Society in England, founded in 1866, became the principal society for aeronautical engineers in that country. The Institute of Aeronautical Sciences was formed in the United States in 1933. As interest in space exploration grew in the early 1960s, the American Institute of Aeronautics and Astronautics was formed, absorbing the Institute of Aeronautical Sciences. This organization grew to more than 36,000 members during the decade of the 1960s.

Fundamental principles. The practice of aeronautical engineering rests upon principles that are fundamental to all flight vehicles: a propulsion system with high thermodynamic and propulsive efficiencies; a structure of minimum weight and maximum strength; an external shape that is stable in flight with maximum aerodynamic efficiency; a precision control and guidance system; and a suitable design compromise among all these elements that allows the vehicle to achieve the required performance. An understanding of these principles depends upon knowledge of the engineering sciences: material science, solid and structural mechanics, thermodynamics, fluid mechanics, and electrical science. In turn, the engineering sciences rest upon the basic sciences: physics, chemistry, and mathematics.

Sciences required in aeronautical engineering

Bréguet's equation, defining the range of flight, may be used as an illustration of how the performance of an airplane depends on the fundamental principles discussed above. Louis Bréguet, a French aviation pioneer, showed that the range of an airplane, or the maximum distance it can travel, depends on two main factors—flight efficiency and the ratio of fuel weight to the gross airplane weight at the end of cruising flight. Range is equal to the flight efficiency multiplied by the logarithm of 1 plus the ratio of the weight of fuel at the beginning of the flight to gross airplane weight at the end of the flight: $R = (\text{flight efficiency}) \times \log (1 + W_F/W_L)$. Bréguet conceived the flight efficiency as a composite measure of the aerodynamic and propulsive efficiencies of the plane, obtained by dividing the product of the speed (Mach number) and lift-to-drag ratio by the engine's specific fuel consumption:

$$\text{flight efficiency} = \frac{\text{ML/D}}{\text{SFC}},$$

in which

$M = V/c$ = Mach number of cruising flight;
L/D = lift-to-drag ratio of the airplane in cruising flight;
SFC = specific fuel consumption of engines in cruising flight;
W_F = weight of fuel on board at start of cruising flight;
W_L = weight of airplane at end of cruising flight;
V = speed of cruising flight;
c = speed of sound at altitude of cruising flight.

It is apparent from Bréguet's equation that, if everything else remains the same, flight efficiency increases with increasing Mach number, M. This is one of the reasons why very high-speed flight is attractive to engineers. The second component of flight efficiency is the lift-to-drag ratio, L/D, of the airplane. The lift-to-drag ratio is a measure of the airplane's aerodynamic efficiency. The lift is an upward aerodynamic force on the airplane that is perpendicular to the path of flight. The drag is the total resistance the airplane encounters in driving through the air. It is a rearward acting force parallel to the flight path. When the airplane is in steady level flight, the lift force is equal and opposite in direction to the downward-acting airplane weight, while the drag force is balanced by the forward-acting engine thrust. Much of the work of aeronautical engineers applies the principles of aerodynamics to the calculation of lift and drag. The wind tunnel, one of the principal experimental tools used by aeronautical engineers, is used to measure airplane lift and drag by placing models in a moving airstream. Well-established similarity rules are used to scale up the measurements made on models to those of full-scale aircraft.

Lift and drag

The third factor in Bréguet's definition of flight efficiency is that of engine specific fuel consumption, SFC. The engineer attempts to reduce the specific fuel consumption to its lowest possible level by improving the overall thermodynamic efficiency of the engine and by improving the propulsive efficiency of the driving or thrusting system.

As the Bréguet equation suggests, flight efficiency is not the whole story in the achievement of range. In order to extend range, a designer must also strive for a high ratio of fuel load at the beginning of the flight to gross weight of airplane at the end, W_F/W_L. This is tantamount to saying that for a given fuel load and payload, he must reduce the empty weight of the airplane as much as possible. This has been the perennial problem of aeronautical engineering from the earliest days of aviation.

Educational programs. A beginning in aeronautical engineering education was made by the College of Engineering of the University of Michigan, where regular courses were introduced in February 1915. Courses in aeronautical engineering were introduced shortly thereafter at the Massachusetts Institute of Technology. In the late 1920s, especially after the stimulus of Charles Lindbergh's solo flight from New York to Paris in 1927, aeronautical engineering curricula were developed in scores of universities throughout the world.

Instruction in aeronautical engineering in the 1920s and 1930s was often based on a curriculum in mechanical engineering with elective courses in applied aerodynamics and airplane design. These courses usually included wing theory, wind-tunnel testing, static stability and control, dynamic stability, performance calculations, and general airplane design considerations. Applied aerodynamics and airplane design characterized aeronautical-engineering education in the 1920s. Airplane structures of this period were similar to bridge and building structures, and airplane structural analysis and design were largely in the hands of engineers educated in university departments of civil engineering. The emergence of stressed-skin monocoque shells in the 1930s, however, caused a departure from traditional civil-engineering practice, and courses in airplane structures were added to aeronautical-engineering curricula.

As aeronautical engineering courses continued to develop through the 1930s and 1940s, separate departments

Development in the 1930s and 1940s

of aeronautical engineering were formed in most schools of engineering. A modern curriculum in aeronautical engineering begins with a firm grounding in the basic sciences, with an emphasis on mathematics, physics, and chemistry. The engineering sciences build on the basic sciences in the areas of material science, rigid-body and solid mechanics, thermodynamics, fluid mechanics, and electrical science. Finally, the aeronautical engineering student studies theoretical and applied aerodynamics, structures, propulsion, stability and control, including automatic control, and design.

In the United States, the standards for university engineering education are set by the Engineers Council for Professional Development (ECPD), which accredits engineering colleges to grant degrees in aeronautical engineering. In 1970, over 40 institutions in the United States were accredited to offer degrees in aeronautical or aerospace engineering. This did not include those institutions offering aeronautical-engineering options in their mechanical-engineering departments.

Although university graduates occupy the key technical positions in aeronautical engineering, there is a large demand for technicians. Being a highly technical enterprise, the aviation industry requires a wide range of scientific and technical education that ranges from the Ph.D. engaged in research to the technician, who supplies mechanical and electronic skills. Technician training is provided by technical institutes, junior and community colleges, and mechanics and flight schools. These institutions, which require relatively short periods of study—a year or two—do not provide the rigorous education in mathematics and basic sciences ordinarily found in university training. The object of such training is usually to prepare the graduate for immediately useful occupations.

Responsibilities of the aeronautical engineer. The largest employers of aeronautical engineers are the large private-aircraft, missile, and spacecraft industries and the supporting subcontractors associated with these industries. Aeronautical engineers employed in this complex of activities are engaged in a wide variety of duties ranging from research in the quest for new fundamental knowledge to hardware design and production. Most engineers engaged in industrial work hold B.S. degrees from four-year university programs in aeronautical engineering, although an increasing number of M.S. and Ph.D. holders are entering these activities.

The second largest employer of aeronautical engineers in most countries is the government. Here the duties lie principally in research and development and procurement. Aeronautical engineers are also employed to a limited extent by the airlines and the universities. Important engineering functions must be carried out by the airlines in the procurement of new equipment and in the maintenance and operation of equipment. University needs for aeronautical engineers are relatively modest, primarily for faculty and research laboratory staffs.

Role of research and development

Research and development is a major part of the aeronautical engineering process. Because of the complex nature of modern high-performance aircraft, the production of new aircraft models is usually preceded by a lengthy research and development process that is culminated by flight tests on one or more prototype vehicles. Research and development activities may be found in industrial, government, and university laboratories. In general, university laboratories engage in research leading to improved understanding of fundamental phenomena. Examples are studies of shock wave boundary layer interaction, studies of the behaviour of materials, development of improved methods of computing flutter speeds in supersonic flight, studies of the complex three-dimensional flow in compressors and turbines, and studies of stability and control in transonic flight. This research is, for the most part, conducted by faculty members and graduate students, although a few universities employ a small number of full-time research staff members.

New aircraft are developed by industry with large design teams usually divided into groups. One group devotes its efforts to preliminary design, or the initial conceptual design of the new airplane. The other groups divide the bulk of the detail design effort among them. The aerodynamics group is responsible for conducting the wind-tunnel tests, designing the overall aerodynamic shape, computing the performance, and taking part in all of the design activities that involve an understanding of the internal and external flows of air. The structures group has the responsibility for designing a structure that conforms to the desired external shape and at the same time is sufficiently strong and light to allow the airplane to achieve its desired performance. The power-plant group has the responsibility for selecting an engine that will give the airplane its desired performance and for designing the power-plant installation. Most aircraft design teams have a stability and control group that is responsible for ensuring that the airplane is stable and fully controllable over the speed range. Modern high-performance aircraft usually require stability-augmentation systems to achieve stability about all three airplane axes throughout the full speed range. Stability-augmentation systems employ electromechanical devices that move the control surfaces automatically in response to motion sensors in such a way as to augment the stability of the basic airplane. Aeronautical engineers who design stability and control systems combine a unique knowledge of aerodynamics, airplane dynamics, electronics, and control theory. The flight testing of the prototype of a new airplane is the responsibility of a flight test group consisting of aeronautical engineers and engineering test pilots. The flight testing is carried out carefully and scientifically, using a variety of sensors to measure the behaviour of the airplane and its systems under a wide range of operating conditions.

Future developments. The role of aeronautical engineers in meeting transportation needs is rapidly increasing. Perhaps the foremost example of new flight vehicles near realization in the early 1970s was the advanced short-haul transport, which incorporates advanced technology to achieve all-weather operations from landing strips no longer than 1,500 feet (450 metres) but which at the same time will retain high cruise efficiencies. These aircraft could provide service from large airports serving domestic and transoceanic jets to smaller, close-in airports in cities 100 to 1,000 miles (about 160 to 1,600 kilometres) apart. They could also be used for trips from outlying suburban areas to major centres. A second advance in air transportation will come with supersonic transports that offer the promise of maximum performance between major world cities. A third major advance will come through increases in total performance of aircraft derived from the present long-range subsonic jets. Another advance that will emerge from new technology relates to the much discussed personal aircraft. Although the personal aircraft has not yet found widespread use, new developments in aeronautical engineering, particularly in the field of stability and control, permit many new possibilities for the increased use of light aircraft.

New flight vehicles

Hypersonic aircraft, powered by air-breathing hydrogen-fuelled engines, represent another example of new potentials and challenges for aeronautical engineering. Three may be mentioned: (1) Earth-to-Earth transport offering higher efficiencies and speeds than existing aircraft; (2) first-stage booster offering flexibility in spacecraft launches; and (3) Earth-to-orbit space shuttle offering the final closure between atmospheric and orbital flight. Although at this writing the technologies required to reach these goals are not entirely in hand, research has progressed to a point at which the uncertainties can now be clearly defined in the areas of aerodynamics, propulsion, structures, stability, and control and design.

BIBLIOGRAPHY. F. ALEXANDER MAGOUN and ERIC HODGINS, *A History of Aircraft* (1931), an early historical survey of aircraft and engines prior to the advent of the all-metal airframe; DARROL STINTON, *The Anatomy of the Aeroplane* (1966), a description of the components of aircraft and engines prior to World War II; P.J. MCMAHON, *Aircraft Propulsion* (1971), a discussion of the principles of jet and rocket propulsion; D.J. PEERY, *Aircraft Structures* (1950), an exposition on the methods of structural analysis and design of monocoque aircraft structures; BERNARD ETKIN, *Dynamics of Atmospheric Flight*

(1971), a comprehensive treatment of the theory of the dynamics of flight vehicles in the atmosphere; EDWARD LEWIS HOUGHTON and A.E. BROCK, *Aerodynamics for Engineering Students*, 2nd ed. (1970), an introductory text on theoretical and practical aerodynamics of flight vehicles; NASA, LEWIS RESEARCH CENTER, *Exploring in Aeronautics: Introduction to Aeronautical Sciences* (1971), a comprehensive introduction to the science of flight from the standpoint of both the aircraft and its engine; H.G. STEVER and J.J. HAGGERTY, *Flight* (1965), an illustrated description of the history and principles of flight.

(R.L.Bi.)

Aerospace Industry

A combining and contraction of "aeronautics" and "space flight," "aerospace" covers all vehicular flight, whether within or beyond the Earth's atmosphere. The aerospace industry is the complex of manufacturing concerns engaged in the production of flight vehicles, including unpowered gliders and sailplanes, lighter-than-air craft, ground-effect machines, heavier-than-air craft of both fixed-wing and rotary-wing varieties, military missiles, space-launch vehicles, and manned or unmanned spacecraft; propulsion systems and other thrusting devices; on-board equipment essential to the design purpose of the flight vehicle; and ground-based support equipment needed for the operation and maintenance of the flight vehicle. An additional area of the industry's effort, expanding rapidly, is fabrication of nonaerospace products and systems derived from aerospace technology.

The aerospace industry, which of necessity devotes a considerable portion of its effort to research and development in the broadening realm of flight, is a leader of world technological advance. Despite the lack of statistical detail from the Communist nations, it would appear that the aerospace industry is the world's largest manufacturing industry in terms of employment and the monetary value of its product output.

Nations with extensive industrial development

The two greatest aerospace industrial complexes are those of the United States and the Soviet Union. France and the United Kingdom round out the Big Four, those nations that have achieved a high degree of aerospace sophistication and have the ability to manufacture home-designed systems embracing the entire range of aerospace vehicles. Japan, whose aircraft industry was entirely destroyed in World War II, did not resume production activity until 1952 but expanded rapidly; it produces few home-designed products, however, relying for the most part on U.S. licenses. Canada has a relatively small but advanced industry. The Federal Republic of Germany has little aerospace production capacity but has developed a strong research and development base. The People's Republic of China had been building aircraft of Soviet design for years and by the start of the present decade may have been producing home-designed aircraft; its demonstrated ability to launch space satellites indicated an advancing technology in other aerospace areas. Sweden's aerospace industry has been very small, but the national policy of selective specialization has brought about a high degree of competence in a narrow focus of effort. All told, there are 30 nations engaged in some form of aerospace production activity, for the most part the manufacture of light and medium aircraft of home design or production of heavier aircraft and high-thrust propulsion systems under license from major-nation companies.

For the major aerospace nations of the world, their own governments and in some cases other governments constitute by far the largest customer. Next in importance as buyers are the world's commercial airlines, a market long dominated by the United States despite recent inroads by the French and British. In terms of units, the production of general-aviation aircraft, private, business, and nonairline commercial planes is high, but the average unit cost of these aircraft is very low by comparison with jetliners; as a result, the sale of general-aviation equipment represents a relatively small part of the overall volume. In the United States, nonaerospace systems, primarily for state and municipal governments, utilities, surface-transportation companies, hospitals, and other segments of the medical community, represent a significant portion of total sales. Although the ratios vary among the major nations, the distribution of shares of the market for 1970 sales volume of the United States aerospace industry could serve as a general indicator of customer breakdown: government (defense), 59 percent; government (space), 11.7 percent; commercial (airlines), 17.1 percent; commercial (general aviation), 1.4 percent; nonaerospace, 10.8 percent.

This article is divided into the following sections:

I. Development of the aerospace industry
II. Aerospace product line
  Major product: flight vehicles
    Aircraft
    Guided missiles
    Spacecraft
    Propulsion systems
  Other products
    Secondary aerospace products
    Tertiary aerospace products
    Nonaerospace products
III. Research and development
  New commercial aircraft
    Computer design
    Wind-tunnel testing
    The prototype
    Flight testing
  Missiles and spacecraft
    General considerations
    Testing
  Other areas of research
  Cooperative programs
IV. Manufacturing facilities and techniques
  Basic fabrication techniques
    Metal processing
    Joining methods
    Use of electronic computers
  Product assurance
  Assembly methods and facilities
  Maintenance facilities
    Airframe and engine overhaul
    Remanufacture process

I. Development of the aerospace industry

The aerospace industry may be said to trace its origin to December 17, 1903, the date on which Orville and Wilbur Wright made the first powered flights of a heavier-than-air machine. An industry, however, is by definition a business activity, and the Wrights in 1903, and for several years thereafter, were engaged in a purely experimental effort. A more accurate date for the industry's origin is therefore perhaps 1908, when the Wright Company sold its patent rights for France to a syndicate called the Astra Company, the first recorded commercial transaction of the aerospace industry. In the same year, the Wrights contracted with the U.S. Army for the first military airplane, a contract successfully concluded the following year after army acceptance tests.

The first aerospace transaction

In 1909 the list of airplane producers was increased by Glenn Curtiss in the United States and Henri Farman, Louis Blériot, Gabriel and Charles Voisin, and Léon Levavasseur in France. The Wrights licensed production of their airplane in England, and experimental groups that would start regular production within a year or two were active in Germany and Russia. The planes being built differed in design but were similar in construction: wing and fuselage frames were made of wood, usually spruce or fir, covered by a coated fabric. This kind of construction predominated until after World War I.

Over the next five years, France and Germany, both aware of the war potential of the flying machine, began aircraft construction on what was then a large scale. By the outbreak of war in mid-1914, France had built some 2,000 airplanes of all types, 1,500 of them military. Germany started the war with about 1,000 military machines. Great Britain was a poor third as regards production, and the United States had completely abdicated its former role as aeronautical leader; in the years prior to 1914 the U.S. had built fewer than 100 airplanes, fewer than half of them military. The French industry, assisted by rapidly expanding facilities in Britain, carried the Allied aircraft-production load through most

of the war. When the United States joined the Allies, two types of British-designed bombers were assigned it for production, along with one American-designed plane (the Curtiss Jenny trainer). The U.S. did not develop enough production by the war's end to have much effect militarily, but the 400-horsepower Liberty engine, a very advanced power plant for its day, provided the foundation of an aero-engine industry.

Worldwide aircraft production went into a sharp decline after the war, until commercial air transportation, started in 1919, gained public acceptance. The airplane itself was considerably refined in terms of design and performance during the '20s, the most notable change being the gradual ascendancy of the monoplane over the biplane. From the standpoint of manufacturing, the largest change was a swing to metal construction, a wartime innovation introduced in Germany by Hugo Junkers, who also introduced the cantilever wing, which required no external struts or braces. The combination of cantilever wing and metal construction met with rapidly widening acceptance. Metal (steel and aluminum alloys) initially was used only in the airframe structure, but, in the latter years of the decade, it was extended to the wing and fuselage skins, and by 1930 the all-metal airplane was appearing with increasing frequency.

In the 1930s a tremendous acceleration of aeronautical research brought the airplane to design maturity. There were hundreds of innovations, among the more important being large increases in aircraft-engine power, new propeller efficiency, cabin pressurization, retractable landing gear, landing flaps, the autopilot, better instruments and controls, and new aids to navigation for all-weather flying. Performance increased; a major reason was the trend toward "cleaning up," or streamlining the airplane for reduced air resistance. Except for light private planes, all-metal construction became standard (see AIRCRAFT; AIRCRAFT, MILITARY).

Coming of the large transports

The most notable impact on manufacturing processes in the 1930s was aircraft size. The growing need for greater capacity in commercial transports, coupled with military requirements for longer range and heavier bomb loads, brought a generation of large-size aircraft. Britain and Germany led the way in the early 1930s with transports capable of carrying as many as 40 passengers; the U.S. took over in the latter half of the decade and introduced the first really modern, highly productive airliners, which were later to see almost universal service. The British, French, Germans, and Russians all produced four-engine bombers of varying efficiency, but again it remained for the U.S., in the latter '30s, to develop and begin production of the airplane that was to become the standard daylight strategic bomber of World War II.

In World War II the production situation of the earlier war was reversed; this time it was the United States that predominated among the Allies. From a 1939 base of fewer than 6,000 planes a year, the U.S. industry, once again bolstered by nonaviation firms, doubled production in 1940 and redoubled it in 1941 and 1942. In the six-year period 1940 through 1945, the U.S. turned out an incredible 300,718 military aircraft.

The transition to jets

The postwar years were marked by a rapid shift to jet propulsion as the power source for military aircraft. The first successful turbojet engine had been tested in Britain in 1937 and a rival engine first flown in Germany in 1939. Both nations introduced the jet in war service, though not in significant numbers, and wartime development paved the way for the great transition of the 1940s and 1950s. Jet power was applied first to military fighters, then to bombers, finally to commercial airliners. Military aircraft progressed from the subsonic (below the speed of sound), 600-mile-per-hour level of the late '40s to supersonic flight in the early '50s, to double sonic (twice the speed of sound; Mach 2) capability in the late '50s, then to Mach 3, which represented the upper speed limit of the fastest operational aircraft in the early 1970s. Each increment of performance gain brought an attendant increase in aircraft complexity, producing an across-the-board impact on the manufacturing process, from tools, materials, and facilities to composition of the labour force. At the same time, it started a trend in which research and development became a much more significant part of the total industry effort. The most notable effect, because greater performance permitted fewer airplanes for the assigned mission, was the transition from mass production of many types to short-run production of fewer but much more complex types. Never since the introduction of jets has aircraft production approached World War II levels. The best post-World War II year, in terms of combined military- and civil-aircraft production, was 1966; total output in that year came to barely 20 percent of the record wartime year.

While the jet transition was moving into high gear, the industry fell heir to a new responsibility, the development and production of guided missiles. It was a logical inheritance, because some of the missiles were aerodynamic, lift-supported like an airplane, and the others shared with the airplane at least a similar set of major components—an airframe, a propulsion system, a method of guidance and control, and a payload.

Wartime Germany had pioneered long-range missiles, introducing as operational weapons two types of surface-to-surface missiles and developing a wide range of other types. German experience provided a departure point from which the victorious nations advanced their own experimental programs in the latter 1940s. Initial missile production got under way in the early 1950s; first-generation missiles included artillery-like battlefield weapons, anti-aircraft weapons, pilotless tactical bombers, and air-launched weapons. In the United States and the Soviet Union the 1950s witnessed the state of production of long-range strategic missiles. Missile sophistication increased through successive generations of automated weapons in the 1960s, and production capability spread to other nations; by 1971 there were 14 missile-building nations, most of them confining their efforts to short-range battlefield- and air-defense systems (see also ROCKETS AND MISSILE SYSTEMS).

The impact of missilery on manufacturing processes was felt in two primary areas: completely new facilities were required, and the labour-force composition, already changing markedly as the result of the jet transition, was further altered as production of automated weaponry dictated greater emphasis on highly skilled workers—scientists, engineers, and technicians—and less on low-skill personnel.

Impact of Sputnik

Before missilery had reached full maturity, a third manufacturing responsibility was thrust upon the aircraft industry, when the Soviet Union sent the first man-made object into orbit, on October 4, 1957. Their success accelerated the American space effort. In 1959 the U.S. Aircraft Industries Association, Inc. changed its name to Aerospace Industries Association of America, Inc., which, the organization stated, "more properly reflected the changing nature of the effort." Initially confined to the United States and the Soviet Union, manufacture of space systems quickly became a significant portion of the total work load, particularly in the rapid evolution of ever more advanced manned spacecraft.

Space-equipment fabrication compounded all the previously established industrial change trends, particularly the trend toward heightened emphasis on research and development. Once simply a prelude to production of a flight vehicle, research and development soon became an integral part of the manufacturing process, extending to all of the major systems that comprise the flight vehicle, to the tools and methods of fabrication and assembly, and to the multitude of testing devices designed to insure an unprecedented level of equipment reliability (see also SPACE EXPLORATION).

II. Aerospace product line

MAJOR PRODUCT: FLIGHT VEHICLES

The product line of the aerospace industry is necessarily a broad one, because each of the primary products—that is, the flight vehicles—requires hundreds, thousands, or in a few cases millions of systems, subsystems, components, and individual parts.

Aircraft. Among the primary products, manufacture of aircraft has continued throughout the years as the major work load of the world's aerospace industry. A majority of those nations with such industries build only aircraft and aircraft engines. In the United Kingdom and France, missile–space efforts are relatively small in comparison to aircraft output. Even in the United States and the Soviet Union, both heavily engaged in missile and space work, aircraft manufacture holds first position by any measure of productivity.

Airplane builders fall into two categories: producers of general-aviation planes and producers of heavy aircraft. General aviation embraces planes and light helicopters used for private pleasure flying, personal transportation, business travel, and short-haul commercial operations, such as air taxis and commuter airliners with low take-off weights (specifications vary internationally). Included in the general-aviation bracket are specialized aircraft such as agricultural sprayers, acrobatic craft, sailplanes and gliders, air ambulances, pipeline-patrol craft, geophysical-survey platforms, and a broad variety of other civil applications. Also included are certain light military aircraft used for observation, liaison, and utility duty and incorporating only minor military modifications to their basic civilian-type design. Although many modern general-aviation planes are jet- and turboprop-powered, all are subsonic. General-aviation aircraft constitute the bulk of airplane manufacture in terms of units but little in terms of sales volume.

Types of heavy aircraft

The heavy-aircraft category includes nearly all military planes, commercial airliners, and large corporate aircraft adapted from airline designs. Among the more technologically demanding craft in this category are the supersonic military planes, huge military and commercial transports weighing up to 800,000 pounds (360,000 kilograms), and supersonic jetliners (production of the latter was confined in the early 1970s to the Soviet Union and the United Kingdom–France in combine). Also belonging to this group are specialized craft such as the vertical takeoff and landing (VTOL) vehicle, initially a military type but one with broad potential in commercial operations; large helicopters, employed as combat weapon systems, as intratheatre military transports, and as passenger-carrying commercial carriers; unmanned reconnaissance drones and target aircraft used for training of gun and missile crews; and ground-effect machines, or air-cushion vehicles, which operate close to the surface, supported by a column of high-velocity air generated by the machine's engines. The term heavy, as applied to this category, is not entirely a matter of weight but refers as well to the heavier demands of work load per unit to produce the stronger structures, more powerful engines, and more complex ancillary equipment required for the vehicles' design missions.

Guided missiles. Second to aircraft in worldwide industry productivity is manufacture of guided missiles and related products. From the standpoint of manufacturing work load, missiles can be grouped in two basic categories. The first group embraces battlefield, tactical, and air-defense weapons and short- to medium-range systems that carry either conventional high-explosive or low-yield nuclear warheads and that have speed capabilities ranging from subsonic to about Mach 3. Production of such weapons is relatively straightforward, within the capabilities of some of the less sophisticated aerospace industries of the world; in the early 1970s, companies in 14 nations were producing such missiles.

Far more technologically demanding are the strategic missiles that carry multimegaton-yield nuclear or thermonuclear warheads at hypersonic speeds (far beyond the speed of sound; above Mach 3) over distances ranging from more than 1,000 miles to intercontinental. Although it is a defensive rather than an offensive system, the antiballistic missile must be included in this category because of its complexity. By the start of the 1970s, only the United States and the Soviet Union had demonstrated capability for producing the antiballistic missile; in addition to these two, France and the People's Republic of China produced strategic missiles.

Strategic missiles and space-launch vehicles

A close relative of the strategic missile is the space-launch vehicle. (In space terminology, the launch vehicle is the rocket system that lifts a payload into orbit; the payload itself, whether manned or unmanned, is called a spacecraft, and the combined launch vehicle–spacecraft is known as a space vehicle.) Only in the United States and the Soviet Union have launch vehicles been produced on a quantity basis. The Soviet Union has followed the practice of modifying long-range missiles for space-launch work; the U.S. used converted missiles for the majority of its space launches but additionally developed families of launch vehicles designed from the beginning for space-launch duty. France, Britain, China, Japan, and the European Launcher Development Organization, a consortium of several European nations plus Australia, have all built launch vehicles on a limited basis.

Another missile-related product is the sounding rocket, known in Europe as the probe rocket. Used for upper-atmosphere and near-space research, the sounding rocket is a small launch vehicle that does not orbit a research payload but sends it instead on a brief, parabolic flight above the atmosphere and back to Earth. Sounding rockets generally have one to four rocket stages that fire in sequence to achieve the altitude desired for the experiment; payloads usually run from 50 to about 500 pounds (23 to 230 kilograms), and altitude capability ranges upward to about 1,000 miles (1,600 kilometres). Many countries are engaged in sounding-rocket programs, but the vehicles are primarily of American origin.

Spacecraft. The third major area of aerospace-industry effort is fabrication of spacecraft. Spacecraft are individually tailored to specific research programs and, therefore, do not fall into logical groupings of systems with similar construction characteristics. Unmanned spacecraft, called satellites when they operate in Earth orbit and probes when launched on a deep-space trajectory, have weighed from three pounds (1.4 kilograms) to almost 19 tons (17,000 kilograms) and have been configured in so many different forms that no basic description is possible. Manned spacecraft, naturally, impose far greater fabrication problems than unmanned systems, because of the equipment necessary to sustain human crews in space and bring them back to Earth. Modern manned spacecraft, weighing up to 100 tons (91,000 kilograms), are the most complex of all aerospace vehicles, and their production has been confined to the United States and the Soviet Union. In addition to these two nations, unmanned spacecraft have been built and orbited (in most cases by means of U.S.-provided launch services) by France, the United Kingdom, Canada, Italy, Australia, Germany, Japan, China, and the European Space Research Organization, a ten-member combine. A number of other countries have built individual systems and experiments for cooperative space programs sponsored by the United States and the Soviet Union.

Propulsion systems. Rounding out the primary-product line are the propulsion systems for all of the flight vehicles described. There are three basic types of power plants: piston, or reciprocating, engines; turbine engines; and rocket engines. No longer used in current-model heavy aircraft, the piston engine still has wide utility as a power plant for light single- and twin-engine general-aviation planes (see also GASOLINE ENGINE). The turboprop (turbine spinning a propeller) finds its primary applications in light–medium commercial and business transports and in military planes when greater range is more desirable than speed. The turboshaft (turbine driving a rotor system) engine is becoming the standard propulsion system for the larger helicopters, particularly those of the twin-engine variety. The pure jets, turbojet and turbofan engines, power all supersonic aircraft except research vehicles and a wide range of subsonic types from light single-engine trainers to the mammoth jumbo jetliners (see also JET ENGINE).

Types of propulsion systems

The primary employment of the rocket power plant is in propulsion for guided missiles and space-launch vehicles. There are two basic types of rocket engines: solid-fuelled and liquid-fuelled. Because of the requirement for long storage, the great majority of missiles are now pow-

ered by solid-fuel systems, as most liquid fuels have to be stored at cryogenic (very low) temperatures. Conversely, most space-launch vehicles employ liquid-propellant rocket systems, although large solid rockets are frequently used as boost-stage auxiliaries for greater thrust in the initial phase of a launch. An additional usage of the rocket engine is for spacecraft on-board propulsion, permitting changes of orbital path. These engines, relatively low-thrust systems, are used only on manned spacecraft and some of the most complex unmanned craft and consequently do not figure prominently in the production work load. Engine manufacturers also produce, in limited quantities, a variety of specialized rockets, including small monopropellant or bipropellant thrusters for adjusting the attitude of an orbiting spacecraft; solid rockets for explosive separation of launch-vehicle states; rockets for high-speed track-borne research sleds; stabilizing thrusters to control roll in missiles and launch vehicles; and aircraft auxiliary rockets.

Nations with substantial aerospace-engine industries, including those that build other nation's designs under license, include the U.S., the Soviet Union, France, Britain, Australia, Canada, China, Czechoslovakia, Germany, Israel, Italy, Japan, Poland, and Sweden.

OTHER PRODUCTS

Secondary aerospace products. The secondary-product line of the aerospace industry is composed of the great many items of on-board equipment required by the various design missions of the flight vehicles. An airframe and a propulsion system sufficed for the initial flights of the Wright brothers, but the modern aerospace vehicle admits of no such simplicity; depending on design purpose, a flight vehicle may have thousands of separate functional equipment units. The list of individual items is too lengthy for even a catalog listing, but a sampling of the more important systems and subsystems illustrates the breadth of the secondary-product line.

Aircraft equipment. This includes aerodynamic controls and actuating devices; engine controls; cockpit instruments and displays, including instrument-landing systems and other aids to bad-weather approaches; communications systems; radar for weather avoidance, target acquisition, search missions, and identification signals to ground stations; navigation equipment other than basic instruments, including radio, optical, and electronic systems and, in the most advanced aircraft, computer-directed inertial navigators; air-conditioning and pressurization equipment; landing-gear brakes and steering controls; data recorders; trouble-warning systems; radar, photographic, infrared-imaging, and other sensory devices for intelligence gathering, together with on-board intelligence-processing equipment; armament for military planes, including armour systems, launchers for airborne missiles, high-speed cannons and gun turrets, bomb and rocket fuses, and arming devices; electronic countermeasure equipment for confusing enemy defenses; engine starters and auxiliary power units for on-board electricity; seats, galleys, and passenger-service equipment for commercial airlines; and escape devices such as parachutes, surface-survival aids, and, for higher speed aircraft, ejection seats and separable cabins.

Missile and launch-vehicle equipment. This includes automatic guidance systems, ranging from heat-seeking homing devices to computerized navigation systems; on-board power systems; controls, including actuating devices for course-changing fins, thrusters, or gimballing engines; communications systems, including receivers for reception of command signals from Earth and transmitters for relaying vehicle-condition data to Earth, sensors to assay vehicle conditions, and trouble-warning devices; abort–destruct systems; engine-ignition and cut-off systems; automatic fuel-flow systems for liquid-fuelled vehicles; tank-pressurization systems; warhead systems; and, for strategic missiles, re-entry vehicles protecting the destructive payload from the friction heat of re-entry into the atmosphere.

Spacecraft equipment. This category includes an extremely broad variety of experiment packages, including instruments ranging from simple micrometeoroid counters to complex automatic telescopes; communications systems, including antenna arrays, telemetry transmitters for reporting data to Earth, recorders to store data for later transmission, encoders to translate data into coded signals, receivers for acceptance of Earth commands, decoders to interpret command signals, and command-distribution devices for translating commands into compliance; power sources, including solar-cell arrays, batteries, fuel-cell power plants, and nuclear generators; stabilization systems and station-keeping devices to maintain a desired spacecraft position or attitude; spacecraft-condition sensors; radio and optical beacons to facilitate Earth tracking; ejection and descent systems for payloads recoverable from orbit; photographic, radar, infrared, and other types of sensory devices for military intelligence gathering and for civil use in meteorological and Earth-resources-survey spacecraft. Additionally, for manned spacecraft there are environmental-control systems; guidance and navigation systems, including sophisticated computers for acceptance and processing of Earth-acquired data; cabin instrumentation and displays, including trouble-warning systems; special-purpose radars, such as those used for in-space rendezvous or lunar landing; pressure suits for extravehicular activity; and galleys, water dispensers, and waste-management systems designed for operation under zero-gravity conditions (see also LIFE-SUPPORT SYSTEMS).

Experimental equipment

Tertiary aerospace products. The industry's tertiary-product line includes those ground-based items necessary for support of flight vehicles, either directly, such as engine starters, or indirectly, such as training devices. A major area of effort is the production of ground radars and antennas with associated data-processing systems (see also RADAR). This equipment is employed for air-traffic control and airport control-tower operations; for detection and tracking of potentially hostile flight vehicles; for the Earth-command type of missile guidance and for interception guidance of air-defense aircraft; and for tracking spacecraft. Another major area is production of automatic checkout equipment for complex aircraft, missiles, and space vehicles; such equipment electronically "interrogates" each of the vehicle's systems and verifies that the system is in proper working condition or pinpoints faults for quick correction. A third area involving considerable work load is fabrication of simulation systems, used for training aircraft, missile and spacecraft crews, and research and development processes. Among other examples of ground-based aerospace equipment are radio transmitters for aircraft navigation; airport lighting and landing aids; ground-based air-conditioners, power units, and airplane-servicing equipment; transporters, erectors, launchers, and fire-control systems for mobile battlefield missiles; and the systems that make up a space mission-control centre, such as communications circuitry, computers, and display consoles.

Ground-based equipment

Nonaerospace products. Nonaerospace products and systems, general-usage offshoots of aerospace technology, constitute a fourth area of industry work load, not significant as a proportion of total work load except in the United States and presumably in the Soviet Union. The United States industry has for several years been engaged in a strong effort to transfer its wealth of technology to general use, in an attempt to further diversify its product line and minimize dependence upon government business. By the early 1970s this program had met with considerable success; the list of nonaerospace items ran into the thousands, and annual sales volume approached the level of $3,000,000,000. Examples of new applications of aerospace-developed technology include many medical aids, ranging from a thimble-sized cardiac-pacemaker power source to a computerized hospital-management system; in transportation, hydrofoil ships and boats, vastly improved railroad cars, truck and bus transmissions, motor-vehicle traffic-monitoring systems, inflatable crash-bag systems for autos, and turbine propulsion systems for both land and water vehicles; in con-

struction, prefabricated low-cost housing and lightweight, superstrong structural materials; in electric power, a number of types of nuclear and nonnuclear power-generating systems; and, in environmental control, a broad range of antipollution systems.

III. Research and development

The ultimate aim of aerospace research and development is the creation of flight vehicles more advanced than their predecessors, but so succinct a statement of goal oversimplifies the enormous scope of research and development activity. It is a fact of aerospace life that expansion of the operating capacity—the "envelope" of a flight vehicle—almost invariably entails corollary advances in on-board systems, ground-support equipment, and even the methods of construction. Frequently, desired performance gains dictate not merely improvement of existing ancillary equipment but entirely new inventions. Aerospace research and development is a continual process of across-the-board thrusting into the unknown and frequently moves beyond the traditional areas of investigation into new byways embracing practically all of the technological disciplines.

Types of research and development

Aerospace research and development is of four types. Basic research involves investigations that have no immediate practical objective; it is the advancement of knowledge for potential rather than direct interest. Applied research is investigative effort aimed at the practical application of knowledge. Development, by definition, is "the use of scientific knowledge directed toward the production of useful materials, devices, systems or methods, including design and development of prototypes and processes"; it is the translation into hardware of the results of applied-research studies. Operational research is concerned with the solution of problems encountered by flight vehicles already in regular service. The aerospace industry engages in a certain amount of basic research, but it is negligible as a percentage of total research and development work load. Operational research is conducted to a greater degree, but the primary focus in the industry is on applied research and development related to the betterment of existing products and the emergence of completely new products.

The level of research and development activity naturally varies a great deal among companies and countries, the determining factor being the types of flight vehicle to be built. Light, general-aviation aircraft, for instance, are not overly demanding of research and development. Like automobile manufacturers, light-plane builders offer their products in a number of price categories and cost is a restricting influence on size, weight, and performance. General-aviation manufacturers, whether airframe, engine, or equipment builders, do conduct considerable research, but it is aimed at small increments of improvement that can be obtained without cost increases that would take the plane out of its competitive category.

NEW COMMERCIAL AIRCRAFT

Development of a new commercial airliner demands extensive research and development, but the requirement is substantially reduced by the fact that the jetliner operates in an already-explored flight regime. Traditionally, military aircraft reach a given plateau of performance a decade or more before the airliner achieves the same capability. Thus, designers of commercial aircraft have available at the outset a vast storehouse of aerodynamic, propulsive, and other knowledge relative to the flight regime in which they propose to operate. This is not to suggest that they can bypass those areas in which prior research has been accomplished; rather, they can use the knowledge available as a departure point for a research and development program in which the emphasis is on utmost efficiency within a known flight bracket rather than initial penetration of an unknown regime.

Jetliner developmental teams focus on two primary areas of research and development: safety and economic viability. Safety involves such major considerations as stronger structures, ultrareliability of engines and on-board equipment, and provision of the most advanced navigation and all-weather landing systems. Economy of operation embraces capacity; maximum flight efficiency, which is translatable into greater range and payload; maximum propulsive efficiency, for reduced fuel costs; lowest possible turnaround time for best utilization of the airplane; longer equipment life for extended periods between overhaul; and ease of maintenance.

Expanding the aircraft envelope

The industry's major research and development effort is expended on those programs that seek major expansion of one or more facets of a flight vehicle's envelope, the bounds of its performance capabilities. More often than not, a speed greater than its predecessor's is the primary performance gain sought for a new vehicle, but the research effort may also be directed toward other elements of improvement. The goal, for instance, might be an increase in range or payload, as is the case in global-ranging military bombers and transports; it might be manoeuvrability, a vital consideration in fighter-aircraft design; it might be greater destructive capability, as in multiple-warhead strategic missiles; it might be duration, as in an extended-stay orbiting vehicle; it might be extremely high altitude, for a reconnaissance airplane; or it might be a combination, as in the case of a new bomber under development, of supersonic speed for long periods at minimal altitudes, a particularly demanding area of research because of different shock-wave characteristics at sea level as opposed to those of high altitude. There are other factors involved, and rarely are the sponsors of a new project willing to settle for improvement in only one area.

High-technology programs of this nature are almost always government sponsored. Because they probe unknown areas of flight, they require exceptionally high funding outlays, millions of research and development man-hours, and an elaborate complex of research and development facilities, some of which may have to be created especially for the particular project. A brief outline of the step-by-step process in the development of a new high-technology vehicle—a high-performance military aircraft, for instance—serves to illustrate the extraordinary human skills required and the complex array of tools and facilities needed to bring such a program to fruition.

Computer design. The design phase is largely theoretical. Drawing upon the general fund of aeronautical knowledge and their own preliminary research, scientific and engineering specialists produce a mathematical model of an airplane that they think will meet all of the project sponsor's specifications. The team's primary tool is the electronic computer, used in a variety of ways. It provides rapid and accurate answers to the millions of mathematical calculations required. It stores in its memory banks the design input fed into it and creates an electronic image of the airplane. Coupled with an automatic plotter, it displays on a TV-like screen pictures of the complete airplane or any section of it, such as the wing, flaps, tail section, or nacelles, affording the team a visual presentation of current analysis. Programmed with atmospheric data, it can "fly" the hypothetical airplane and report to the team the aerodynamic response of the design to any condition of flight, from takeoff to landing, from straight-and-level cruise to combat manoeuvres. Through a lengthy series of such simulated flights, the team refines its design, until the electronic airplane meets or exceeds all specifications.

Wind-tunnel testing. The next step is wind-tunnel testing of a physical model, a carefully constructed scale replica of the computer-aided design, for verification of the performance predictions. The wind tunnel, which actually predates powered flight by more than 30 years, is an apparatus in which air is blown over a model mounted in a test section, creating an effect comparable to airplane flight. There is wide variance in the size of wind-tunnel test sections, depending upon airflow-velocity requirements. Some low-speed tunnels have test sections large enough to accommodate a complete small airplane or a wing–nacelle section of a larger plane; in a higher speed system the test model may be only the size of a desk model, and in the very high velocity tunnels used

for missile and space testing the model may be as small as a match. Tunnels are classified according to airflow-velocity range in five categories: subsonic, up to Mach 0.8, or 80 percent of the speed of sound; transonic, Mach 0.8 to Mach 1.2; supersonic, Mach 1.2 to Mach 6; hypersonic, Mach 6 to Mach 12; and hypervelocity, above Mach 12.

Methods of operation of wind tunnels

There are two basic methods of feeding air to the test section. In the continuous-flow tunnel, airflow is generated by a battery of fans or compressors, forced through the tunnel to the test section, then directed through a curving tunnel section back to the starting point for reuse. The alternate method involves storing high-pressure air in a tank, releasing it by valve controls to pass through the tunnel and test section into a capture tank. This type of system, the intermittent-flow tunnel, is generally used for very brief bursts of high-velocity airflow.

Some wind tunnels are almost as complex as the flight vehicle under development. They have variable throats to permit changing airflow velocity, conditioning systems to change air characteristics for simulation of different altitudes, and heating devices to simulate the friction heating encountered by high-speed flight vehicles. Mounted within the model are tiny transducers that acquire a variety of data and transmit it to recording systems outside the tunnel. Specially designed colour cameras photograph shock waves and other airflow phenomena. The computer digests the data from the test run, analyzes it, and presents a report on the flight vehicle's performance.

Most major manufacturers have some type of wind tunnel, usually of the intermittent-flow type, and for more demanding programs they have access to the complex government-operated supertunnels that no single company could afford. After a series of wind-tunnel "flights," the design may be further refined to incorporate the findings of tunnel research; then the airplane is ready for prototype construction.

The prototype. In the prototype-construction phase, emphasis shifts from research to testing. A customary procedure is to build one or more static test airplanes, designed not to fly but to verify the structural integrity of the design. This requires another array of facilities, including ovens for applying high temperatures to materials; acoustic chambers, in which engineers can study the effect of high-frequency engine noise on structures; rigs for measuring landing impacts; and rapid-cycle vibrators for investigations of wing-panel flutter. A major facility is the load-test rack. A flight load is the force experienced by the airplane in its various manoeuvres; designers provide for an ultimate load factor, which represents the greatest load the airplane may encounter in the most violent manoeuvre plus a comfortable safety margin. The load-test rack consists of a series of instrumented hydraulic jacks that push, bend, and twist the airplane to the point of failure, verifying that the ultimate load factor called for in the design has been met in construction.

While the prototype airframes are being built, the manufacturers of on-board equipment are going through a parallel process to devise and test the airplane's many systems and subsystems. The broad variety of ancillary equipment precludes description of the testing process, since it differs with each system. Generally, equipment is checked out, repetitively and exhaustively, by a battery of electronic test equipment tailored to the system being tested. In the laboratory, the test set electronically monitors the system as it is run through its performance cycle and signifies okay or detects and isolates faults for correction. In many cases, complete systems are further checked in altitude chambers that simulate the environment in which they will later be operating. Engines are tested in the propulsion research equivalent of the wind tunnel, the test cell capable of simulating the conditions of flight; test engines are heavily instrumented, and the data findings are transmitted to a computer for processing. To qualify for installation in a prototype airplane, a new engine is normally put through several hundred hours of cell testing, embracing the entire range of speed and altitude capacity of the airplane for which it is intended.

Flight testing. As a general rule, prototype flight testing is conducted at government-operated installations in lightly populated areas, due to the possibility of accident to an experimental airplane. Flight testing is necessary to validate what has been to this stage really only theory, but so thorough and extensive are modern procedures of computerized design and tunnel testing that the flight-test phase rarely dictates a major design change. Rather, it serves to debug the airplane, to correct the many minor faults that somehow find their way into a new airplane despite the most painstaking efforts to avoid them.

The modern test pilot

Modern flight testing bears little resemblance to the once-familiar motion-picture portrayal of the daredevil pilot trying to fly the wings off the airplane. The modern test pilot is a thoroughly experienced flyer who is also an engineer; he is selected for his flying ability and for his ability to observe and report in technical detail on the handling qualities and peculiarities of his airplane. He is backed by a forest of sensitive test instruments, gathering volumes of data, which are relayed by radio telemetry to computers on the ground. The flight program is carefully planned, again with the help of a computer. It begins with ground-taxi tests and progresses to brief flights at the lower end of the performance spectrum, then to longer flights of gradually increased performance until the limits of the design envelope have been reached or surpassed. The test program for an advanced airplane may take several months and complete evaluation of the data additional weeks or months.

MISSILES AND SPACECRAFT

General considerations. The research and development effort that goes into development of missiles, launch vehicles, and spacecraft parallels that of the airplane in the design and ground-test stages but differs when the flight-test stage is reached. The shorter range battlefield and tactical weapons pose no major testing problems; they are fired more or less like artillery systems at large-area land ranges. Major launch vehicles and strategic missiles are more demanding as regards test techniques. The absence of man on board, the great expense of a single launch, and the inability to recover and re-use the test vehicle all combine to exert two major influences: more elaborate instrumentation, both in the vehicle and on Earth, and extraordinarily intensive preflight checkout to prevent a costly abort.

Major world testing ranges

Additionally, the distance covered in a major unmanned flight-vehicle test, which is several thousand miles in the case of an intercontinental ballistic missile, requires the use of spread-out test ranges with downrange tracking and data acquisition stations. These ranges are necessarily government operated, but generally contractors have their own facilities at the launch centre for prelaunch testing. In the United States the U.S. Air Force Eastern Test Range includes both the John F. Kennedy Space Center and the Cape Kennedy Air Force Station, launch bases for civil space flights and military missile testing; surface stations extend 10,000 miles down the South Atlantic. A second major United States facility, the Western Test Range, centred at Vandenberg Air Force Base, California, extends into the Pacific. The Soviet counterpart of Cape Kennedy is the missile-space-launch centre at Tyuratam, on the Aral Sea, with impact areas for long-range flights in the Pacific Ocean. The Soviet Union operates two other major facilities, one at Plesetsk, near the White Sea port of Arkhangelsk, the other at Kapustin Yar, north of the Caspian Sea. France operates a range that extends into the Atantic from a base at Kourou, French Guiana. Other major ranges are operated by Australia and Japan, with launch centres at Woomera and Kagoshima, respectively. There is no information available on the Chinese range facilities.

Unmanned spacecraft are unique in that they undergo no prototype test flights; if their launch vehicles perform successfully, they go into orbit, so they carry their full complement of operational instrumentation. In a sense, then, all unmanned spacecraft are test vehicles. since

they report details of their own operation as well as scientific-data findings, providing the basis for design refinement in later models of the same family. The substitute for prototype testing is Earth-based simulation, usually in two types of simulators: the space simulator, which duplicates all the environmental conditions in which the spacecraft will operate, and the mission simulator, which permits conduct of the entire range of manoeuvres and systems operations that would be performed on an actual flight.

Manned spacecraft are not only prototype tested but are subjected to the most rigorous ground-test program experienced by any flight vehicle. The program defies detailed description, since it involves literally millions of separate tests. Some idea of the intensity of the test effort and the broad variety of research and development facilities needed may be obtained by a recounting of the major ground-test operations of a manned spacecraft, such as the United States' Apollo mooncraft.

Testing. *Aerodynamic testing.* Although most of its lunar mission takes place outside the atmosphere, Apollo does fly within the atmosphere briefly en route to orbit and during re-entry and parachute descent to the ocean. To verify design predictions as to aerodynamic loads, acoustic effects, friction heating, and lift-to-drag ratios at the hypersonic speed of re-entry, Apollo was put through more than 7,500 hours of wind-tunnel testing.

Vacuum chamber tests. In these the environmental-control system was tested at conditions ranging from sea-level pressure to hard vacuum.

Impact tests. In these a full-scale command module was slammed into a water tank by a huge pendulum arm, simulating various angles of splashdown and impacts much greater than would be encountered on a normal descent.

Space-simulation tests. In these the complete spacecraft was placed in a large facility capable of simulating the total space environment, including hard vacuum, solar radiation, and temperature extremes, to determine effects on systems and materials.

Oven–freezer tests. In space, Apollo is at times positioned so that one side is in the undiluted sunlight while the other is in cold shadow. For simulation of these simultaneous extremes of temperature and observation of structural effects, the spacecraft was electrically roasted on one side at 600° F (320° C), while the other side was sprayed with liquid nitrogen at −320° F (−200° C).

Load tests. In a facility that is the spacecraft equivalent of the hydraulically operated airplane load test rack, Apollo was subjected to immense loads far greater than any normally anticipated.

Dynamic tests. These include acoustic, vibration, shock, and acceleration investigations. A huge "cocktail shaker" rapidly shook the complete spacecraft to simulate the vibration and buffeting of the launch phase.

The above are a few samples of a broad range of ground tests conducted repetitively over a period of several years. Running concurrently was a series of unmanned suborbital and orbital flights. For the initial series the contractor built a number of test vehicles known as boilerplates, replicas of the actual spacecraft in size, shape, structure, mass, and centre of gravity but not equipped with the myriad functional systems. The boilerplates were used on ten flights to test the spacecraft's compatibility with its launch vehicle, the launch escape system, and the general stability and structural integrity of the spacecraft in orbital flight. The boilerplate tests were followed by six additional flights in which completely equipped spacecraft were flown in orbit on an unmanned basis to check out all systems under actual space conditions and to verify the performance of the all-important re-entry heat shield. For these flights, the stand-in for the human crew was a robot controller capable of directing systems operation either automatically, by preprogrammed instructions, or by command from Earth monitors. The meticulous preparation for each flight, coupled with the evaluation time for the vast amount of data acquired, stretched the 16 unmanned flights over a period of four and a half years.

The boilerplate test series

OTHER AREAS OF RESEARCH

In addition to project-oriented research and development, the effort involved in a specific flight-vehicle program, the aerospace industry conducts a tremendous amount of research and development of a general nature designed for competitive readiness for the next generation of aerospace products. Among the more active areas of general research are:

Propulsion. Anticipating continuing demand for greater thrust, particularly in the airplane field, engine manufacturers are designing and developing new families of superpowered propulsion systems. In addition to thrust ratings, research is aimed at lower engine weight in relation to power output, better combustion efficiency, and further extension of the time between overhauls. A special line of research in recent years has been concerned with reduction of engine noise and emissions.

Electrical power. Continually increasing flight-vehicle performance means more on-board systems and greater capability of the individual system, hence a requirement for more electrical power to operate the systems. Manufacturers are paying particular research attention to fuel-cell power plants and thermoelectric converters, along with a variety of conventional generators of more advanced design.

Materials. Probably the greatest single influence on aerospace design is weight, and researchers are constantly looking for new materials that offer weight reductions without sacrifice of strength. The quest extends to materials for all parts of all types of flight vehicles, but the greatest emphasis is on temperature-resistant materials for high-performance vehicles. Conventional materials weaken under extended exposure to combustion-induced heat, solar-induced heat, friction-induced heat, and hot–cold extremes. Seeking simultaneous improvement in temperature resistance and strength-to-weight ratio, researchers are exploring in particular such exotic metals as columbium, molybdenum, and tantalum; ceramics; composites, combinations of two or more materials in which the resultant has better properties than the individual constituents; and vehicle-skin coatings for prevention of oxidation.

Significance of weight in aerospace design

New systems. "New," in this sense, means not developmental improvements of existing equipments but completely new systems that do not exist in operational status and for which a need is clearly indicated. An example is an automatic collision-avoidance system for aircraft.

Miniaturization. A traditional method for reducing weight and at the same time cutting down on systems volume requirements is making things smaller. The missile–space era brought such great strides in this area that manufacturers call it microminiaturization. This type of research is continual, first because the trend in high-performance flight vehicles is generally toward more equipment and second because the initial version of a system, designed primarily for performance, usually leaves considerable room for later miniaturization.

New manufacturing processes. A large portion of general research and development work is devoted to research in new manufacturing processes, in order to combat ever-mounting costs, reduce fabrication time, and assure greater product reliability. Additional effort goes into development of machines and techniques for shaping and forming the new materials being introduced to production status, which almost invariably are more difficult to handle than their predecessors.

COOPERATIVE PROGRAMS

As aerospace flight vehicles and systems grow in performance capability and consequently in complexity, research and development costs mount steadily. Even the United States and the Soviet Union, with their vast resources, are finding it difficult to finance essential research and development, and among the smaller nations financial limitations would make it necessary to bypass promising areas of research and development were it not for the cooperative-program approach.

In western Europe the cooperative approach has been growing for two decades and has reached significant pro-

portions. In these programs, two, three, or more nations pool their financial resources and their facilities on projects that would be beyond the means of any one of them. In some cases, developmental responsibility and financing are divided equally among the participants; in others, one nation assumes the major burden but is backed by associate participants. Project management may be assigned to a single major company, or the project may be directed on a consortium basis, with alternating national leadership.

In this fashion, European countries are accomplishing considerable research that would otherwise not have been undertaken. Several have banded together as the European Space Research Organization; a parallel group is the European Launcher Development Organization, which seeks to develop and produce space-launch vehicles. Other projects are more directly oriented toward commercial markets; the British–French Concorde supersonic plane is the prime example, but there are many others, involving jetliners, military aircraft, helicopters, missiles, turbine engines, and a communications satellite for a contemplated European network. Participants in cooperative programs include France, the United Kingdom, the Federal Republic of Germany, Italy, Sweden, Belgium, Switzerland, The Netherlands, Spain, and Denmark; Australia, which supplies a launch range, is also a participant in the European Launcher Development Organization.

The Concorde project

IV. Manufacturing facilities and techniques

No one company builds an entire flight vehicle, not even the simplest light plane, because of the exceptional range of skills and facilities required. Aerospace manufacturers are specialists, operating as either independent companies, divisions of very large corporations, or bureaus of a nationalized industry (among the major producers the Soviet Union is completely nationalized, France almost exactly half and half as regards the personnel division between the *secteur nationalisé* and the *secteur privé*, and the United States and United Kingdom entirely private industry). The primary areas of specialization are airframes, propulsion systems, airborne electric and electronic systems, other airborne systems, and ground-based equipment. Within these broad areas there is sharper focus of specialization; for instance, one electric–electronic company has separate divisions manufacturing guidance and navigation systems, control systems, electric-power systems, instruments and life-support equipment, and scientific-data-acquisition systems.

A major aerospace production program may involve literally thousands of individual suppliers operating as subcontractors to a team leader, usually the airframe builder, known variously as the prime contractor, principal constructor, "mother" manufacturer, systems integrator, or design bureau, the last a Soviet term that embraces not only design but also supervision of the complete manufacturing process. Subcontracting covers not only the on-board equipment but, in most major programs, the airframe itself. Production of the British–French Concorde serves as an example: the airframe is broken down into 21 separate components built by 11 different companies and divisions, then delivered to the team leader's plant for mating. Work-load sharing of this nature takes advantage of additional pools of skill and facilities and helps compress production time. On commercial projects it is also employed as a method of spreading the requisite financing for materials, tooling, and other production preliminaries.

BASIC FABRICATION TECHNIQUES

The variety of the product line, from individual parts to major systems to the complete flight vehicle, precludes description of a typical manufacturing plant, since each factory is tailored to the items produced. In general, the primary work areas of any plant are the fabrication facilities, in which the individual parts and components are made, and the assembly facilities, ranging from moderate-size rooms to huge, high-bay buildings covering several acres.

Metal processing. In the fabrication stage, each of the thousands of parts must be precisely shaped, and there are a number of ways of doing this. Fabrication tools and techniques fall into two major categories: chipless operations, performed without cutting tools, and material-removal operations. Chipless operations include die-casting, or pouring molten metal into a mold; molding, which is casting under pressure; and a number of methods of pressure forming. Aircraft skins, for instance, are made by rolling machines that squeeze and flatten hot sheet metal to the desired thinness; other parts are made from metal sheet, plate, foil, ingot, and bars by stamping; by forging, or pressing the metal over dies; by extruding, or forcing heated metal through a die, like squeezing toothpaste; by stretch forming, giving the metal a permanent bend by stretching it over a curved die; or by braking, pressure bending a clamped metal workpiece. A relatively new method is explosive forming, in which an explosive charge is detonated against a water-submerged workpiece and the resulting blast slams the metal against a shaped mold (see also MATERIALS PROCESSING).

Metal-removal operations

Metal-removal operations are primarily those in which metal is trimmed away from a workpiece by tools such as milling machines, lathes, shearing machines, borers, punch presses, knives and saws, sanders, grinders, and polishers. In addition to these cutting and friction tools, metal removals are accomplished by electrical devices, such as an electrical-discharge borer, and by chemical means; in the latter method, an acid bath eats away the excess metal while a coating protects the desired shape. In either chipless or metal-removal operations, the particular technique selected for a given part is dependent upon a number of factors; for example, the type of basic material, the intricacy of the shape, requisite strength, and, when it becomes a part of the flight vehicle, the degree of exposure to load forces and temperatures.

Joining methods. Individual parts, once fabricated, must be mated, or joined, with other parts. The most familiar method is mechanical joining by means of rivets, bolts, screws, pins, and other special-purpose fasteners. In cases in which a tighter seal is necessary or where the assembly will be subjected to great loads, manufacturers employ welding, which in most cases is accomplished by melting the metal at the joint, applying molten metallic filler, and allowing the joint to cool. Electric-arc welding, in which heat is supplied by an electrical circuit created between the workpiece and the welding tool, is in wide aerospace usage. Other methods include cold welding, or joining by pressure without high temperature, and forge hammering with heat application. A more advanced technique, diffusion bonding, involves the application of both heat and pressure in a nitrogen atmosphere, creating an exceptionally strong, lightweight joint. Another joining method in wide use is adhesion bonding, or simply gluing the parts together by means of supercapable adhesive coatings, such as plastic resins.

The honeycomb sandwich

In the production of components that must bear high loads yet be as light as possible, aerospace fabricators have evolved a technique for changing the properties of a material by engineering. The most notable example is the so-called honeycomb sandwich. The core of the sandwich is the honeycomb, a structure composed of row upon row of framed cells, or holes, resembling the honeycomb storage facility of a beehive. To each side of the core are bonded extremely thin sheets of metal, creating the sandwich, which is far lighter yet has greater resistance to bending than a comparable thickness of metal plate. Aluminum is the most used metal, in both the core and the facing sheets, but the technique is applicable to a great many metallic and nonmetallic materials. Sandwich construction is a major part of aerospace manufacturing, one used to some degree in almost every type of flight vehicle.

Most of the basic fabrication techniques have been employed for years in aerospace manufacture, and the difference today is in the degree of sophistication of the tools and techniques; like the product line itself, tools and techniques have progressed through several plateaus of technological advance. For example, simple hand-

torch welding would not suffice for the stringent joining requirements of a current spacecraft; it is necessary to enclose the weld station in a pressurized, temperature-controlled, contamination-free portable room. Similarly, while forming machines perform the same functions as their forebears, they differ considerably because of new and more difficult to handle materials, such as titanium, magnesium, and beryllium in airframe construction and the superalloys of iron, nickel, and cobalt in engine production. Additionally, the modern machine must work to incredible tolerances to produce parts of greater precision than the finest watch movement. Not long ago, machining to a thousandth of an inch was considered a tremendous accomplishment; some modern systems demand tolerances measured in millionths of an inch.

Use of electronic computers. The primary difference between today's manufacturing facility and its World War II or immediate postwar counterpart is the machining precision and efficiency made possible by the electronic computer. In a process broadly known as computer-aided manufacturing (CAM), computers are employed to direct the operations of machine tools. CAM embraces a number of types of computing equipment and various methods of controlling the machine. In general, the computer serves as a translator between the human programmer and the machine tool. The design of the part to be manufactured is reduced to symbolic computer language, and the programmer feeds the computer instructions as to the necessary machining operations. The computer in turn relays the instructions, in the form of command signals punched on tape, to an automatic controller, which directs the machining with extreme precision. Advanced CAM systems include multiple-operation machines, equipped with a battery of different tools; by computer-direction, the machine can perform any combination of operations in a prescribed sequence and even change the tools as required. Some huge tools can turn out large single-piece components and structures of intricate shape, eliminating the need for joining several individual parts. Computer-aided manufacturing shortens production time, thereby reducing cost, and in addition makes possible greater accuracy in parts fabrication.

The CAM system

PRODUCT ASSURANCE

Reliability is an important factor in the manufacture of any type of vehicle, but it assumes obvious urgency when the vehicle is one that flies. In aerospace production, exhaustive measures are taken to ensure utmost reliability and to guarantee that the completed vehicle or system or subsystem meets all the specifications prescribed by the end user of the product. The terms vary by company and by nation, but generally the collective precautionary measures are known as product assurance, which includes two subdivisions of effort: reliability engineering and quality assurance.

Reliability engineering, which begins in the design phase, essentially is a process of intensive study of alternative methods of performing a particular function and selection of the approach with the best predictable reliability. Reliability-engineering teams are frequently at odds with other designers, because the most reliable system is often one that demands compromise in such matters as cost and weight. An example is the practice of assuring reliability by means of redundant systems; a particular unit or system is provided an identical backup, together with a monitoring package that automatically activates the backup when it detects a malfunction of the primary. In some flight vehicles triple redundancy (two backups and a primary) is employed.

Reliability through redundancy

Quality assurance encompasses the entire manufacturing effort. It begins, as a management function of the project-team leader, with selection of the multitude of suppliers. The team leader studies the suppliers' own capability estimates, inspects their facilities, and investigates their history of performance before assigning a production task.

Quality-assurance groups focus a good deal of their attention on people. Despite the high degree of automation and the sophistication of fabrication tools, human effort still plays the dominant role in aerospace production. Quality assurance reaches into the processes of selecting skilled personnel and training employees. By lectures, bulletins, and posters on the potential consequences of carelessness, the quality-assurance group seeks to reduce human error. Some plants adopt extraordinary error-reduction measures; for example, some producers, to guard against the possibility of a potentially damaging tool being left behind in an engine assembly, provide workers with fitted toolboxes that are inspected after each shift.

The major aspect of quality assurance is quality control, the process of monitoring every step of the manufacturing cycle, from reception of raw material through delivery of the finished product. Quality-control technicians inspect new material coming into the plant to make sure that it is the specified type and has adequate properties. As the material is formed into parts and then subassemblies and assemblies, items that are considered critical are inspected after each work step, sometimes visually, sometimes by X-ray, microscopic, chemical, and ultrasonic devices that uncover invisible or subsurface cracks, voids, and other defects. Noncritical items, those whose failure could not influence the safety of the flight vehicle, are inspected on a spot-check basis.

Functional systems are subjected to repetitive tests after each workstep to insure that they operate exactly as they are supposed to when incorporated into the flight vehicle. There are hundreds of different types of testing equipments, each specially designed for a particular product, the bulk of them electronic testers that check out the operation of the product and report any deviation from specified performance. There are tests of individual parts, of assemblies formed by a number of parts, then of complete systems. Systems are checked out separately in their assembly facilities, later in their actual position in the flight vehicle.

ASSEMBLY METHODS AND FACILITIES

Assembly of an aerospace vehicle and the systems that go into it is a pyramidal process. The base of the pyramid corresponds to the mating of two or more individual parts designed to work together. Moving up the levels of the pyramid, these parts assemblies are joined with other assemblies to form subsystems, then systems. The terms system and subsystem are not sharply defined, other than that either is an assemblage of parts or equipment that *in toto* perform a particular function in the operation of the flight vehicle. Thus, in military terminology, a strategic missile is a system, and its inertial navigator is a subsystem; in the plant of the guidance-equipment manufacturer, however, the inertial navigator is considered a system, composed of such subsystems as the computer, the gyroscopic sensing platform, and the electrical-power source and circuitry. Few on-board systems are manufactured in a single plant; in most cases, parts and subsystems are produced elsewhere and delivered to the prime-system contractor for assembly.

At the top of the pyramid is assembly of the complete flight vehicle in the plant of the team leader, to which flow from a great many plants the airframe components or subassemblies, on-board systems and subsystems, and a variety of individual items that do not merit classification as systems.

The types of facilities employed in the assembly of on-board equipment differ considerably, as do assembly-line methods. Generally speaking, this type of aerospace product, small by comparison with airframe components, is assembled in low-bay buildings containing a number of separate work stations at which specific steps in the assembly process are accomplished. The growing assemblage is moved from station to station by conveyor or by wheeled cradling fixtures, frequently inspected or tested along the way. Some relatively simple equipment needs no assembly line; it is put together, inspected, and tested all at one work station.

In extensive use among aerospace manufacturers is the "clean room" type of facility, where assembly is conducted in a contamination-free atmosphere that rivals the surgical theatre. Clean-room assembly is imperative

"Clean room" assembly

for a great many systems that go into missiles, launch vehicles, and spacecraft (and, to a lesser extent, aircraft), complex equipments that are especially sensitive to environmental contamination; a particle of dust or a temperature outside certain rigid limits could cause a malfunction.

A typical clean room has an atmospheric-control system that rigidly controls temperature and humidity and bars entrance, by means of filters, of all but the tiniest mote of dust. Walls and ceilings are of one-piece plastic with no cracks where dust might collect and are washed and vacuumed daily. Maintenance cannot be done within the room: plumbing, wiring, and lighting are so arranged that maintenance can be handled in crawl spaces above the ceiling. The room has no sharp corners; they are rounded off to forestall dust collection.

Before entering, workers don special clean suits, including head covering and boots, and pass under an "air shower" that removes all loose particles of matter. The parts that make up the assembly are thoroughly cleaned and polished before delivery to the clean room, which they enter through an air lock.

Clean-room assembly is expensive and is used only when necessary, in most cases for the late stages or the final phase of the assembly procedure. The size of the room varies with the product; it may be from living-room size to a huge area considerably larger than a football field. Introduced in the latter 1950s with the development of sophisticated strategic missiles, clean-room assembly expanded over the years as systems became more complex.

Final assembly

The final assembly of the complete aerospace vehicle usually takes place in a large structure with a very lofty ceiling and no interior partitions or columns. Major aircraft require the tallest bays; some have tail sections reaching more than 65 feet (20 metres) above the floor. Additionally, the facility must accommodate a network of overhead rails over which ride heavy-lift cranes capable of moving large segments of the airplane.

Although flight-vehicle assembly bears little resemblance to systems assembly, the technique is the same in terms of movement through progressive work stations. At each work station, equipped with its own tools and workstands that provide access to all parts of the segment being worked upon, a certain number of assembly operations are performed according to a carefully considered manufacturing plan.

Assembly of an airplane begins with subassemblies. A typical subassembly for an airline-type plane is the rear fuselage section, itself composed of several segments. The segments are dollied or craned to the subassembly area, fitted together, and joined into a unit, and certain of the interior equipment is installed. In similar manner, plant teams put together other subassemblies; that is, other fuselage sections, wing sections, the tail section, and engine nacelle. The various subassemblies then join the main assembly line.

Mainline operations normally start with the joining, or mating, of two fuselage sections craned onto a supporting jig, or fixture. When all of the attendant joining and interior operations are completed, the embryo airplane moves to the next work station for mating with another subassembly. Light- and medium-weight aircraft may be moved on wheeled fixtures; heavier aircraft may be craned; and movement of modern supersize planes is accomplished by an adaptation of the air-cushion technique. The supporting fixture is a massive steel structure with a great many holes in its underside. Compressed air is pumped into the structure to escape downward through the holes. The thrusting force of the escaping high-pressure air lifts the entire structure and assembly off the floor a fraction of an inch, enough to permit movement of the assembly by an ordinary tractor. In this fashion, the budding airplane moves through a series of work stations, acquiring additional subassemblies one at a time and gradually accumulating its on-board systems and the maze of wiring, ducting, control cables, and other interior plumbing. Major steps include the matings of nose and tail sections, the wing, the engines, and the landing gear. Upon completion of work at the last station, the airplane is rolled out of the assembly plant to the flight line for production flight test, a process that by no means compares with the exhaustive research and development testing but that nonetheless involves a thorough check-out of specified performance.

There are many types of small missiles that require no such elaborate techniques or facilities. Composed basically of a cylindrical shell, a warhead, a guidance system, and a rocket motor, they are readily assembled in a simple low-bay plant. The larger missiles of the ballistic type and space-launch vehicles are assembled aircraft fashion in a high-bay facility, although most of them, horizontally mounted on the line, do not require the extreme bay heights of modern jumbo jet planes.

Spacecraft are unique among the flight vehicles because final assembly takes place under clean-room conditions. Preliminary assembly covering the mating of the structural elements is carried out in a conventional work area, after which the spacecraft goes through an air lock into an atmosphere-controlled chamber, where it is vacuumed and, in some cases, machine tumbled to shake loose dust particles. It then moves on to the final assembly clean room. Some spacecraft are assembled at progressive work stations; others remain in a fixed position while teams of specialists successively install the myriad on-board systems. Since the spacecraft gets no flight test, the equivalent, an intensely detailed acceptance checkout, is handled in the clean room by automatic test equipment. Then the spacecraft is encased in a steel container for shipment.

MAINTENANCE FACILITIES

The maintenance support provided by aerospace-industry firms revolves mostly around aircraft of the corporate, airline, or military types. Light-plane maintenance is generally handled by local fixed-base operators, not considered a part of the aerospace industrial complex. Launch vehicles and spacecraft, although they are maintained throughout their prelaunch life by constant checking and correction, are single-use systems. The smaller military missiles are maintained in the field by specialists in the operating unit. Major missiles similarly undergo routine maintenance at their field installation or silo base, but certain types of work are beyond the capabilities of field units. Overhaul of an inertial guidance system, for instance, requires special equipment and clean-room assembly, so it is usually returned to the plant of the original builder for re-manufacture.

Routine maintenance, normally accomplished by the civil or military operator, includes frequent inspections, sometimes after every flight, sometimes after a series of flights or a stated interval of time, and minor maintenance, such as replacement of a part or repair of a faulty item of equipment. This type of maintenance can be handled at most airline terminals and military bases.

Major maintenance of aircraft

Major maintenance involves complete rework of an airplane or engine that has had considerable service time. Some of the larger airlines have their own extensive facilities for major overhaul, and major military air forces have similar facilities. Even in these cases, however, some major maintenance work is contracted, as is the work of all other operators who do not have adequate facilities, to maintenance-specialist firms operating either independently or as a division of a major aerospace company.

These firms require facilities almost as elaborate as those of the original airplane or engine manufacturer: high-bay plants or hangars, machine tools, many types of inspection and test systems, engine-test cells, disassembly and assembly tools, parts- and components-cleaning equipment, equipment for overhaul of many of the on-board systems, and a complete inventory of spare parts. Maintenance companies are closely monitored as to capability and safety standards by appropriate government agencies. A maintenance firm must be certified by the agency for each type of product it overhauls, and a repair operation must be supervised by an agency-certified mechanic. Maintenance companies operate under

two additional guidelines: the original manufacturer of the equipment provides a manual outlining maintenance requirements and detailing specifications that must be met, down to the tolerance of an individual part, and the operator provides a similar manual, which lays down his own demands as to maintenance procedures and standards.

Airframe and engine overhaul. In a major airframe overhaul, the entire structure is disassembled, and each component is visually inspected for wear or damage. Additionally, structures are examined by X-ray, fluorescent, ultrasonic, and dye-penetrant methods to detect defects not visible to the eye. Corrosion is removed by sandblasting or vacuum blasting. Defective components are repaired or replaced, sometimes requiring machining operations to make a part. The airplane is reassembled, in much the same manner as it was originally assembled, repainted, then rolled out to the flight line for post-maintenance ground and air test, which is as thorough as the original production test.

Engine overhaul procedure

The procedure is similar in engine overhaul: complete disassembly, down to individual parts; removal of grease or carbon from each part; inspection of individual parts and measurement by precision equipment to insure conformance with the tolerance limits tabled in the manufacturer's manual; repair or replacement of any part that is even marginal; then reassembly via the progressive-work-station technique. After assembly the engine is mounted in a test cell and run through a lengthy series of tests as directed by the manuals. In all maintenance and overhaul operations, whether on airframe, engine, or accessories, technicians follow the same quality-control procedures that were in effect during original manufacture.

Remanufacture process. The most elaborate type of program under the general heading of maintenance is the remanufacture process, a military measure that combines a general overhaul with an updating of the airplane's performance capability. It is customary practice in aircraft manufacture to progress through several models of a basic airplane type, the later models incorporating design changes and better on-board systems dictated by service experience with the original model. Thus, older aircraft differ considerably from advanced models of the same type. If it is decided that a particular model, in service for several years, still has years of useful life, it is cheaper to upgrade its performance by remanufacture than to build additional aircraft of the advanced model.

Remanufacture is usually handled by a division of a major airframe company—perhaps, though not necessarily, the one that built the plane originally. The remanufacture facility needs additional capabilities; it must, for example, have design and engineering groups for the extensive modifications needed, and it must have the ability to manufacture large structural segments of the airplane if design changes so dictate. In this type of work the airplane is disassembled and stripped of its interior equipment and its engines, which almost always are replaced by improved models. After inspection and repair, the airframe segments that can be retained move through progressive work stations for step-by-step mating with new components. Similarly, the on-board equipment that is still usable is overhauled and delivered to the work stations for reincorporation, along with the new systems that design progression has specified. Reassembled, painted, and production-tested, the airplane begins life anew, a machine in most respects superior to the original.

BIBLIOGRAPHY. *The Illustrated Encyclopaedia of Aviation and Space*, 14 vol. (1971), a general source written in layman's language; JAMES J. HAGGERTY and H. GUYFORD STEVER, *Flight* (1965), a general treatise on aviation for the layman; CHARLES H. GIBBS-SMITH, *The Aeroplane* (1960), the best source for historical aerospace data; F. ALEXANDER MAGOUN and ERIE HODGINS, *A History of Aircraft* (1931), an excellent source for early aviation history; WALTER T. BONNEY, *The Heritage of Kitty Hawk* (1962); and *The American Heritage History of Flight*, ed. by ALVIN M. JOSEPHY (1962), two general historical works: *Jane's All the World's Aircraft* (annual), the best reference work for aircraft production data on types produced—also contains international information on space vehicle and missile types produced; *Aerospace Yearbook* (annual, 1919–70), the standard reference work on the aerospace industry in the United States.

(J.J.Ha.)

Aerospace Medicine

Beginning with determinations of oxygen requirements for pioneer balloonists by Paul Bert, a late-19th-century French physician–physiologist, aerospace medicine has evolved as a specialized branch of medicine with innumerable responsibilities in support of man's flight in the atmosphere—aviation medicine—and beyond the atmosphere—space medicine. The ultimate aim of this specialty is to promote the safety and effectiveness of man while he is exposed to the stresses encountered in aerospace flight.

In the role of clinician, specialists in civil and military aerospace medicine establish and apply appropriate medical standards for the certification and selection of pilots and other flight personnel to assure that they have no physical limitations or medical conditions that could impair their performance under usual or emergency flight conditions.

The work of aerospace medicine specialists

Aerospace medical specialists plan and help to carry out flight crew training in first aid and in the prevention of illness and injuries among passengers; they also assist in training paramedical personnel in the aerial transportation of patients. Such specialists also apply the principles of preventive medicine against the spread of disease by air travel and for the protection of aircraft crews and passengers from serious illnesses during trips to developing areas. They also conduct preflight and postflight medical evaluations of astronauts to assure fitness for space flight and to identify any adverse effects of space flight on the body. During space flight, they monitor physiologic responses of astronauts and advise them on the management of in-flight medical problems with the aid of the spacecraft's limited medical provisions. When any hazardous research or equipment testing causes exposure of human beings to simulated or actual aerospace environments, physicians establish safety standards and provide monitoring and therapeutic medical support as needed.

Beyond the scope of clinical medicine, aerospace medical specialists work with the multidisciplinary teams that develop the vehicles, emergency systems, and protective equipment for manned aerospace flight. With physiologists and psychologists, the medical specialists determine man's safe tolerance limits and performance capabilities. With the aid of other specialists—including those in biology, nutrition, and radiation—they provide aerospace engineers with data required for designing systems compatible with effective pilot functioning, comfort, and health. In the investigation of aircraft accidents, their role is to cooperate in the identification of causes of accidents, injuries, and death.

STRESSES OF AEROSPACE FLIGHT

Aerospace medicine is based on an understanding of the physiological, pathological, and psychological effects on man of the various stresses encountered during flight.

Gaseous stresses. Although the percentage of oxygen in the Earth's atmosphere remains fairly constant, its partial pressure decreases along with the decrease in atmospheric pressure during ascent and thus produces oxygen deficiency, called hypoxia, in the exposed individual. At 10,000 feet (about 3,000 metres) hypoxia can significantly impair pilot performance. Above 10,000 feet, hypoxic symptoms, such as impairment of judgment, deterioration of fine muscular coordination, drowsiness, light-headedness, and unwarranted self-confidence, appear earlier and with greater frequency and severity after exposure to relatively small changes in altitude. Hypoxia can be aggravated by such factors as physical activity, fever, ingestion of alcohol, illness, and the taking of any of a number of drugs. Hypoxia can limit useful function within ten to 15 minutes of exposure to the atmosphere at an elevation of 12,000–15,000 feet. At 20,000–

Symptoms of oxygen deficiency

23,000 feet, unconsciousness can occur within several minutes; at 35,000 feet, within one-half to one minute; and at 43,000 feet and above, within 13–15 seconds. The time until death also diminishes rapidly with altitude, ranging from several hours at 22,000 feet to a few minutes at the highest altitudes, where breathing ceases and irreversible brain damage begins within four minutes of exposure. Ebullism, or boiling of body fluids, occurs immediately after exposure to altitudes above 63,000 feet and may significantly increase hypoxic effects.

Hypoxia can be prevented by maintaining an adequate atmospheric pressure in aircraft and spacecraft cabins, by supplying oxygen by mask, and by inducing acclimatization to altitude. For the safety of flight crews and passengers, the air pressure in cabins on airliners and military transport aircraft must be kept greater than the atmospheric pressure at 8,000 feet. Oxygen must be provided by mask if significant failure in cabin pressurization occurs. Spacecraft cabin atmospheres have varied considerably, from use of 100 percent oxygen at five pounds per square inch absolute (atmospheric pressure at sea level is 14.7 pounds per square inch absolute), in Mercury, Gemini, and Apollo spacecraft of the United States, to use of a mixture of 20 percent oxygen and 80 percent nitrogen at near sea level pressure in the Vostok and Voshkod spacecraft of the Soviet Union.

In small unpressurized and partly pressurized civilian and military aircraft, oxygen is provided by mask, either by continuous flow or on demand from an oxygen regulator. To conserve oxygen, most demand regulators gradually increase the concentration of oxygen in the inspired air, providing 100 percent oxygen at 34,000 feet altitude, at which the pressure of pure oxygen is equivalent to that of oxygen in air at sea level. For prolonged exposures from 34,000 to 43,000 feet and brief exposures up to 50,000 feet, oxygen must be supplied under pressure. At even higher altitudes man must be enclosed in a pressure suit to enable him to breathe oxygen at a life-sustaining pressure. Space suits used in some space missions have been pressurized with 100 percent oxygen at about 3.7 pounds per square inch absolute. Finally, even though improvement in altitude tolerance occurs after several days of exposure in a low-pressure chamber, such altitude acclimatization has not been practical in light of the protection offered by pressurized cabins, oxygen systems, and pressure suits.

Symptoms of hyperventilation

Hyperventilation—excessive rate and depth of breathing—with associated symptoms of dizziness, numbness and tingling, muscle tremors and spasm, and, finally, unconsciousness caused by the removal of abnormally large amounts of carbon dioxide from the blood, is much more likely to occur during exposure to high altitude than at usual elevations. Predisposing factors are anxiety, decreased air density, unfamiliarity with oxygen-breathing equipment, and hypoxia. To avoid serious impairment of performance, all aviators must learn to recognize hyperventilation at its onset and so take corrective action by consciously controlling rate and depth of breathing.

Re-entry problems during descent

During ascent, expanding gas in the middle ear always leaves it by way of the eustachian tube without difficulty, so that the middle ear and surrounding atmospheric pressures remain the same. During descent, however, gas re-entry can be impaired by inadequate opening of the eustachian tube; this often occurs in conjunction with the common cold. Aerotitis media—inflammation of the middle ear—with severe pain and hearing loss, can then occur from the effects of negative pressure on tissues surrounding the middle ear cavity. Another condition known as aerosinusitis, with associated excruciating pain, can be produced by inadequate pressure equalization of the nasal sinuses during descent, and occasionally ascent, in persons with severe inflammation of nasal tissues. Gas trapped in the stomach or large bowel can cause agonizing pain and fainting, and expanding gas trapped under dental fillings can produce excruciating pain. Rapid expansion of internal gas, as in sudden loss of aircraft, spacecraft, and space suit atmospheric pressure ("explosive" decompression), can damage lung tissues, even with respiratory passages open during decompression.

A potentially dangerous effect of exposure to high altitude is the formation of bubbles of nitrogen and other gases that had been dissolved in body fluids. This can lead to one or more manifestations of decompression sickness, including joint pains and, with much less frequency, chest pains and cough, skin symptoms, bizarre nervous system problems, and often fatal circulatory failure, or shock. Although 30,000 feet is generally considered the critical altitude for decompression sickness, strenuous physical activity, obesity, rapid ascent, and prolonged exposure to altitude have contributed to the occurrence of symptoms at as low as 18,000 feet. Decompression sickness is prevented by the maintenance of an adequate atmospheric pressure during flight or by "prebreathing" 100 percent oxygen for a period to reduce nitrogen in body tissues to a safe level before exposure to altitude (see COMPRESSION AND DECOMPRESSION INJURIES).

Toxic gas accumulation in the atmosphere is of particular concern in space flight. Carbon dioxide, a waste product of body metabolism, can reach toxic levels and severely limit mental and physical functions within minutes in space suits, or within hours in spacecraft cabins, if systems that remove carbon dioxide from their atmospheres fail or are temporarily overloaded. Highly toxic gases can also be introduced in aerospace atmospheres by accident and as by-products of combustion and thermal breakdown of materials. Possible toxic effects of trace contaminants that are produced in sealed environments by man, materials, and machinery assume increasing importance as space missions become longer in duration.

Kinetic stresses. The unique kinetic, or motion, environment of aerospace flight can be misinterpreted by the organs of equilibrium, leading to spatial disorientation. Failure to monitor or believe aircraft instruments that provide orienting information when flight situations produce sensory illusions is a common cause of fatal aircraft accidents.

Excessive stimulation of the organs of equilibrium, particularly by complex angular, or turning, head motions during aerospace flight, can produce motion (air) sickness, characterized in progressive order by stomach awareness, drowsiness, dizziness, cold perspiration, nausea, vomiting, and prostration. Apprehension, overheating, disagreeable odours, and high noise levels may be contributory factors. Motion sickness may be prevented by adapting to flight motions through training, by restricting head movement in flight, and by taking a motion sickness drug, such as scopolamine, meclizine, or cyclizine, well before flight (see MOTION SICKNESS). Motion sickness drugs frequently produce blurred vision and drowsiness, however, and thus they cannot be taken by pilots.

Effects of acceleration

The forces of acceleration and deceleration, as encountered in aircraft banking turns, nosing over and pulling out of a dive, and in catapult takeoff, spacecraft engine firing, ejection escape, arrested landings, and parachute openings, can impair pilot performance and produce injuries. Accelerative forces of flight are stated in terms of multiples of the force imposed by the Earth's gravity, or *G*. Application of accelerative forces in the head-to-foot direction, or positive *G*, at levels of 4.5 to 6.0 *G* for longer than six seconds not only restricts muscular actions markedly because of the greatly increased body weight but also produces loss in peripheral vision (gray-out), then loss of central vision (blackout), then unconsciousness, because of the pulling of blood away from the head to pool in lower parts of the body. Tolerance to the high positive *G*-levels encountered in flying high-performance combat aircraft is considerably increased by applying pressure to lower parts of the body with a constricting garment, or *G*-suit. Human tolerance to headward, or negative, acceleration is only 2.5 to 3.0 *G* for a few seconds.

Transverse *G*, or accelerative forces applied at right angles to the body's long axis, is tolerated well at levels up to 20 *G* for several seconds; this is particularly relevant to determining safe body positions for the sustained high *G*-loads imposed during launch and re-entry phases of

space flight, and for high impact forces, as occur in crashes. Tumbling and spinning from loss of control of aircraft and spacecraft, as well as from inadequate ejection system stabilization after escape from high-speed aircraft, can prevent physical attempts to regain control and produce marked spatial disorientation, motion sickness, and injuries. Severe vibration throughout the body is encountered, in particular, in high-speed, low-altitude flight, helicopter operations, and the boost and re-entry phases of space flight. Although this stress can seriously impair pilot performance, it rarely inflicts injury.

Aerospace systems, including propulsion units of aircraft and spacecraft and auxiliary maintenance apparatus, produce the most severe noise environments in existence. Excessive noise not only annoys, fatigues, diminishes performance, and interferes with speech communication but also produces temporary and permanent hearing loss and ear pain. Control of noise at its source is not practical in many aerospace situations, so hearing conservation programs must be instituted to monitor and control duration of noise exposure, provide personal sound protectors (such as earplugs, earmuffs, and helmets), and periodically evaluate hearing.

Exposure to weightlessness during space flight has resulted in a temporary intolerance of standing upright on return to Earth because of an excessive pooling of blood in the lower extremities. If this problem is indeed significant and is not counteracted by donning a constrictive lower body garment on return to Earth, simulation of *G* effects on blood vessels of the lower extremities by such measures as periodically applying negative pressure or accelerative forces to the lower part of the body may be necessary during long space missions. Significant, but by no means dangerous, loss of minerals from the bones and reduced work capacity observed after space flights may also require preventive measures during flights of long duration. Nausea and disorientation reported in Soviet space flights and stomach awareness and nausea experienced during early phases of American space flights have been attributed to excessive complex angular (turning) head movements while moving about freely in the weightless space environment. This problem has been prevented by intensive preflight training in high-performance aircraft and simulators and by controlling head movements for a period of hours to a few days into space flights.

Radiation stresses. The risk of exposure to microwaves with frequencies and intensities sufficient to damage the lens of the eye must be considered whenever radar and microwave ovens are used in aerospace systems. Thermal stresses of aerospace flight range from the high heat exposures that can result from inadequacies or breakdown of temperature control systems in high-speed aircraft, in spacecraft, and in space suits to the exposures to extreme cold that can occur during inadequately protected ejections from aircraft and during extra-vehicular operations in shaded space environments. Man's range of thermal comfort is rather narrow, and insufficient thermal protection in aerospace flight can lead to discomfort, heat disorders, burns, lowering of body temperature, and cold injuries.

Risks of eye injury

In high-altitude and space flight, the glare can be dazzling, with associated eye discomfort and reduced visual efficiency. The greater irradiance from the Sun in near Earth space than at sea level and absence of atmospheric scattering of light to give a warning glow as the line of sight approaches the Sun impose significant risks of transient and permanent visual impairment during space flight unless adequate light shielding is provided. Absorbed high in the atmosphere, ultraviolet radiation is of little concern in atmospheric flight, but in space it can produce so-called arc-welder's eye injury, or snow blindness, in relatively short unprotected exposure.

Harmful effects from exposure to the particulate radiations (*i.e.*, radiations of subatomic particles) in space are of concern in space mission planning and in spacecraft and space suit design. Avoidance of prolonged flights within the Van Allen radiation belts (zones of geomagnetically trapped protons and electrons surrounding the Earth) minimizes the risk of skin injury that could result from prolonged exposure. Essentially unpredictable in frequency, time of occurrence, and intensity, solar flares are the most uncertain and probably by far the greatest radiation hazard in space. Radiation measurements in near Earth aerospace environments indicate that even with the shielding provided by present-day manned spacecraft hulls, high-energy protons and possibly alpha particles emitted by several large solar flares in a few days, with astronauts beyond the protective umbrella of the Van Allen belts, might produce sufficient skin injury and possibly influenza-like symptoms to impair performance. It is not known whether galactic cosmic radiation, consisting of extremely high-energy atomic nuclei, is a hazard in space and whether virtual absence of magnetic fields in space will impose a significant stress on man, who has been exposed for aeons to the Earth's strong magnetic field.

Temporal stress. Man's circadian, or 24-hour, rhythm of wakefulness and sleep, with associated changes of practically all physiologic functions, is synchronized with the local geographic, light–darkness, day–night cycle. Although apparently not injurious, desynchronization of these cycles by flights across several time zones and into space can produce discomfort for many days from hunger, sleepiness, and wakefulness at the wrong times of day. Therefore, whenever flight personnel must perform at peak proficiency, their rest–activity cycles must either continue to synchronize with their originating time zone or be allowed time to resynchronize with a new time zone before flight. Various drugs for hastening resynchronization are being evaluated.

System stresses. Overestimation of man's mental and physical capabilities and inadequate provision for his comfort and basic needs in the design of aerospace systems can compromise his performance and health. Prolonged monitoring of flight panel instruments that are difficult to read, too numerous, provide too many data for continued comprehension, or are poorly arranged in terms of importance and range of effective visual scan can lead to serious piloting errors from visual and mental fatigue. Flight controls that are difficult to reach and manipulate or that coordinate poorly with aircraft and spacecraft response can also be mentally and physically fatiguing. Such psychologically and physically stressful factors as confinement, isolation, lack of privacy, uncomfortable seating, poor eating and sleeping arrangements, poor illumination, uncomfortable surrounding colours, extremes of humidity, noxious odours, hazardous equipment and protrusions, and insufficient accommodations for leisure and recreational activities are to be avoided if at all possible in both aircraft and spacecraft design. Particularly in spacecraft, and to a degree on military aircraft that fly missions of long duration, adequate supplies of food and water, provisions for elimination of bodily and other wastes from the environment, facilities for personal hygiene, and medical provisions all are necessary for optimum efficiency and health.

Psychological and physical stresses

AVIATION MEDICINE

History. As early as 1910, the German military was selecting pilots only with superior physical and mental qualifications and then making every effort to keep them at peak fitness and morale. This program placed both the American and the British air forces, which had not fully recognized the need for a special medical service for aviation, at a distinct disadvantage during the early years of World War I. Faced with an unbelievably high accident rate of one crash per 241 flying hours in 1917, the following year the United States Army established the Air Service Central Medical Research Laboratory, at Hazelhurst Field, Long Island, New York, "to investigate all conditions which affect the efficiency of military pilots and to consider all matters pertaining to their selection and their physical and mental fitness." This organization immediately began to improve medical standards for pilot selection, examine pilot candidates, identify and study stresses of flight, and train medical officers in the specialized duties of flight surgeon.

Early work at School of Aviation Medicine

Having essentially the same responsibilities as flight surgeons of the 1970s, the first flight surgeons were to make medical support directly and personally available to all flight crew members, living and flying with them and understanding all stresses of flight in order to make medical care directly applicable to flying situations. In 1922 all instruction and research in aviation medicine in the American military services were conducted in the new School of Aviation Medicine at Mitchell Field, Long Island, with Louis H. Bauer, author of the first text on aviation medicine, as its first commander. This facility was the predecessor of the great existing military aerospace medical institutions in America, including the United States Air Force (USAF) School of Aerospace Medicine at Brooks Air Force Base, Texas; the United States Naval Aerospace Medical Center at Pensacola, Florida; and the USAF Aerospace Medical Research Laboratory at Wright Patterson Air Force Base, Ohio.

As medical director of the new Aeronautics Branch of the Bureau of Air Commerce, Bauer developed the first medical standards for the certification of civil airmen and designated and trained the physicians to whom all civil airmen were to report for periodic medical certification.

Another milestone in the history of aviation medicine was the founding of the Aero Medical Association by Bauer in 1929. Now called the Aerospace Medical Association, this organization has grown from a group of 35 physicians to an international, interdisciplinary, active membership of more than 5,000 medical specialists, life scientists, and engineers in the field of aerospace medicine and bio-astronautics. Its constituent associations, which also participate in its annual scientific meetings, include the Airline Medical Directors Association, the Civil Aviation Medical Assocation, the Aerospace Industrial Life Sciences Association, the Society of U.S. Air Force Flight Surgeons, and the Space Medicine Branch. The Aerospace Medical Association publishes a monthly journal, *Aerospace Medicine*, formerly the *Journal of Aviation Medicine*.

The growth of aviation medicine has followed parallel routes in most countries. War efforts gave rise to intensive aeromedical research and flight surgeon training in such facilities as the Royal Air Force Institute of Aviation Medicine, Farnborough, England; the Royal Canadian Institute of Aviation Medicine, Toronto, Canada; the Kaiser Wilhelm Institute, Heidelberg, Germany; and the Army Aero Medical Laboratory, Tachikawa, Japan. Many other governments have established similar standards for medical certification of airmen, with the International Civil Aviation Organization, a specialized agency of the United Nations, serving as international coordinator, especially with respect to certification of airline pilots. The large European airlines have been active in all phases of aviation medicine. Chapters of the Aerospace Medical Association throughout the world and numerous other aeromedical organizations, including the International Academy of Aviation and Space Medicine, the International Astronautical Federation, and the European Congress of Aviation Medicine, sponsor international meetings for the exchange of knowledge in aerospace medicine. Other aerospace medical journals are *Aeromedica Acta*, *Space Life Sciences*, *Japanese Journal of Aerospace Medicine and Psychology*, and *Voyenno-Meditsinsky Zhurnal*.

International Civil Aviation Organization

Medical standards. In America, Part 67 of the Federal Aviation Regulations applies to the medical certification of civil airmen—including crews of fixed-wing powered aircraft, helicopters, powered gyrocopters, airships, and free balloons—by more than 7,000 designated aviation medical examiners across the country. Their primary concern is with any physical impairment or medical condition that could compromise safe pilot performance. The type of airman activity determines which minimum class of medical certificate must be held. A first-class certificate is required only for airline transport pilots, is valid for six months, and must be obtained from a specially designated senior aviation medical examiner. A second-class certificate is required for other commercial pilots, for flight engineers, and for flight navigators and is valid for 12 months. The third-class certificate is required for private, student, and free balloon pilots and is valid for two years. For medical certification, there is no age restriction, which is already determined by the various airman certificates required for specific flight activities. Citizenship is not a requirement for medical certification. Notably, all airmen must obey the regulation that states that

> no person may act as a pilot in command, or in any other capacity, as a required pilot flight crew member while he has a known physical deficiency that would make him unable to meet the physical requirements for his current medical certificate.

The airlines currently do not conduct certifying medical examinations on their pilots for the Federal Aviation Administration (FAA). Airlines having medical departments or contract arrangements for medical services conduct more comprehensive medical examinations, often including psychological testing, and apply more stringent standards in pilot selection than in FAA certification. In other countries many airlines conduct certifying examinations for their governments.

The military services conduct comprehensive physical and psychological examinations and apply stringent medical standards in pilot selection. Periodic medical examinations are given to all aircrew members by flight surgeons. When a flight surgeon must deny recertification for flight duties, further medical evaluation is usually conducted at a major military aerospace medical institution and final decision rendered by a panel of senior flight surgeons, all aerospace medical specialists.

Medical criteria for passenger flying. The ever-increasing number of passengers travelling by air includes a certain proportion who have various types and degrees of clinical disorders. In 1961 the Committee on Medical Criteria for Passenger Flying of the Aerospace Medical Association established guidelines for the airline industry and all physicians to improve the proper selection of patients for transportation by air. The guidelines were expected to minimize the risk of adverse effects of flight stresses on patients, disturbance of fellow passengers, requirements for special attention during flight, and the interruption of schedules and great expenses imposed by unscheduled landings to disembark ill passengers.

The committee advised physicians of the need to focus attention in particular on effects of atmospheric pressure changes on pre-existing clinical problems. It emphasized the risks of air travel for persons with serious heart or lung conditions or severe anemia; for infants under seven days of age; and for some women in the ninth month of pregnancy. The committee suggested that a person subject to epilepsy or unable to take care of his own physical needs be accompanied and that the safety of other passengers be considered before acceptance as a passenger of any person in a psychiatric state. It was advised that persons with contagious disease or whose behaviour might be hazardous be denied passage.

Medical training of flight crews. Instruction provided to airline cabin attendants, such as stewards and stewardesses, in the recognition and first-aid treatment of the more common medical problems that can occur during flight varies considerably. In general, airlines flying long, overwater routes have provided comprehensive aeromedically oriented first-aid courses, whereas short-route airlines have usually given minimum instruction in the basics of first aid. Flight deck crews—pilots, flight engineers, and navigators—of many airlines also receive aeromedical indoctrination, with particular emphasis on consequences of loss of cabin atmospheric pressure and instruction in the recognition of medical problems. Notably, no airline currently specifies that a nurse be a member of any cabin crew.

Indoctrination of military flight personnel

In the military services, all flight personnel initially receive thorough instruction in the stresses imposed by atmospheric flight and are thoroughly familiarized with the personal protective equipment available to them. Periodic refresher training may be required, depending on the type of aircraft flown. Instruction is conducted in an aviation physiology training unit by a team usually con-

sisting of a flight surgeon, aviation physiologist, aviation physiology technicians, parachute riggers, and other protective equipment technicians. A typical program includes discussion of flight stresses, their effects and protection from these effects, familiarization with various types of oxygen-breathing and other protective equipment, including *G*-suits, full-pressure suits, emergency escape systems, and parachutes. In high-altitude indoctrination, aviation personnel briefly experience hypoxia and, if required, training concerning the ejection seat, night vision, parachute and water survival, and flash blindness.

Aerial transportation of patients. Large-scale transportation of the ill and injured by specially equipped aircraft with in-flight medical attendants has mainly occurred in military services. For example, the Military Airlift Command of the United States Air Force operates a global network of aeromedical evacuation flights returning patients from overseas theatres to the United States and transferring patients within the United States to appropriate treatment centres. The air evacuation of battle casualties may be carried out in one or more highly coordinated steps, depending on severity of injuries. The injured soldier may be taken quickly by air ambulance from the battlefield to a forward hospital, where his condition is stabilized and treatment of his wounds is initiated. When he is ready to be moved, a cargo aircraft adapted for the purpose may transfer him to the nearest major or general hospital serving the theatre. If he is not likely to return to duty within a specified period, the soldier may be transported to the special military treatment centre nearest his home. All of these steps require the support of trained aeromedical crews. Medical technicians on air ambulances have sufficient training to permit them to continue first-aid treatment in flight and to recognize and cope with emergencies that may arise before the patient reaches the first medical treatment centre. In aircraft used in subsequent steps of the evacuation process, in-flight care is provided by specially trained flight nurses and aeromedical technicians. At each stop along any aeromedical evacuation route, physicians and nurses are informed on problems that may be encountered in flight. Also, the physicians who release patients for flight must know what medical complications can occur from flight stresses.

The civil aerial transportation of patients has not reached the magnitude of the corresponding military operation, since most illness or injury occurs when persons are reasonably near their homes and patients seldom have to travel far for special care. Nevertheless, there has been a marked increase in the use of air taxis to transport patients from one city to another and in the use of helicopter ambulance services to move patients quickly over short distances from accident sites to emergency hospitals. Helicopter ambulance services generally carry trained paramedical personnel and necessary medical equipment and supplies on board. Few air taxis, however, provide comparable aeromedical support. Consideration is being given to establishing standards for operation of any air ambulance service.

SPACE MEDICINE

History. In 1948 the United States Air Force School of Aviation Medicine, then at Randolph Air Force Base, Texas, sponsored the first symposium on space medicine; among the participants were several leading German scientists in the field of aviation medicine and space science. The following year, the school established the first department of space medicine in the world, with the former director of aeromedical research in the German Air Force, Hubertus Strughold, as its first professor of space medicine. Now acknowledged as the father of space medicine, Strughold had theorized for many years on the medical implications of space flight. Aware of the space flight ambitions of Gen. Walter Dornberger and rocket specialist Wernher von Braun of the V-2 rocket program at Peenemünde, on the Baltic, he estimated that intensive space medical research, paralleled by great technological advances, would make manned space flight feasible within 15 to 20 years. At that time, it appeared that with the exception of solar radiation and weightlessness, the stresses of space flight would be logical extensions of the not insurmountable stresses of atmospheric flight. This was borne out by data from ground-based research on the effects of acceleration, deceleration, and vibration noise, and atmospheric control, and from observation of pilot performances in the rocket-powered aircraft, as well as from knowledge gained from high-altitude balloon flights. Also in the pre-Sputnik era, American Aerobee rockets and captured V-2 rockets were used for the first time to expose living creatures to space flight conditions and to study effects of radiation and weightlessness in space on seeds, fruit flies, mice, and monkeys. Prior to his launch into space, man could be exposed to weightlessness only for periods of less than a minute in high-speed aircraft flown in parabolic trajectories.

First symposium on space medicine

Medical selection of astronauts. Early astronauts were a highly select group, in a medical sense, in that they were all military test pilots who had volunteered their services and were homogeneous in terms of flight experience and keen interest in the space program. After a medical records review and initial psychiatric and psychological screening, medical examination included conventional medical, family, and sociological histories; investigation into the attitudes of the immediate family toward hazardous flying; a physical examination conducted by flight surgeons, internists, and many other medical specialists; many laboratory tests; X-rays of the teeth, sinuses, chest, gastrointestinal system, heart, and lower backbone; and tests of lung function and exercise capacity. Many psychological tests were administered to determine personality and motivation, intelligence and special aptitudes. The astronaut candidates were subjected to a large number of stress tests of both a physiological and psychological nature.

Biomedical monitoring during space flight. Early in the American space program, aerospace medical specialists realized the need for continuously monitoring physiologic responses of astronauts during flight in order to make decisions about the medical feasibility of additional flight objectives, to assure the health and safety of astronauts, and to obtain biomedical information for predicting the effect of space flights of extended duration on man. The monitoring instrumentation had to provide reliable physiologic data, could not interfere with crew comfort or function, and had to be compatible with other spacecraft systems and be under virtually automatic control. Great advances subsequently made in developing new biosensors (devices for monitoring physiologic processes), in miniaturizing instrumentation, in transmitting physiologic data great distances, and in immediate analysis of these data by computer not only made possible the monitoring of man's physiologic state during aerospace flight but also provided the field of general medicine with tools for continuously monitoring patients undergoing diagnostic studies while they are confined to bed and during day-to-day activities.

Reasons for monitoring of astronauts

In later flights, body temperature, blood pressure, heart rate (determined by means of electrocardiograms), respiration, and voice transmissions of the astronauts were monitored by aerospace medical specialists while spacecraft were over tracking stations. Brain wave patterns and heart sound patterns were also recorded in Gemini flights on an on-board recorder. Electrocardiograms and respiration were recorded for one astronaut at a time during the first two Apollo flights and in later flights for all three astronauts at a time; their temperatures could be ascertained at any time. While spacecraft were behind the Moon, data were stored on a tape recorder and were transmitted to Earth as soon as the spacecraft came back into radio contact. When an astronaut was outside the spacecraft, his electrocardiogram and respiration were determined either through the space suit "umbilical" or by means of telemetry from equipment carried by the astronauts.

Medical experience in manned space flight. Over the years, many harmful consequences of space flight have been predicted on the basis of experiences in atmospheric

flight and on conjecture about effects of known and unknown stresses of the space environment on man. Man has tolerated all stresses thus far encountered in space flights without impairment of his performance or deterioration of his health. Medical experience in Mercury, Gemini, and Apollo missions is briefly summarized below.

Spacecraft cabin and space suit pressures were maintained without difficulty, and the astronauts lived in a comfortable, "shirt-sleeves" environment on later flights.

Effects of prolonged exposure to pure oxygen

The only identified effect of prolonged exposure to the atmosphere of 100 percent oxygen at five pounds per square inch absolute of the pressure used in the spacecraft cabins has been a slight decrease in blood cell mass; this rapidly returned to normal in the postflight period. Removal of nitrogen from the blood by breathing 100 percent oxygen for three hours before flight has prevented decompression sickness.

Accelerative forces imposed during the launch and reentry have been well within man's limits of tolerance. Effects of weightlessness have been touched upon in the section above on *Stresses of aerospace flight*. Otherwise, adaptation to unusual sensations and effortless movement in the weightless environment has occurred rapidly, without any effect on proficiency. All extra-vehicular tasks in the lunar environment were also accomplished without significant difficulty. Predicted problems in eating and drinking were avoided by proper packaging of food and water. There was no difficulty in swallowing, digestion, or urination in space. Although greatly improved waste-management systems are required, collection and disposition of urine and feces under weightless conditions was accomplished without environmental contamination.

Even with penetration of the Van Allen belts, all radiation doses received during space missions were exceedingly small. No solar flares occurred while man was in space. Harmful effects of light and ultraviolet radiation were prevented by visual shielding and avoidance of gazing directly at the sun.

Passing through a day and night every 90 minutes in orbiting missions did not disrupt circadian rhythm. As more was learned about sleep in the space environment and more confidence gained in systems reliability, work–rest cycles of all astronauts were synchronized whenever possible with prelaunch schedules, and spacecraft were artificially darkened by covering windows. Sleep in the weightless environment was normal, except for being somewhat hampered early in missions as the crews adjusted to the desire for contact with some object during sleep.

Modified isolation of astronauts for 21 days before flight was employed in Apollo lunar missions in an attempt to reduce the risk of communicable illness during flight to a minimum. Colds, gastrointestinal upset with vomiting, eye irritation, minor skin problems, and urinary infection occurred in space; most symptoms were controlled effectively with medications from on-board medical kits. To preclude the possibility of contaminating the Earth's biosphere with lunar organisms, Apollo astronauts were subjected to strict decontamination and isolation procedures and were quarantined under close medical observation for 21 days after returning from lunar explorations. There was no evidence of infectious diseases that resulted from lunar material. This material contained no identifiable organisms and had no adverse effects on animal and plant life. The isolation procedure was abandoned in later missions.

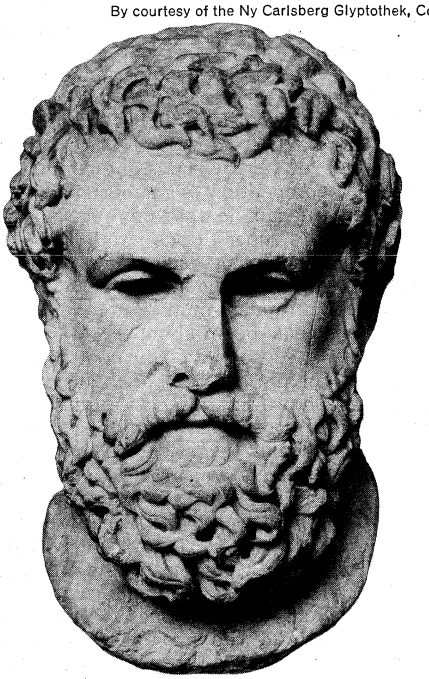
BIBLIOGRAPHY. HUGH W. RANDEL (ed.), *Aerospace Medicine*, 2nd ed. (1971), is recognized as the principal reference textbook for specialists in aerospace medicine and aeromedical scientists throughout the world. JOHN A. GILLIES (ed.), *A Textbook of Aviation Physiology* (1965), provides detailed information on human responses to the aerospace environment and on aircrew protective equipment. ROBERT J. BENFORD, *Doctors in the Sky* (1955), traces the growth of organized aviation medicine in America as a narrative history of the Aero Medical Association. The USAF SCHOOL OF AVIATION MEDICINE, *German Aviation Medicine, World War II*, 2 vol. (1950, reprinted 1971), a historic classic, continues to serve as a basic reference for researchers in aerospace medicine, biomechanics, and human tolerance. DOUGLAS E. BUSBY (ed.), *Recent Advances in Aerospace Medicine* (1970), well demonstrates the scope of aerospace medicine in presenting key papers from the *Proceedings* of the 18th International Congress of Aviation and Space Medicine.

(D.E.Bu.)

Aeschylus

The first of classical Athens' three great writers of tragedy, Aeschylus lifted the art from its dim origins as choral and largely static recitative to fully developed drama, the presentation of action. At centre stage of most of Aeschylus' dramas stands man as a political being, reflecting the playwright's concern and the time in which he lived. Aeschylus was born under tyranny (in Greece, a government—not necessarily abusive—in which absolute power is vested in a single ruler), grew up in the turbulent period when the Athenian democracy had to prove itself against self-seeking politicians at home and invaders from abroad, and himself took part in his city's first struggles against the Persians. He later watched the political battle between liberals and conservatives, each fighting to turn Athens into their idea of what it should be. It is the tension between the often conflicting concerns of a leader for his people and for himself that provides the focus of tragic action in Aeschylus' plays.

By courtesy of the Ny Carlsberg Glyptothek, Copenhagen



Aeschylus, Roman copy of a Greek original, second half of the 4th century BC. In the Ny Carlsberg Glyptotek, Copenhagen.

Birth and background. According to the generally reliable Parian Chronicle (an inscription containing important dates in Greek history set up in the 3rd century BC), Aeschylus was 35 years old in 490 BC, the year of Marathon (the battle in which the Athenians routed Persian invaders), which puts his birth in 525 BC. His youth was spent in a time of ostensible if repressive political peace, when the sons of Peisistratus, a tryant who had held sway earlier in the century, held supreme power. A tradition that Aeschylus' family resided at Eleusis (northwest of Athens) may be true though it may rest only on a misunderstanding of a line in *The Frogs*, by the comic playwright Aristophanes. The historian Herodotus described a person whom other sources call the brother of Aeschylus as a resident of Pallene, which may indicate that the family really lived there. Of their father nothing is known beyond his name, Euphorion.

Dramatic career. In about 501 or 500 BC, Athens' major dramatic competition, the urban Dionysia, seems to have been reorganized; and, in one of the immediately

succeeding years, Aeschylus is recorded as having competed—probably for the first time. He won his first victory in the theatre in the spring of 484 BC (the title of the winning play is unknown).

In the meantime, he had helped to fight his country's battles. At Marathon, in 490, the poet, according to one tradition, was carried wounded off the field, his participation in this battle being singled out years later for mention on his epitaph. In 480 the Persians again invaded Greece, and once again Aeschylus saw service, fighting at Artemisium (on northernmost Euboea, off Greece's east coast) and at Salamis (near Athens), his role in the latter battle being attested by the poet Ion.

Aeschylus thus saw at firsthand how close Greece had come to annihilation, how truly miraculous had been its deliverance, above all, how much it owed to the foresight and planning of Themistocles (a statesman and general). These impressions took root in his imagination and blossomed in his play *Persians*, the earliest of his works to survive. Produced in the competition of the spring of 472 BC, it won first prize. The whole panorama of the enemy's onslaught and the Greeks' repulse is here: the colossal ambition of Persia's king thwarted by a force of democratically governed Athenians responding with vigour and precision to the commands of their inspired leader.

Earliest surviving work

Aeschylus acceded to the request of the Sicilian tyrant Hieron I to present *Persians* again at Syracuse (in Sicily), where, either on this same occasion or, as seems more likely, some years before, he had also staged a pageant, *The Women of Etna*. Ancient biographical traditions contain a number of elaborate but self-contradictory stories about the poet's visits to Sicily. According to one tale, he went there out of pique at losing to Simonides, a well-known writer of victory songs and commemorative dedications, in a competition to memorialize the heroes of the Marathon; or again he was said to have been escaping his countrymen's superstitious wrath when the theatre's seats collapsed during one of his performances. Neither account rings true: the first is likely to be nothing more than an embellishment on the known fact that Aeschylus wrote an elegiac poem about the Marathon campaign; the second was also told of the contemporary dramatist Pratinas. The presence of Aeschylus at the court of Hieron I does, however, raise large questions, about how a poet whose loathing of tyranny resounds through every line of *Prometheus Bound* could have written *The Women of Etna*, a piece honouring the foundation of the tyrant Hieron's new city of Etna. An answer may be found in the tone the poet took: if a papyrus fragment published in 1952 contains, as is now widely believed, a passage from Aeschylus' lost pageant, the prominence of the figure of Justice in it may be significant; Justice, in Aeschylus' words, "sits by the throne of Zeus . . . honouring the just life of those who are just." This is more admonition than praise, an ideal to which the dramatist is urging Hieron to aspire. A production of *Persians* at Syracuse may also be interpreted in this light. The tyrant's brother and predecessor, Gelon, had refused to answer a Greek summons for help before Salamis but had successfully routed an invading force from the North African city of Carthage that coincided with the Persian invasion, and Hieron himself had repulsed the Etruscans at Cyme, Italy, in 474 BC. Rivalry and animosity must now be laid to rest, Aeschylus may be saying, and, with the eastern and western barbarian invaders driven off, Greece must now unite under the standard of freedom and justice.

The sources attest Aeschylus' acquaintance with a number of other writers, some of whom he must have met, though probably not for the first time, at Hieron's court. The Byzantine scholar Eustathius records Aeschylus' association with the lyric poet Pindar, who was seven years his junior and who, as a young man, had studied at Athens, where he won a dithyrambic (dithyrambs were types of choral ode) contest in 497/496 BC. The two poets show certain similarities of theme and treatment (a taste for unusual metaphors and graphic personifications of Mt. Etna's eruption, for example), and Pindar was later said to have acquired his "grandiloquence" from Aeschylus. Aeschylus' city-centred and essentially democratic view of man, however, is a world away from Pindar's praise of manliness as manifested in the athletic contests, a virtue that is supranational and, so ultimately, rootless. Apart from a probably spurious story of Aeschylus' defeat by the prolific and versatile lyric poet Simonides, no direct evidence links the two men. It is likely, however, that they had met in the circle about Themistocles, for both manifested in their own ways the extent to which each had fallen under the general's spell; and, since Simonides is mentioned as an habitué of Hieron's court, the two may have renewed their acquaintance there. Just as Aeschylus had taken over (particularly in *Persians*) some techniques from the older tragic dramatist Phrynichus, Sophocles, another of the great Athenian tragedians, fell under Aeschylus' influence, as he himself was later to admit when he described his early style as one of Aeschylean "weightiness." (There are signs that, during the decade or so when the two were competing at Athens, some reverse influences may also have been at work; for *Prometheus Bound*, considered by many critics to be Aeschylus' last play, shows a surprisingly large number of Sophocles' tricks of style.) It was probably in the 460s that Aeschylus first met Ion, a young poet from the Greek island of Chios, who had recently arrived in Athens.

Association with Pindar and Sophocles

Aeschylus' later career is a record of his dramatic success—and one stunning defeat, at the hands of the neophyte Sophocles, whose maiden performance at the Dionysiac festival of 468 BC was victorious over the older poet's entry. But suspicious circumstances surrounded the victory. The Greek biographer and historian Plutarch (*c.* AD 46–*c.* 119) records that the usual judging procedures were suspended, and the board of generals, headed by the conservative politician Cimon, received the honour of deciding the winner in the drama competition. It is hard to avoid the conclusion that politics clouded the general's critical faculties and that Cimon may have been reacting to Aeschylus' support—four years earlier, in *Persians*—of Themistocles, his bitter rival. A palpably erroneous tradition has the dramatist resentfully going off once more to Sicily; in fact, Aeschylus recouped the loss with a victory in the year 467, with his Oedipus trilogy, of which only the finale, *Seven Against Thebes*, survives. "A drama full of Ares [the Greek God of War]," the professor of rhetoric Gorgias called it, and the vivid picture it presents of Thebes (about 45 miles northwest of Athens) under attack by an invading force from Argos (Sparta's rival in southern Greece) was probably coloured by memories of the Persian threat to Athens in 480 BC. Eteocles, the play's hero, acts according to his lights to defend his city but goes off to slay his brother Polyneices and in turn be slain by him under the curse of blood from which the ill-starred family of Oedipus suffers. Also dating from the 460s is the Danaid (named after Danaus, a major figure in the legend dramatized by Aeschylus) trilogy, of which only the opening play, *Suppliants*, survives. The drama presents the agonizing dilemma of a leader who is approached by refugees demanding protection and asylum; if the King grants their request, he is likely to embroil his own people in bloody war. The anguish felt by the mythical Pelasgus is similar to that which Athens must have experienced when it was approached by the Ionian revolutionaries in 499 BC and also, more recently, by the Argives, who had given shelter to the outcast Themistocles; neither parallel can have been far from Aeschylus' mind when he wrote his emotion-filled piece. In 462 the Athenian constitution was reformed along radical lines and the conservative Cimon exiled, not without the infliction of deep political wounds by the opposing factions and the actual shedding of blood. Aeschylus' Orestes trilogy of 458 reflects this atmosphere of citizen armed against citizen, "tribal war" in the poet's own phrase; but, in the Aeschylean resolution, strife gives way to harmony, bloodshed to peace, revolution to the rule of law. Old divinities such as the Furies are transformed and given a constructive function in Athena's city, and they thus become its Eumenides, or "Well-Wishers."

The Oedipus and Danaid trilogies

After the *Oresteia*, Aeschylus went to Sicily for the last time; Aristophanes has his comic Aeschylus say in *The*

Frogs that he "did not get on with the Athenians," which need not be taken entirely seriously. A later, sensationalizing tradition invented a full-scale trial before the Areopagus (the Athenian council of justice)—much like the trial conceived by Aeschylus himself for Orestes in *Eumenides*—for revealing the content of the secret rites held at Eleusis. All of this seems to be sheer fiction. The chronographers recorded his death at Gela (on Sicily's south coast) in 456/455, aged 69. A ludicrous story that he was killed when an eagle dropped a tortoise on his bald pate was probably fabricated by a later comic writer. At Gela he was accorded "heroic" honours, a public funeral with sacrifices and dramatic performances held at his grave, which subsequently became a place of pilgrimage for writers. A later tyrant of Syracuse, Dionysius I, was said to have taken pains to acquire the poet's writing desk, which he revered as a sacred relic.

Reputation and influence. Aeschylus' career had been outstandingly successful. One account, perhaps based on the official lists, assigns him 13 first-place victories; in short, out of over 80 known titles, 52 won first prize. The Athenians took the unprecedented step of decreeing that anyone who wished should be allowed to reproduce his works in the festival competitions; and it may have been on this occasion that he was awarded the title "Father of Tragedy," which is mentioned by the Sophist philosopher Flavius Philostratus (*c.* AD 170–*c.* 245). No specific dates of repeat productions are known, but the sources speak of victories won by posthumous performances of his plays, and comic reminiscences in Aristophanes (especially in *The Birds* and *The Frogs*) suggest that there may have been revivals during the Peloponnesian War (431–404 BC). On the motion of the Athenian politician Lycurgus, bronze statues of the three great classical tragedians were set up in the Theatre of Dionysus (below the Acropolis in Athens) in the 330s. Several Roman copies are thought to derive from this "official" statue, which in any case could have been nothing more than an idealized impression. A bust in Rome's Capitoline Museum, which portrays a bald-headed pensive man well beyond middle age, may derive from a true likeness, the original perhaps having been commissioned by the citizens of Gela.

Aeschylus' two sons also achieved prominence as tragedians. Although one of them, Euphorion, is said to have won four victories with posthumous new productions of his father's plays, he also won first prize in his own right in 431 BC over Sophocles and Euripides' *Medea*.

Aeschylus' artistic innovations

Aeschylus' influence on the development of tradedy was considerable. He is said to have introduced into Greek drama the second actor who had an individual character and dialogue, thereby supplanting the theatrical convention of having only a single performer and a chorus. This innovation allowed much more variety and dynamic tension in plot construction and dialogue. Although a stage designer, Agatharchus, is mentioned as an associate of his, it is doubtful whether full-scale painted scenery (*e.g.*, a realistic Causasus mountain backdrop for *Prometheus Bound*) could have been in use so early. His works were in any case known for their spectacular scenic effects (*e.g.*, the bizarre modes of conveyance for characters in *Prometheus*; the sleeping, snorting Furies tableau in *Eumenides*) and exotic, often terrifying masks and costumes; the first appearance of his chorus of *Eumenides* was said to have struck terror in the spectators. It appears that he reduced the size of the chorus and standardized their number at 12, as well as altering the chorus' function. Although the dominance of the chorus in early tragedy is a mere hypothesis, it is probably true that, in the words of the Greek philosopher Aristotle, Aeschylus "reduced the chorus' role and made the plot the leading actor." At the same time, he dispensed with the services of a choreographer, training his choruses himself and even devising new dance steps for them to perform. He probably acted in most of his own plays, following the practice usual among dramatists at the time.

Although the Roman statesman and orator Cicero cites and even translates numerous passages from Aeschylus' works, by the 1st century AD the Romans' grasp of his difficult language began to slip; the rhetorician Quintilian, for example, found him "rough and disjointed." By the 2nd century AD, Aeschylus had slipped to a position of secondary popularity. In 1475 the papal library of Sixtus IV contained but three manuscripts of Aeschylus. In the late 18th century, however, Aeschylus was "rediscovered"; the *Agamemnon* was a favourite of Napoleon's, and the distinguished French writer Victor Hugo fully appreciated the poet's genius. With Percy Bysshe Shelley's *Prometheus Unbound* (1820), Aeschylean inspiration entered the mainstream of English poetry. In the 20th century the writers who have fallen most deeply under his influence are the French poet and dramatist Paul Claudel and the U.S. playwright Eugene O'Neill, whose *Mourning Becomes Electra* (1931) captures better than perhaps any other modern work the brooding grandeur of Aeschylus.

MAJOR WORKS

TRAGEDIES: *Persai* (472 BC; Latin, *Persae;* Eng. trans., *Persians*); *Hepta epi Thebas* (467 BC; Latin, *Septem contra Thebas;* Eng. trans., *Seven Against Thebes*); *Hiketides* (463 BC; Latin, *Supplices;* Eng. trans., *Suppliants*); the trilogy known as the *Oresteia* (458 BC), comprising *Agamemnon*, *Choephoroi* (Eng. trans., *The Libation Bearers*), and the *Eumenides* (Eng. trans., *The Furies*); *Prometheus desmotes* (date uncertain, probably late; Latin, *Prometheus vinctus;* Eng. trans., *Prometheus Bound*).

BIBLIOGRAPHY. The Greek text of the complete plays is available in editions by ULRICH VON WILAMOWITZ-MOELLENDORF (1914, reprinted 1958); and GILBERT MURRAY, 2nd ed. (1955); a new edition by D.L. PAGE appeared in 1972. H.J. METTE has edited the fragments (1959), in German. H. WEIR SMYTH provides a literal prose translation in his "Loeb Classical Library" edition, 2 vol. (1922–26; reprinted 1956 and 1963), of which the second volume also contains a translation of the major fragments, and HUGH LLOYD-JONES contributes an appendix of recent discoveries. The "University of Chicago Greek Tragedies Series" provides verse translations by RICHMOND LATTIMORE *et al.* (1953; also available in paperback) and the "Prentice-Hall Greek Drama Series" is issuing up-to-date translations with literary commentary (1970–). Studies of special topics relating to Aeschylus' plays are: A.J. PODLECKI, *The Political Background of Aeschylean Tragedy* (1966); and A.F. GARVIE, *Aeschylus' "Supplices": Play and Trilogy* (1969). Articles on specific subjects include: E.R. DODDS, "Morals and Politics in the *Oresteia*," *Proceedings of the Cambridge Philological Society*, 186:19–31 (1960); HUGH LLOYD-JONES, "Zeus in Aeschylus," *Journal of Hellenic Studies*, 76:55–67 (1956); and C.J. HERINGTON, "Aeschylus in Sicily," *ibid.*, 87:74–85 (1967). The various candidates for Aeschylus' portrait are illustrated by G.M.A. RICHTER in *The Portraits of the Greeks*, vol. 1, figures 577–610 (1965). For a useful collection of modern criticism, see MARSH H. MCCALL (ed.), *Aeschylus* (1972).

(A.J.P.)

Aesthetics

Nature and scope of aesthetics as a discipline

The discipline called aesthetics may be described broadly as the study of beauty and, to a lesser extent, its opposite, the ugly. It may include general or theoretical studies of the arts and of related types of experiences, such as those of the philosophy of art, art criticism, and the psychology and sociology of the arts. The word general is emphasized because a narrowly specialized study of a particular work of art or artist would not ordinarily be regarded as an example of aesthetics, although it might provide data for aesthetics. Aesthetics has often been defined more specifically as the science of the beautiful, a definition implying an organized body of knowledge covering a special field of subject matter. Although this raises the question of what properly is a science and what is its subject matter, it at least has the advantage of bringing aesthetics into some relation with the familiar special sciences of today and suggesting that it may be a recent development—which in fact it is.

The arts may include the visual and theatre arts, music, dance, and literature. In the ancient world, there was no clear distinction between aesthetic and useful art.

Aesthetics as a philosophic or scientific discipline is not to be confused with art, though it may undertake to study the arts in a more or less intellectual, logical way. To develop an empirical science of aesthetics or a philosophi-

cal approach to one, it is not enough to create or enjoy works of art. It is necessary also to try to understand, explain, and perhaps evaluate them. Aesthetic evaluation, however, has been slow to develop toward science.

For related articles dealing with the broadest aesthetic principles in all of the arts, see ART, PHILOSOPHY OF, which stresses the experience of artistic creativity (while the present article stresses that of the viewer and appreciator); ARTS, CLASSIFICATION OF THE; ARTS, CRITICISM OF THE; ARTS, SOCIAL AND ECONOMIC ASPECTS OF THE; ARTS, PRACTICE AND PROFESSION OF THE; ARTS, STYLE IN THE; PRIMITIVE, FOLK, AND POPULAR ARTS. For discussions of aesthetic principles operative in the major mediums of the arts, see LITERATURE, ART OF; DANCE, ART OF; MUSIC, ART OF; THEATRE, ART OF; ARCHITECTURE, ART OF; SCULPTURE, ART OF; PAINTING, ART OF; PHOTOGRAPHY, ART OF; MOTION PICTURES, ART OF. (Th.M./S.C.P.)

NATURE AND ORIGIN OF MODERN AESTHETICS

Philosophy as the mother discipline

For the Greeks there were no special sciences; yet they had a voracious appetite for knowledge. There were special skills and techniques such as shoemaking, carpentry, shipbuilding, sailing, fishing, farming, and animal breeding. But among the men who were pursuing knowledge for its own sake, the so-called philosophers, though there were vigorously opposed schools of thought on numerous issues, there were no divisions of subject matter. The special sciences did not break away from philosophy till many years later, largely in the 17th century and thereafter. The breaks came when there was so much to be known in a special field that it was all that one man could do to master it.

Philosophy, however, continued as a centre for nonspecialized knowledge and general comprehension and as the seat for potential specialties yet to come. The last of the special sciences to break away from philosophy was psychology, a break that occurred in the mid-1920s, previous to which time psychology had been a special subject within the field of philosophy. This position, however, is held today by aesthetics: it remains a part of philosophy, rather highly specialized but not so specialized as to break away completely.

Baumgarten's naming of the field. Aesthetics was even late in recognizing itself as a special field within philosophy and in acquiring its name. The name aesthetics first appeared in *Reflections on Poetry* (1735), by Alexander Baumgarten, a follower of the Rationalistic school of philosophy under the influence of René Descartes, the father of modern philosophy, and Gottfried Wilhelm Leibniz, a polymath German philosopher. These men had not given attention to aesthetic matters; their concern had focussed upon what Descartes had distinguished as "clear and distinct" ideas, by which was meant systematic thought as developed in logic and mathematics. This distinction restricted cognition to conceptual knowledge. Baumgarten noticed, however, that sensory and perceptual cognition was thereby excluded. It appeared to him that here was to be found just the sort of cognition developed in poetry and the other arts and that the Cartesian restriction resulted in a serious omission; thus he set to work to correct this oversight.

In Descartes's terms, "clear" meant what was unquestionably apparent to an attentive mind and "distinct" what was precisely differentiated from other things or internally articulated. In a geometrical system all of the terms and propositions are presumably clear and distinct. Ordinary discourse is usually just the opposite: obscure and confused. Sensations and perceptions were regarded as intrinsically confused. They could be clear, however; and when they were so, men had a definite kind of cognition (though of a lower order than clear and distinct logical cognition). This perceptual cognition is what Baumgarten found in poetry (and by extension in the other arts). Drawing upon the Greek word for perception (*aisthēsis*), Baumgarten coined the word aesthetics for the science of perceptual cognition—*i.e.*, of perception as clear but confused knowledge. His book is a systematic exposition of this subject matter, with many examples from Latin and Greek poetry indicating that the greater the perceptual clarity achieved in a poem the greater its poetic perfection. Thus Baumgarten's term aesthetics has become a permanent part of the vocabulary of philosophy.

Proportions of clear and confused ideas

The basic conclusion of his analysis can be stated in these terms: the aesthetic value, the beauty, of a poem is proportional to the extent of clear, as compared with confused, ideas that it contains; or—paraphrased in the terms of contemporary theory—the aesthetic value of a work of art is proportional to the intuited vividness of the fused quality of the experience that it evokes.

Baumgarten's work is important in the history of aesthetics. In spite of its strange terms and its forbidding mode of exposition (presented like a Euclidean geometry with axiomatic definitions and deduced propositions) and in spite of its poetic illustrations in languages that almost no one can understand today, it is well worth reading, for it marks the beginning of aesthetics as a special field of investigation.

Toward a well-rounded conception of aesthetics. To get a well-rounded conception of how aesthetics has developed since Baumgarten started it out on its own, it is necessary to review some of the principal aesthetic theories that have come to prominence and maintained themselves. The crucial question is what constitutes the subject matter for this sort of special study. In a general way Baumgarten has already demarcated it: its subject matter is the sort of responses and experience that works of art were created to produce, to which may be added natural objects and any other things that can evoke the same sort of experience. The central material here is the aesthetic experience.

But works of art themselves cannot be neglected either. Actually, for any extended investigation of the field, there is a reciprocity between the works of art and the aesthetic experience. No aesthetic theory will stand up if it cannot explain why certain works of art are outstanding exemplifications, embodiments, or stimuli for the aesthetic experience. The works need not be limited to the so-called classics (though these manifest a high degree of persistence in demanding recognition). They may be artifacts of primitive people; they may be fairly enduring rituals or traditions or customs of certain sorts or routines as for a dance or techniques. The justification for an aesthetic theory, the evidences for it, will almost inevitably in the end be sought in the aesthetic arts.

Baumgarten constantly appealed to passages of poetry to support his theory. It is as if the instigating problem for aesthetics were to account for the values attributed to these arts—to poetry, drama, dance, music, painting, sculpture, architecture. Having explained the values of the arts, a theory may then expand these values as widely as they may go to natural objects, to acts of living, eating and drinking, playing and loving, and even to dreaming. But it is the explanation that can be given for deeply prized works of art that stabilizes an aesthetic theory.

Philosophical theories reflected in aesthetics

The various conceptions of aesthetic value (or aesthetic experience) may be summarized as they have developed under the guidance of four or five philosophical theories that have been widely influential in modern times. Their development as aesthetic theories has been relatively recent, though traces of them—or even partial systematizations of some of them—can be found as far back as the ancient Greeks. There is now no single dominant theory of aesthetic value. Nor can these four or five views be consistently amalgamated into one; for each holds to be dominant a certain feature that the others would accept only as subordinate. But a sympathetic understanding of each will illumine the ways in which aesthetic values arise in this special field. There is a fair amount of agreement as to the core of the field and even as to the possible expansions of it, though the precise demarcation of the boundaries varies with each view. The chief difference among these views seems to lie in the way they distribute their value emphases within the field.

NATURALISTIC OR SCIENTIFIC APPROACHES TO AESTHETICS

Hedonism: the pleasure theory. The simplest and most liberal of these views is that which identifies the aesthet-

ic experience with some distinctive sort of pleasure, satisfaction, or liking. As a starting point for discussion, this theory has the advantage that it spreads out before the reader practically the whole field of the psychological data with which any sort of aesthetic theory will find itself concerned. Not all aesthetic subject matter is psychological, but a large proportion of it is—obviously all that has to do with aesthetic appreciation. There is no more direct way of initiating this survey than through this simple psychological approach, which merely says in principle that an aesthetic fact is a way of finding pleasures and satisfactions in things.

If pleasure is allowed to include satisfactions of any kind, this theory can properly be called the pleasure theory, or hedonism. It has a long history. It comes out of a naturalistic philosophical approach. Pleasure is equated with positive value and pain or unpleasantness with negative value. Aesthetic value is distinguished from other values, such as the ethical, by one or more special traits. In the broadest sense aesthetic value is taken as intrinsic pleasure or the liking of a thing for itself, as against liking it as a means to an end. Aesthetic values are immediate (direct) as opposed to mediate or practical values.

Element analysis of aesthetic value

This view lends itself easily to a typical mode of scientific analysis—that of element analysis such as one finds in chemistry, in which atoms combine into molecules and these into living cells and cells into organisms. Thus, aesthetic values can be conceived as building up from elementary sensory pleasures and likings, to combinations of these into aesthetic forms and modes of composition, and thence to the organizations of works of art, which in turn get gathered together into styles of art. This conception of aesthetic analysis accorded well with an important school of so-called structural psychology, which had been producing the most constructive introspective and physiological research from the time of the 17th-century Empiricist John Locke until the mid-20th century, so that the data and laboratory results of this school could be utilized for an analytical aesthetics.

One of the virtues of a naturalistic aesthetics is the emphasis it spontaneously gives to the simplest aesthetic delights.

Santayana's version of the pleasure theory. This emphasis comes out in *The Sense of Beauty* (1896), by George Santayana, a Spanish-born Harvard naturalist, in which he presented the first fully systematized aesthetics couched in hedonistic terms—and, in fact, the first such naturalistic theory. The book is a classic literary work of art in itself. Its analysis is divided into three parts: the materials of beauty, form, and expression.

Materials of beauty. The first sort of materials that Santayana mentions are those connected with man's vital functions. "They are loose," he says,

> and unlocalized, having no special organ, or one which is internal and hidden within the body. . . . Particularly important are they for the *value* of our experience. They constitute health, without which no pleasure can be pure.

Then he presents a whole section on the aesthetic importance of sex from both its direct and indirect expression. These passages read like a poem at the same time that they give a condensed psychological analysis of wide significance. The following passage may be quoted both for its beauty and its insight:

> Sex is not the only object of sexual passion. When love lacks its specific object, when it does not yet understand itself, or has been sacrificed to some other interest, we see the stifled fire bursting out in various directions. One is religious devotion, another is zealous philanthropy, a third is the fondling of pet animals, but not the least fortunate is the love of nature, and of art; for nature also is often a second mistress that consoles us for the loss of a first.

Pleasures of the senses

On turning next to the pleasures of the senses, Santayana admits that, as aesthetic materials, the lower senses—touch, taste, and smell—lack the capacities to combine into large works of art as sound and the visual materials can. For that reason the term beauty, as denoting aesthetic value in the highest (and perhaps most proper) sense, could be withheld from them (though many naturalistic aestheticians would not agree with this restriction). But the stress on the aesthetic value of pure simple sensuous pleasures is ever present in Santayana. He writes:

> Sensuous beauty is not the greatest or most important element of effect, but it is the most primitive and fundamental, and the most universal. . . . Taste, when it is spontaneous, always begins with the senses. . . . When sincerity is lost, and a snobbish ambition is substituted bad taste comes in. The essence of it is a substitution of non-aesthetic for aesthetic values.

Form. Aesthetic form comes when an aesthetic pleasure arises from a combination of elements that individually have little or no value: thus a triangle may be exciting when its individual lines are dead; and a musical chord can give a satisfaction far beyond what its component notes can yield separately. He distinguished between two sorts of form: compositional form and what he called types (see below).

Design and pattern

In later work in naturalistic aesthetics, two varieties of compositional form were distinguished, viz., design and pattern. The function of design is, negatively, to avoid monotony or aesthetic fatigue. Out of this come aesthetic principles or laws of contrast and of theme and variation. When a particular kind of aesthetic satisfaction begins to pall, interest in it can be revived by a contrasting pleasant stimulus. Theme and variation is even more effective in avoiding monotony and maintaining interest because it can continue longer. Some whole musical compositions are built on this principle only. An interesting theme or motive or tune is announced and then repeated with variations of rhythm or accompaniment or key or even of some of its notes. With such variations the theme can be repeated many times without loss of interest. And finally the original melody can be repeated, as it was first given, and it will come as something fresh and welcome. Contrast and theme and variation can cooperate with each other, as they do in the familiar three-part form of music—the first part developing a certain theme, followed by a contrasting part developing new material in a different key or mood, followed in turn by a repetition of the first part, which gives a natural and comfortable ending. These principles are equally effective when used simultaneously as in the visual arts. Sustained visual design depends greatly on their mutual cooperation. Architectural design makes impressive use of them.

Turning next to pattern, it is clear that the compositional function of pattern is, negatively, to avoid confusion—as the function of design was to avoid monotony. The pattern depends greatly on what the Chinese called "the rule of five," which refers to the limitations of the attention span. A viewer can take in five elements at a glance without counting. In fact, he can take in even a few more; but five is easy, and an artist can count on it with assurance. Aesthetic order (not mathematical order, which is quite different) depends entirely on this principle, unless some other principle has already bound the elements together. Thus aesthetic order can be obtained by groupings that fall within the range of the attention span—*i.e.*, by groups of groups of groups of not more than five or so. The stanzas of regular versification, for example, are groups of lines, which are groups of verse feet—little patterns of rarely more than two or three syllables with different accents or durations—which, in turn, are groups of uttered syllables. The verse feet are gathered into lines of rarely more than six or seven feet, and when they are that long there is generally a secondary break in the middle known as a caesura. Stanzas are often four lines long. If they are longer than that, their patterns are internally divided by rhymes or regular variations of lengths of line that keep them within the attention span and so aesthetically well-ordered. Large organizing patterns of this embracing sort or of other sorts can bring whole works of art with hundreds of details into transparent order. Where such order is missing there is confusion, unless other connections are also present to help produce an overall unity of pattern.

Santayana's notion of type

What Santayana called type is another such mode of aesthetic organization. Types, which are produced as a result of learning, are said to be the concepts that men

have developed in making adjustments to their social and physical environment. They are of many kinds. What may be called natural types are a person's means of recognizing types of objects and features of his natural environment—animals and plants, clouds and seas and mountains, and the characters of the men about him and their characteristic ways of acting (whether friendly or hostile). Aesthetically, these are the materials that make up men's recognition of representations of reality in pictures, novels, dramas, and the like. There are also types of utility consisting in their recognizing the uses of objects and how they ought to function for maximum utility. The appreciation of function in architecture comes from these types. And there are types of technique derived from men's knowledge of the skills needed for production or performance, particularly observed in the appreciation of the performing arts. A type may be described as the psychological concept of a group of features recognizable as a whole when presented for attention. The aesthetics of types is the psychology of recognition applied to the arts.

The importance of types for a naturalistic aesthetics lies partly in their organizing power; for they consist in acquired patterns of connected elements that can give unity to a work of art. A work of architecture, for example, can be largely unified by the transparent fulfillment of its function throughout—in plan, construction, and design. But, even more, it is valued as a specific source of the pleasure that it yields, viz., the satisfaction derived from the fulfillment of a type. For a type once stimulated operates like a purpose demanding fulfillment. And the satisfaction involved, particularly if the fulfillment is difficult and rare, can be very great. This experience can easily be appreciated in the perfection of performance achieved in music or acting. But it is equally present in supreme achievement in any work of art. Wherever there is a glow of admiration in the excellence of an aesthetic work, there the observer is showing his appreciation of fulfillment. Because there is a conceptual background for this sort of aesthetic valuing, it is sometimes called intellectual beauty.

Expression. Of Santayana's tripartite division for a naturalistic aesthetics into pleasurable materials, form, and expression, the pleasures of expression remain to be noted. In this area, he refers to the aesthetic contributions of symbols, in which the beauty is not so much in the symbol as in what it symbolizes. It is the aesthetics of meaning. All language is involved in this realm so that literature is the outstanding art of expression. But meaning saturates all of the arts. In religious art, this is particularly obvious, for it leads deeply into the aesthetics of the emotions, where soon the pleasure theorist finds that he has to deal with tragedy. For a pleasure theory, however, tragedy is a great paradox; for in tragedy the artist deliberately seeks out unpleasant, even horrible, subject matter. The paradox of the horrible as the beautiful in a theory in which beauty is based on pleasure is the more acute in that many of what are reputed to be the greatest works of art are tragic.

Symbols and meaning

Catharsis theory of tragedy. Santayana and other hedonists have offered many suggestions to account for this paradox, not the least of which is the theory of catharsis. This theory was originally put forward by Aristotle, who in the *Poetica* made a careful analysis of the art of tragedy having the works of Aeschylus, Sophocles, and Euripedes in mind. His statement about catharsis is brief:

> Tragedy, then, is an imitation of an action that is serious, complete, and of a certain magnitude; in language embellished with each kind of artistic ornament, the several kinds being found in separate parts of the play; in the form of action, not of narrative; through pity and fear effecting the proper purgation of these emotions.

This is unquestionably the most discussed passage in the history of aesthetics. There are dozens of interpretations of it.

Freudian interpretation of tragedy

One view that accords with the hedonistic approach and employs a recent development in psychology is the Freudian interpretation, which is essentially this: that in dealing with extreme mental conflicts and emotional sufferings, man is endowed with a protective capacity to inhibit them so that they do not have to rise into full consciousness. They nevertheless are not really forgotten but, lying in the back of the mind as complexes, constantly seek relief, often coming up in weird disguises in dreams. There is a satisfaction in such relief; and one of the ways of getting it—without having to face the full consciousness of a trauma held back by the inhibitions—is through what are called tragic works of art. In a roundabout way men enjoy the horrors depicted because, without being fully aware of why, they thereby obtain relief from similar experiences they have had themselves and have inhibited but that constantly tend to express themselves. Tragic art is thus, as Aristotle put it, like a purge for pent-up emotions.

A hedonistic aesthetics, as has been seen, is psychologically oriented. In Santayana the approach is by way of a structural method of element analysis. But as a psychological type of aesthetic theory, hedonism readily accepts new psychological developments. How easily it can absorb the Freudian data has just been shown. In fact, Freudian discoveries have contributed greatly to an understanding not only of the peculiar emotional satisfactions possible through the art of relief in contrast to an art of delight but also of many of the perplexities connected with aesthetic creativity.

Gestalt psychology: its amplification of naturalistic aesthetics. Yet another recent psychological movement that has been absorbed into hedonistic aesthetics is that of Gestalt psychology, a movement that began in violent opposition to the analytical methods of structural psychology on which Santayana and many other naturalistic aestheticians were learning. Actually, the two movements are not necessarily contrary to each other; it is better to say they are complementary.

The Gestalt movement was inaugurated by a group of German psychologists, Max Wertheimer, Wolfgang Köhler, and Kurt Koffka, whose laboratory experiments cast serious doubts on the structuralists' long-standing assumptions that simple sensations are the ultimate building stones of mental structures, that these elements get related into wholes by the laws of association, and that all complex wholes are thus acquired by associative learning. The Gestaltists presented evidence that there are instinctive unlearned patterns and that there are special Gestalt laws that cannot amalgamate with the traditional associative pattern laws. Gestalt is, in fact, the German term for pattern, and this term has been retained to distinguish the nonassociative patterns discovered by this school.

The Gestaltists demonstrated that many mental patterns, being self-contained with intrinsic internal relations, cannot properly be described as reducible to combinations of smaller elements. Moreover, perception in children does not begin with simple elements and work up to larger combinations; it begins with large, easily comprehensible patterns and may later analyze out internal features, which may or may not be able to stand as independent patterns. These matters were particularly manifest in the visual field. The principal aesthetic work so far written from the Gestalt viewpoint is *Art and Visual Perception* (1954), by Rudolf Arnheim, a noted psychologist of art. In it, the emphasis throughout is on visual wholes that hold together and visually keep intact in combination with other such wholes. The book demonstrates how somewhat unfinished wholes tend to complete themselves in perception by closure and calls attention to psychological tensions intrinsic to certain shapes and to the tendency toward simplicity of shapes, the pervasiveness of the figure–ground relationship, the internal demands of balance, and so on. It is not that those using structural psychology as a background were unaware of most of these things in their aesthetic analyses but that they failed to give them the aesthetic emphasis that they deserved.

Arnheim's *Art and Visual Perception*

It has been remarked that exclusive emphasis on the Gestalt features in turn covers up many aesthetic discriminations of the structuralists, notably the immediate subtle sensuous discriminations that run through Santa-

yana's *Sense of Beauty* and the works of D.W. Prall, another Harvard philosopher, Walter Pater, a late 19th-century art critic, and others. The Gestaltists are sometimes accused of putting so much emphasis on form that they fail to enjoy the simple materials. There is no necessary conflict, however, between the aesthetic contributions of men with the analytic approach of Santayana and those with the Gestalt approach of Arnheim. Both are dealing with intrinsic satisfactions—as against practical instrumental values; and both seek to explain them by psychological principles working on mental materials generally stimulated by external objects such as works of art. The Gestaltists merely forced the structuralists to liberalize the field of aesthetic materials. Though it may be true that simple sensations are the ultimate discriminable sense materials, they need not necessarily be the ultimate mental building blocks of aesthetic perceptions. They may come internally built into instinctive perceptual patterns.

Complementary theoretical approaches

In short, hedonistic aesthetics can welcome Gestaltists and Freudians as well as traditional structuralists, provided only that these schools are tolerant of one another. Several works of men who have contributed notably to the hedonistic psychological approach to aesthetics are cited in the bibliography.

Nonphilosophical aesthetic specialists. Mention must be made of a large number of men who, though not professed philosophers, have contributed enormously to the discrimination of data within the aesthetic field, viz., the musicologists, the art historians, and the literary historians. These men are aesthetic specialists—so specialized, in fact, that they are often naïve in each other's fields and even in the underlying aesthetic philosophy of their subject. Yet their aesthetic contributions are the most factually detailed and extensive materials available. The philosophical aestheticians turn to them when they want expert information about such things as musical scales, rules of harmony and counterpoint, recognized forms, historic styles of art, cultural influences, cycles of taste, techniques of composition, and production and performance.

Thomas Munro: culminating exponent of a scientific aesthetics. Through the work of these specialists, the analyses of psychologically inclined aestheticians, and some experimental work in psychological laboratories, a large quantity of factual material has accumulated through the last century, which needed only to be systematized to demonstrate that a distinct science of aesthetics actually exists today. Such a systematization was accomplished in a compendious volume entitled *Form and Style in the Arts* (1970), by the philosopher Thomas Munro, who, ever since his *Scientific Method in Aesthetics* (1928), has argued that aesthetics should be treated as a science and not as something opposed to a science.

Scientific systematization of aesthetics

Munro's influence has been widely felt. He was the founder of the American Society for Aesthetics and of *The Journal for Aesthetics and Art Criticism*, both of which were set up as scholarly and scientific undertakings. In his later work he produced what could be regarded as a model scientific text in the field. His main headings deal with "the psychological components of art" and "the modes and dimensions of organization."

His materials are taken directly from works of art themselves as empirical objects like plants and animals and the historical styles of art, and his book reads like a botanical treatise and seems almost to be value free and objective (in that limited provincial sense bruited about by many scientists). But he says explicitly of his book, "I do not claim that it is 'value-free.'" He believes only "that traditional aesthetics has suffered from too much haste in approaching evaluative problems without the necessary factual knowledge and without an adequate methodology."

Critics agree that Munro is correct in saying that the material of his book is not value free, that every bit of it consists of facts observably related to pleasures and satisfactions and other experiences of intrinsic value, since all of it is lifted out of works of art. What Munro calls his "naturalistic, functional definition of art," which controls his selection of material throughout, reads in part as follows:

> Art is skill in making or doing that which is socially used or intended as a stimulus and guide to satisfactory aesthetic experience . . . especially in such a way that the perceived stimuli, the meanings they suggest, or both, are felt as beautiful, pleasant, interesting, emotionally moving, or otherwise valuable as objects of direct experience.

Munro arises directly from the naturalistic, hedonistic (in the wide sense) tradition of aesthetics to which Santayana belongs. His work stands as one sort of culmination of it. It shows in a monumental way how aesthetics may be presented as an empirical science.

NONSCIENTIFIC APPROACHES TO AESTHETICS

Influence of Immanuel Kant. There are other views, however, with often a completely opposed tradition. By way of transition, Kant's *Kritik der Urtheilskraft* (1790; *Critique of Judgment*) comes to the fore. Strangely, in spite of some conclusions quite at variance with those of the typical hedonists, Kant can still be regarded as an aesthetic hedonist because he held that aesthetic value exhibits itself in a sort of pleasure.

His stress on superior nonscientific factors. For Kant, however, it was a very special sort of pleasure. In order to attain to beauty, the pleasure must be (1) disinterested, (2) universal, (3) necessary in a uniquely specified way, and (4) must give the effect of purposiveness without actually being the satisfaction of a purpose.

Most hedonists have distinguished aesthetic pleasure from pleasure in general. There has been wide agreement that it has to be an immediate pleasure found in the object, not the practical liking of a thing as a means to an end. This distinction was very nearly the same as Kant's disinterestedness. But many hedonists wanted to reserve the term beauty for a special higher sort of immediate pleasure, of a kind not associated, for instance, with the pleasures of the lower senses. In spite of Santayana's aspersions against those who disdained pure sensuous pleasures, he himself made a sharp distinction between what he called objectified pleasures, which alone he called beautiful, and others that were only agreeable. By objectified pleasure he meant such pleasure as occurred when a person was so absorbed in a piece of music or a picture that the feeling of pleasure seemed to be emerging from the piece and not coming out of the observer. "Beauty," he said,

Beauty as a higher type of pleasure

> is pleasure regarded as the quality of a thing. . . . It depends upon the degree of objectivity my feeling has attained at the moment whether I say "It pleases me," or "It is beautiful."

Kant had the same feeling that beauty was a value derived from pleasures far above the merely agreeable. He was particularly concerned about the demand people make for the universality of judgments of beauty to comport with the high prestige of the arts. This demand (he felt) was also made for morality, which, without it, seemed threatened with disintegration. Moreover, universality was very urgent for the stability of scientific knowledge in the avoidance of Skepticism.

As regards scientific knowledge, Kant preserved its universality and necessity by his revolutionary theory of the categories. David Hume, a Scottish Skeptic, had just given a devastating blow to the concept of causal necessity by an analysis of man's ideas as based in sense impressions. Hume could find no introspective evidence for necessary connections between any two impressions of which he was aware.

Universality and necessity

Kant reinstated the necessity and universality of causal connections by invoking the action of a mental faculty of Understanding (as he called it) common to all men, which Hume had not recognized. Understanding is the faculty of concepts. Among its leading concepts is the category of causality by which the mind molds unordered sense impressions into the order of nature discovered by the physical sciences. Thus, Newton's spectacular cosmic organization in terms of the laws of motion and gravitation could be universally justified.

Kant was convinced that a similar justification could be found for the widespread belief in the universality of judgments of moral and aesthetic values. He accepted the naturalistic view that aesthetic value is to be found in pleasure. But most pleasures are clearly subjective and do not yield to molding by the concepts of the Understanding. Accordingly, the universality of aesthetic pleasure could not be accounted for by the categories of the Understanding. His solution was that it is due to representations or impressions that set up a harmonious interplay between two whole faculties of the mind. Aesthetic pleasure arises when the faculty of Imagination comes into harmony with that of Understanding. Because these faculties are common to all minds, being the very structure of mind, the aesthetic judgment of the universality of aesthetic pleasure is automatically justified. For the same reason it is seen to be necessary. It is disinterested because practical purposive interests have to do solely with concepts of ends, which belong only to the Understanding. And it is not a satisfaction in the achievement of a purpose because, as has been shown, purposes are based on concepts. It contains a sort of purposiveness, however, because the manifold of material gathered up by the faculty of the Imagination is brought to conform harmoniously with the Understanding as the faculty of concepts.

The distinction of aesthetic pleasure as superior to all of the common pleasures of mere sensuous enjoyment, carnal indulgence, and entertainment is thus established by its universality, necessity, disinterestedness, and purposiveness without explicit purpose.

Worth and importance of Kant's Critique. Kant's theory in his *Critique of Judgment*, when presented in stark outline as above, has struck some critics as hardly plausible and even as fantastic or absurd. Nevertheless, it has been acclaimed as the first major work in aesthetics and as one of the greatest works in aesthetic history. Yet there were no prominent succeeding Kantians; no one now believes in faculties of the mind; and faculty psychology (a heritage from Plato) has long been dead. Moreover, practically no one now believes in Kant's a priori categories of the Understanding, and very few believe in any sort of a priori claims of necessity or certainty at all except perhaps in pure logic.

Significance of *The Critique*

It is reasonable, then, to question why it is so important a work. It is partly because it gathered together and assimilated the scattered aesthetic opinions of Kant's own generation and of many preceding ages; it made a drastic distinction between the activities of the sciences and those of the arts, which prepared the way for the distinction between facts and values; and it raised the possibility of two kinds of cognition—scientific, conceptual cognition and the insight of the genius into deeper values. But, perhaps most of all, Kant's theory of categories and the control furnished by strong guiding concepts in determining men's conceptions of the world gave a new turn to the whole scientific and philosophical enterprise. It is a recognition of Kant's categories, in fact, that gives order to the discussion on aesthetics in this article—except that these categories are not here presumed to be a priori. The presumption is merely that the different outstanding theories of aesthetics are distinguished by certain groups of categories in terms of which their findings are organized. These different sets of guiding categories operate to give alternative relatively adequate and consistent hypotheses about aesthetic experience and modes of evaluation.

Though Kant deserves much intensive consideration, it is necessary here to pass to other important aesthetic theories and their categories. Because the so-called contextualistic theory has some affinities with the hedonistic theory with which the discussion began, this theory is dealt with next. Here again Kant's influence is discernible in its claim that there are two sorts of cognition—the scientific and the intuitive.

Contextualistic contribution. It is necessary here to be somewhat critical of the idea that aesthetics can be a science of beauty, though it can still remain an orderly study of beauty—a study of theories about the special sort of values called aesthetic. Aesthetics then becomes a general term covering all of these aesthetic theories, including the theory that aesthetics is a special science dealing with aesthetic values as its subject matter.

In considering the contextualistic theory, it is notable that, when dealing with hedonistic aesthetics, attention was called earlier to the close connection between pleasure and the sensuous material in which the pleasure was taken. Of these two aspects, the hedonist emphasizes the pleasure whereas the contextualist emphasizes the intuited quality in which the pleasure is taken. The latter defines the aesthetic field as that of the quality of events: great beauty is great enhancement of quality. It is surprising what a great change of perspective develops from this change of emphasis.

The contextualistic view is the most recent systematic aesthetic theory to arise over the historical horizon. There were intimations of it in the past. Baumgarten's concept of confused cognition was one of them. The paraphase earlier offered for his definition of aesthetic perception is a contextualistic one. Aside from him, however, the French intuitionist Henri Bergson, a philosopher of biology, and the Italian Idealist Benedetto Croce were the earliest philosophers to give impetus to this new theory. Bergson's emphasis on intuition and instinct in his *Évolution créatrice* (1907) and his book *Le Rire* (1900; Eng. trans., *Laughter*, 1917) had many of the features of contextualistic aesthetics. He pronounced himself a Pragmatist, which was another name for contextualist.

Croce's aesthetic theory

Intuition of quality and fusion. It was Croce, however, who rather unintentionally had the most profound influence in disseminating the doctrine of aesthetic value as the intuition of quality. Croce's book *Estetica come scienza dell'espressione e linguistica generale* (1902; Eng. trans., *Aesthetic*, 1922) aroused tremendous interest during the early decades of this century. His view was, however, by no means a pure contextualism. Croce arose from the organistic Hegelian school (to be described presently) and his principal philosophical ideas reflected this origin. In aesthetics, however, his novel concept of intuition fed directly into contextualistic thought. He made an unusual identification of intuition with expression, which split his influence two ways: his emphasis on intuition as immediate qualitative sensitivity was highly contextualistic, while his emphasis on expression led to a quite different expression theory of aesthetics (which will be referred to later).

The close connection that Croce felt to exist between intuition and expression emerges from an observation that can be regarded as at least half true, viz., that a creative artist does not really know what the quality of his creation will be until he has fully expressed it. In this view, the intuition of the quality of his aesthetic object comes to light only in the process of giving it expression: a poet finds out what his poem is only as he makes it. He does not have to make it on paper, however; he can make it in his mind. But it is only in such expression of the poem that he realizes his intuition of it. And a reader in turn only gets the intuitive feel of the poem by a sort of re-expression of the poet's original expression. Croce goes on to draw many conclusions from this identification of intuition and expression, which constitute the subject matter of his *Aesthetic*. One of these conclusions is to reject all concepts from aesthetic intuition: concepts produce science, he holds, and intuition of immediate qualitative expression produces art. Thus any pure aesthetic intuition must be completely drained of concepts—a dichotomy reminiscent of Kant.

The mature contextualist in aesthetics does not go as far as this: he admits concepts into his intuitions of vivid quality, for even concepts have qualitative overtones, and they can help in connecting groups of qualities together. He seeks qualitative vividness in aesthetic intuition, however, not scientific analysis or prediction. So the contrast between scientific knowledge and the intuitive cognition of immediacy is retained. But the contextualist will welcome any means of increasing the

range and intensity of qualitative immediacy even to the utilization of concepts and analysis.

Stress on the quality of the whole in contextualism

Contextualism is unique among aesthetic theories not only in focussing on vividness of quality as the central character of beauty or aesthetic value but also in stressing fusions. It is the quality of the whole event or object that is emphasized—the total intuited quality of the picture, the poem, or the melody. What is beautiful is the whole picture, not the heap of its details. This means that in the aesthetic intuition the qualities of the details are fused into the quality of the whole.

The fact of fusion in man's everyday practical living is generally overlooked, even though it is omnipresent. In immediately recognizing a person's face, the observer does not put together a lot of details about the shapes and positions of the eyes, nose, and mouth, chin, forehead, and hair. He instantaneously intuits the character, the total quality of the face. All of the details count in the individuality of that face; and if one of them were much altered or omitted, he would notice it with the shock of a new quality. Similarly, with most of the objects that a man perceives, the details are fused in the comprehension of the whole. The fused quality, moreover, is different from any of the details even when the perceiver is taking up their qualities into the whole. The process and result is transparent in hearing any simple chord followed by the notes that compose it played in isolation: the quality of the whole is different from, yet clearly derived from, the qualities of the notes that are fused in the chord. The Pragmatist William James's famous example is the taste of lemonade: one can analyze out the tastes of the sugar, lemon, and water that go into it; but they are all fused into the taste of the mixture, which is distinctly different. Consequently, analysis is opposed to fusion.

Dangers to quality vividness and its value. In the process of analysis, the perceiver tends to undermine the felt quality of the whole and destroy it. This destructive tendency is what Croce noticed and is why he was inveigled into excluding concepts and analysis from aesthetic intuitions altogether. Another process that tends to be hostile to the vividness of quality is habit or regularity. A third is, as usual, practical problematic activity. Strangely enough, all three of these intrinsically hostile processes, however, can be used in certain ways to increase the vividness of quality.

Aesthetic value of conflict

The most interesting of these procedures is conflict, the source of most of man's practical activity. In contextualism, conflict is stressed as a positive aesthetic value up to the point at which it begins to precipitate overt practical action. No other aesthetic theory can be said to do this. The usual emphasis is continually on harmony. In contextualism with the emphasis on vivid quality, one of the dangers is the threat of monotony, excessive regularity, and the deadening results of habit. There is nothing so effective in counteracting these results, however, than a shock of conflict cutting across a habit or routine. But even when there is no danger of aesthetic fatigue, the contextualist can recommend the skillful use of conflict for positive aesthetic purposes, notably in an explanation of the high value of tragedy in art. Here, deeply moving drives are set in action such as ambition, love, jealousy, and fear and deliberately confronted with frustrations and conflict, up to the limit that a spectator can endure without an overt practical action (to close the book or leave the theatre). The aim is to increase the vividness of the depicted experience to the highest endurable intensity, for the greatest aesthetic value that it can yield. Tragedy as an art form produces no difficulties for the aesthetics of contextualism.

Aesthetic value of vividness

Having indicated the unique contributions made by contextualism, it may be well to close the discussion by briefly answering the question of the value of mere vividness of quality. The value must rest on something more than the pleasure or satisfaction to be found in it. One answer has been given in a contextualistic book, *Art and the Human Enterprise* (1958), by Iredell Jenkins of the United States. Jenkins argues that the value of vividness lies in its bearing on biological adaptation, on its human survival value: pleasure is subordinated to survival. Indeed, the ultimate human value of pleasure is its helpfulness to human adaptation, but the adaptive value of art in presenting vivid events and situations consists in its making man keenly aware of his environment and so preparing him to cope with it in emergencies and, indeed, in his everyday life as well. This service is most obvious in novels and drama. But all art makes men aware of themselves and their surroundings. As could be guessed, mere entertainment is disparaged in this approach.

Though Jenkins, as one of the latest of the contextualistic aestheticians, most clearly brings out the adaptive value of this approach, John Dewey in *Art as Experience* (1934) was by far the most influential contextualist aesthetician. In this book the social bearings of art are expatiated. All of the characteristic features of contextualism are found there expounded with the deep earnestness of the social reformer that Dewey was in all of his philosophy. But the purity of his contextualism is somewhat overlaid by strong, organistic ingredients (to be taken up presently).

Organistic contribution. Another important recent approach is that known as organicism, in which the chief value of art is found in the integration of experience. Integration also has adaptive value: personality psychology has exhibited the integrated personality as the mentally healthy one. Thus, many have found it tempting to combine an organistic approach with the contextualistic, and this, as just mentioned, is what Dewey did in his *Art as Experience*. A full-fledged organicist, however, concentrates entirely on the integrative function of art culminating in the concept of the organic whole.

Integrative function of art

An organic whole is a structure or form such that every detail in it is so intimately related to every other detail that the whole is destroyed or badly damaged by the loss or disturbance of any one of them. Thus an animal organism depends for its life and health on the mutual functioning of all of its organs together. If the heart, the liver, the lungs, or the brain gives out, the whole organism is injured if not destroyed. And, on this theory, so it is with any great work of art: every detail in it calls for every other detail; all must function together. Indeed, this is probably what Kant was calling attention to in his aesthetic requirement that a work of art should exhibit purposiveness without being an explicit purpose. In an organically structured work of art, every detail functions as a sort of purpose demanding satisfaction in the complementary functioning of the other details. Thus, in the well-balanced composition of a picture, each line and mass and colour reaches across the picture for its compensating mate to find there its own little purposive satisfaction. And so on all over the picture. The result is a cumulative satisfaction of these many little purposes finding their many satisfactions through the organic structure of the whole. There is a feeling of satisfied purposiveness provided internally throughout. And yet the work of art is not itself a purpose in pursuit of an external goal.

In this way a work of art yields a special sort of satisfaction not to be found except in an organic whole. Setting up an organic whole as the ideal of aesthetic organization, this sort of internal satisfaction—to be found only in such an organization of experience—may fittingly be called the aesthetic satisfaction. Practical satisfactions are automatically excluded. Loosely organized entertainment and crude isolated sensual enjoyments are also looked down upon. There are, to be sure, degrees of organicity. An organicist would hesitate, however, to call a work beautiful that was not rather highly integrated in the direction of an organic whole.

Art as a dynamic activity

A characteristic of this conception of the aesthetic experience is that it takes a certain stretch of time to fully realize, perceptually, an organic whole, especially if it is a work of even moderate complexity. The act of creating such a work is a dynamic activity. The artist has to feel his way through it to reach the total internal functioning of all the details that the whole is demanding. Similarly, an appreciative critic must imaginatively follow through the same process to attain the total satisfac-

tion potential in it. Thus it is said that the act of aesthetic appreciation is a re-creation of the artist's creative act; and criticism consists in observing how far the artist succeeded in realizing the organic demands of the details. If a shortcoming is discovered, the failure will be specifically indicated, showing just where the artist failed to satisfy a demand that he himself set up in the composition of his work.

The organistic approach has been steadily winning more and more adherents as time goes on. In its full development, it is a relatively recent aesthetics. Though the principle of the organic whole can be found in the writings of Plato and Aristotle and in other ancient writers, it is there conceived as a rather static sort of thing. Regarded as a dynamic activity with internal purposiveness as above described, it can hardly be said to have been realized before the advent of G.W.F. Hegel, perhaps the greatest of the German Idealists.

Formistic contribution. One of the oldest types of aesthetic theory is that of formism, more popularly known as the imitation theory. It begins with the idea that the artist's aim is to make a copy in some medium such as marble or bronze, paint on a smooth surface, or action and words as in a drama of existent objects and events in the environment.

Imitation theory in Plato and Aristotle. Plato presented the theory in a famous passage of *The Republic.* A picture is conceived as intended to be a faithful representation of its object, the closer the better—even to the point of being mistaken for the real thing. Having described the artist in these terms, Plato promptly degraded him as a socially superfluous person. Many scholars cannot believe that Plato was quite serious in this passage, beyond disparaging a type of illusionistic painting current at the time. Besides, aesthetics as a special study was not even thought of at that time.

Art as expressing the universal

Aristotle soon followed Plato and gave the imitation theory a new twist, which was developed and carried along through the succeeding centuries. He suggested that the artist might imitate the universal form or essence in which particular objects share rather than the physical particular. Aristotle made many other wise comments based on his observations of the arts, so that his complete aesthetic theory should not be identified with any one of them. But this suggestion—that the artist may, and for the noblest results should, imitate the universal form rather than the physical particular of a species—became a central principle of formistic aesthetics. By this means art acquired, as he once said, a universal cognitive role resembling science rather than history. The artist represents the form, not the particulars; he seeks, in this view, to look through the particular to the form. As a great painter brings out the character of his sitter in a portrait rather than a momentary appearance at a certain place, so he seeks out the essential form of a tiger, a lion, a horse, or even of a certain kind of chair, pairs of boots, a social scene, or modes of human action in comedy and tragedy.

Since Plato had already developed a theory of Forms, the general theory is often called Platonism. An important world theory emerged from Platonism, the force of which can be felt when trying to give serious thought to how to explain the kinds of similar objects that are found all about one. The problem is how to explain the relation of species to individuals in classifications of plants and animals. In a flock of sheep all of the same form and a herd of cows all of the same form, every particular instance partakes of the same appropriate form. Though particular animals come and go, the universal form remains and is thus the more stable reality. Plato held that the form is the basic reality, that it has, in fact, a sort of being distinct from that of particulars, which are only the appearances of forms in a lesser sort of being caught up in a flux of change. Aesthetically (when he was promulgating his imitation theory), Plato made the mistake of conceiving the artist as imitating the appearance of a particular. Aristotle corrected this by conceiving the artist as imitating the universal form as seen through the particular.

This insight of Aristotle's has been widely held and developed ever since. Among other attractive features of this theory for aesthetics is the fact that it furnishes a ground of objectivity for the aesthetic judgment and experience: for a form not only explains the relation of particular occurrences and appearances to universal types and laws of nature but also produces objective ideals of evaluation; the particular individuals of a species fulfill their forms with various degrees of completeness. A good specimen of a species is often hard to find. A collector of butterflies, for instance, will often have to discard many individual specimens before he gets a "perfect specimen" to include in his collection. A form sets up an ideal for its species.

Styles and social connections of art. The same principle carries over into the arts. Here also there are forms of art. These are easy to see in the crafts in terms of their technique in the service of norms of utility. The forms constitute ideals of perfection, which are sources for the production of styles of art. Formistic aesthetics has thus given a great incentive to the study of styles of art. It can be shown in detail how styles of art gradually grow and, in a sense, discover themselves and come to a culmination. The greatest works of art are presumed to be those produced at the culmination of a style, such as the Parthenon of Athens or the Cathedral of Amiens. Yet the innovators are not without praise, as well, in their insight into a new ideal to be aimed at, though their experience is still inadequate to attain it. Some formists offer evidence that the culminating works of a style exhibit the principle of the organic whole. This whole is conceived by them generally, however, as a static, passive, organic whole already subsisting in the norm—in contrast to the dynamic organic whole of the organicists.

Pioneering, growth, and culmination of styles

The formistic interest in styles of art leads naturally into an interest in the social connections of art; for a style of art tends to reflect a type of social organization as well. In architecture the mutual interplay of art forms and cultural pattern is particularly obvious. Such interrelationship permeates all of the arts, however, at least to some degree. This social phase of the aesthetics of formism is exemplified in the work *Philosophie de l'Art* (1865; Eng. trans., 1865) by Hippolyte-Adolphe Taine, a French deterministic psychologist. That he thought of himself as a Positivist—*i.e.*, as a purely scientific aesthetician—is just that much more interesting and to the credit of certain possibilities of the formistic approach. Taine was concerned with merely describing the types of art that exist, correlating them with their physical and social environments. And the outcome, as he abundantly disclosed, was the exhibition of the essence, the central form, not only of each style of art but of the individual superlative representative works done in that style. Great art, he held, emerges as an approach to perfection in the representation of the essence of a style.

The classic work on the aesthetics of formism, however, is Aristotle's *Poetica.* Probably no book has been so widely studied and so steadily influential as this one. It gives the deceptive impression of being just a series of reflections on the reasons for man's interest in the arts, particularly in drama and in the great tragedies of the period. But the whole philosophy of formism permeates it. In spite of vigorous criticism, the formistic theory of art is alive today. It is, for example, the official view of the Scholastics, represented by two of the foremost authorities on medieval philosophy, Étienne Gilson, in his *Painting and Reality* (1957), and Jacques Maritain, in his *Art et scholastique*, new ed. (1927; Eng. trans. 1930). In his later years, Santayana became a formist with a philosophy centred on essence. Though no exclusive treatment of aesthetics appeared during this phase of his thought, his writings on the "Realms of Being," including *Scepticism and Animal Faith* (1923), frequently refer to the aesthetic value of essence as immediately intuited.

Recent expressions of formism

Other contributions. Two other approaches to aesthetic value are those that proceed by way of emotional expression and mystical experience.

The expression theory. It was noted earlier that Croce's theory of the identification of Intuition and Ex-

pression split his successors two ways: one group stressing the intuition of quality, leading to the contextualistic aesthetic; and the other stressing expression, leading to a movement led by E.F. Carritt (*Theory of Beauty*, 1914) and R.G. Collingwood (*The Principles of Art*, 1938). The central doctrine of expressionism is that a work of art is an expression of the artist's emotion, which, if successful, can be communicated to the public or at least becomes capable of communication for those prepared to respond to it.

Emotional power of art

This is by no means a novel theory. The power of artists to arouse and communicate emotion and to embody it in works of art was noticed by Plato and Aristotle and by the Greek and Roman rhetoricians. What may be new is the construction of a whole aesthetic theory out of the observation. It was done, even before Croce's followers picked up the idea, by Leo Tolstoy, a Russian novelist and philosopher, in a much read book of his later years, *What is Art* (1898). Here Tolstoy pleaded eloquently for an art that elevated men with a deep moral and religious emotion free from worldly frivolity and carnal dissipation. In it he even repudiated his own great novels. A genuine work of art, he now held, should move men, by a sort of contagion with the emotion embodied by the artist in his work, toward faith in the universal religion of the brotherhood of man. His book *What is Religion* then became a companion piece to *What is Art*, showing the ultimate criterion of the value of art. The simple peasant by his closeness to nature, uncontaminated by the conventions and hypocrisies of the upper classes, becomes the only reliable critic of the essential sincerity of a work of art.

R.G. Collingwood, a correlator of philosophic and historical thought, has much the same theory—relieved, however, of Tolstoy's evangelism. In his view, the essential embodiment of sincere emotion and the importance of its communication with the work of art as the vehicle are both retained. Art becomes a medium of emotional communication on a level with the conceptual verbal communication of the sciences and of practical life—or, better, of commercial life: for nothing could be more practical than emotional communication when this is needed. And art can effectively communicate many things that are conceptually incommunicable.

Authenticity of emotions

The expressive theory of art is easy to explain in principle but leads to many confusing ramifications in detail. The heart of the theory is always an appeal to the genuineness of the emotion expressed. Superficial emotional expression or mere stimulation in entertainment is frowned upon. And insincere emotion is the ulcer of ugliness. The Existentialist term of an "authentic" emotion may be the most appropriate.

The problem is to describe or designate this crucial evaluative emotion and its expression. Tolstoy, as shown above, tied it in with his conception of a supposedly universal religious expression. Croce stressed its completeness of expression, a special sort of intuitive emotional clarity that can only disclose itself (even to the artist) by its fullness of expression. And he insisted on the exclusion of all conceptual or practical material.

Collingwood was not so exacting. He admitted intellectual material if it was sufficiently fused into the emotion. (Indeed, once in a while Croce did the same.) He had much to say about the creative activity of Imagination and sometimes came close to the organistic criterion of genuine aesthetic emotion in terms of an organic whole. His central opposition was that between the "corrupt" and the "uncorrupted" emotion—representing repressions, according to the evidences of the psychoanalysts, as a particularly insidious source of corruption. He allowed also for successful private expression of genuine emotion, so that communication is not essential; but some sort of embodiment of emotion is.

The mystic theory. When emotion is given so central a position in aesthetics, it is not far to go to the mystic theory of art based on the mystical experience as the immediate experience of God and ultimate Being. That mystical experiences do occur to many individuals and that they have the noetic (or knowledge) character ascribed to them—involving a conviction of an indubitable disclosure of the total significance of the universe in an ecstatic all-embracing intuition—is a well-attested fact. How far it may be credited as a true revelation is another matter. Moreover, the full experience—or some approximation to it—has inspired fine poetry and painting and music in man's cultural heritage. Artists as a group are fascinated by the theory. It is the dominant aesthetic view underlying a large part of Oriental theory and practice in the arts. With such an impetus behind the view, it is surprising that only one or two outstanding philosophers have espoused a mystic aesthetics in Occidental thought—the 3rd-century philosopher Plotinus, the foremost representative of Neoplatonism, and the German philosopher Arthur Schopenhauer, known for his pessimism. Moreover, both of them offered what amounts to a Platonic formist theory of aesthetic experience. This experience consists in the disinterested contemplation of the Platonic Forms in their calm detachment from particulars that are caught in the flux and turmoil of the world of appearance. The only mystic motive is that these Forms may act as symbols or rungs on the mystic ladder leading to communion with the unity and peace of ultimate Being.

The aesthetic and the Holy

Mystic aesthetics may be classified as a kind of emotive theory in which the criterion for the supreme experience is that of the mystic experience itself in complete ecstasy. Yet all mystics in the exposition of their aesthetic views seem to agree that the supreme experience itself is not an aesthetic experience. Thus Rudolf Otto, a German phenomenologist of religion, has set up, as a fundamental category, "the Holy," as distinguished from the aesthetic. It is as if the mystics feel that an identification of these two would be a sacrilege. Besides, the experiences of beauty in communion with nature and through the arts have been thought to be clearly distinguishable from the ultimate ecstasy that, when attained, absorbs all of these lesser experiences into its ineffable unity. The possibility of attributing a special high value to these lesser symbols of the ultimate Reality would appear to depend on their not being conceived as merged and lost in an ultimate ineffable One. Thus, there has been an inevitable tendency for mystical philosophers to borrow the details for the lesser levels of being from other world hypotheses.

The more simple religious mystics have borrowed profusely from the myths of animism, the view that natural objects are living and conscious. Their more sophisticated colleagues have borrowed from the more adequate world theories—particularly from Platonism and organicism. There are many organicists with strong mystical leanings, such as the Idealists F.H. Bradley and William Ernest Hocking; but none of them has shown much interest in aesthetics. So, again, Plotinus and Schopenhauer seem to be the most adequate examples of aesthetic mysticism. William Butler Yeats, Irish poet and dramatist, may be cited as a pure mystic who may almost be called a philosopher.

In mystical aesthetics beauty is found in the symbols of mystic peace and unity to be felt in nature and in works of art. For some appreciation of the efficacy of these symbols, one is best served by the poems and the paintings of the mystics themselves. In particular, classic Chinese painting is saturated with such mystical symbols. One of the most striking of these is the presence of great empty spaces in this art. When the viewer concentrates on these spaces in the context of a serene ink painting, for example, they cease to be mere vacancy. The mystic artist renders them alive in their emptiness: they are the symbols of the negation of all appearances and of union with the undisturbed unity of Nirvana.

LINGUISTIC APPROACHES TO AESTHETICS

The foregoing expressive emotive theories facilitate the transition to a vigorous new trend in aesthetics that goes under the general heading of the Linguistic movement. There are many facets to this movement; all are centred on questions of meaning and language, which, in turn, may be fairly said to be concentrated in the question of

Nature of an aesthetic judgment

what is involved in an aesthetic judgment. Answers to this question have led to close analyses of the words, sentences, and meanings used by critics and aestheticians in their discussions of art and aesthetic value.

The outcomes of these analyses often do not differ greatly from conclusions to be found in the theories already reviewed. There are two conclusions of Linguistic Analysis, however, that have been regarded as revolutionary in effect and have been very influential in English and American circles. One of these is the doctrine of the "naturalistic fallacy," and the other the emphasis on "the open question" and "the open textured" concept.

The Linguistic movement and the naturalistic fallacy. Though the naturalistic fallacy has had its chief impact on ethics, it is equally applicable to aesthetics and to value theory in general. It is primarily associated with an analysis of what can be meant by the term "good" developed in a book *Principia Ethica* (1903), by George Edward Moore, a careful pioneering Analyst in England. Briefly, his argument was as follows: the meaning of good cannot be identified with any natural experience such as pleasure or satisfaction of desire, for it makes sense to ask the question "Is pleasure good"; hence the proposition that good is equated with pleasure is either meaningless or false. The same argument would apply to any other natural entity—to satisfaction of desire or expression of authentic emotion, to harmonious integration of feeling and impulse, to perfection of form, or to mystic ecstasy. It is always meaningful to ask whether any of these is good. Hence "good" cannot mean the same as any one of these natural entities. The proposal or assumption that it can is "the naturalistic fallacy." Moore thought that the only way to escape this fallacy is to resolve that "good" means good and nothing else. One can have an undefinable "intellectual intuition" of the universal, formistic Platonic essence good itself. And this, according to Moore, is the basis of value.

It was soon pointed out that the identification of the meaning of "good" with a Platonic universal was also open to question by any philosopher who questioned the truth of the Platonic theory of universal Form. Moore himself had committed a naturalistic fallacy. Besides, the hypothesis that an intellectual intuition exists to be aware of this universal good had little evidence to support it.

Emotive theory of normative statements

But to many, Moore's linguistic analysis of "good" still seemed to have something in it. A Logical Positivist, A.J. Ayer, another English philosopher, in his *Language, Truth and Logic* (1936), proffered the idea that the naturalistic fallacy consists in thinking that a value judgment such as "Generosity is good" or "Lying is bad" may not be true or false like a scientific proposition or common declarative sentence. True, the word forms look like declarative sentences such as "The water is boiling" or "Diamonds cut glass," but actually value judgments are merely emotive expressions like uttering "Generosity, WOW!" "Lying Ouch!" In aesthetics, likewise, "The sunset is beautiful" really means only "Sunset, WOW"; and "A gangrenous leg is ugly" really means "Gangrene, OOF!" The upshot of this sort of analysis is to draw a sharp line between facts and values and to declare that a science of values or any study of values purporting to be true (or false) is impossible. Books on ethics and aesthetics are little more than expressions of likes and dislikes, wishes and dreads, demands and rejections. In technical terms they are made up of optative, imperative, and exclamatory sentences—which are neither true nor false—concealed under the appearance of true (or false) declarative sentences about matters of fact.

The emotive theory of value, including aesthetic value, appears to be the most skeptical theory of value ever promulgated, though this is not a reason for rejecting it.

Pluralistic nature of value

Many scholars now hold that the supposed naturalistic fallacy simply calls attention to the natural ambiguity of a very important word of ordinary language that is the general term for values of all sorts. There is no necessary need to identify the meaning of "good" with one kind of value alone and reject all other meanings—nor, indeed, to be led finally to reject all meanings, leaving the term apparently meaningless. The problem of value is rather to discriminate clearly the various meanings and kinds of good and the occasions for their appearance and application and to determine the relative priorities of these various values for each occasion according to the relevant evidence available. Whenever the evidence is not conclusive, the reasonable procedure is said to be to hold the various hypotheses of value priorities in suspense; *i.e.*, to hold them open pending additional evidence. As regards aesthetic value, one hypothesis gives priority to pleasure of some specified sort, another to vividness of quality, another to organic unity, and so on. But each recognizes that there is a contribution to aesthetic value in what the others hold in priority. The spectator, the critic, and the artist may all profit from taking each view seriously but not dogmatically. There may be biological or social conditions in which each of these values legitimately acquires priority in fact: under prosperous conditions when social pressure is low, pleasure may be the value that can safely be given priority over other aesthetic values; but under conditions of social stress, the adaptive value of a vivid feeling for the actual quality of events may gain a priority and pleasure might seem superficial. A reasonable conclusion may be to hold the hypotheses open for whatever relevant evidence may come.

The Linguistic movement and the open question. The most constructive contribution of the linguistic movement for aesthetics and other value disciplines may be its stress on the "open question." The trouble here was that even this could be carried too far. When the concept of the open question became a dogma, it ceased to remain open. It can be carried to such an extreme that aesthetic evaluations of better and worse in art become impossible. For a definition of aesthetic value or of a type of aesthetic value is the sole linguistic basis for any criterion of evaluation.

Definition of value as a standard

The question may be asked, for instance, how judges could decide on a prize dog or bull without a definition understood by them all of the "points" for the sort of dog or bull being judged. The same question may be asked of art juries selecting pictures for exhibitions and for aestheticians writing about ultimate aesthetic values. Their theories culminate in definitions of aesthetic value that constitute the ultimate linguistic standards of beauty from their point of view. These definitions, as has here been urged, are not to be taken as dogmatic but as hypothetical. Their adequacy depends on the evidence available to support them. They are descriptive definitions summarizing the conclusions of systematic studies in aesthetics.

If the open-question attitude, however, is carried to the extreme of discounting evaluations of art made from firmly formulated definitions of any sort, then the liberal intention deteriorates into a destructive dogma. An evaluative definition need not be loose and open-ended itself. If it is regarded as an hypothesis open to revision as new evidence may direct, it is still open. But it is at the same time firm and unambiguous in its evaluations.

Some linguists insist that the relevant concepts of art are all of the type called "family resemblance concepts." Ludwig Wittgenstein, a leading philosopher of language, brought this notion into prominence. Using the concept of "game" as his paradigm, he showed that a sharp boundary line cannot be drawn around the field of instances that all men would be willing to denominate "games," that there are no common traits characteristic of all games. All games have traits common to some of the others, but not to all of them. The resemblances overlap. Yet in common living men gather them all together into a single concept because they overlap so easily. Ordinary language is full of such concepts. This is therefore one particular way in which a concept may be open. In fact, by virtue of its very overlapping structure, such a concept cannot be closed.

Some aestheticians argue that all aesthetic concepts are, or should be, of this family-resemblance type. Many such concepts unquestionably are, especially in ordinary language. And there is a school of linguists who have

been saying that ordinary language concepts (except in the sciences) are the only correct concepts. Thus value concepts must be left open in the family-resemblance way. It is claimed that closed concepts would dam up the creative urge of artists, and it may generally be agreed that there is a modicum of truth in this view, because the established schools of art generally have developed clear standards of style, whereas the innovators are not yet fully aware of what the culmination of their new style will be. But, of course, any new style is actually seeking for its new standards, which will sometime become clarified. It still remains true that relatively adequate theories of aesthetics have been developed with clear, firm definitions of what constitutes excellence in art. Though closed in themselves, these definitions are best taken as hypotheses open to confirmation and revision in terms of any new available evidence. It is thus not necessary that aesthetic standards be kept deliberately vague and indecisive in order to be also free from dogmatization and open to revision and reconsideration.

Importance of definitions of aesthetics

It was stated earlier that a definition of value can act as a standard of value for the field of values that it defines. Some simple instances were shown. The point needs to be amplified a bit, however; and this can be done by summarizing the importance of firm, clear, closed definitions for a fully developed aesthetic theory. A definition acts as a standard of aesthetic value in two ways: first, it demarcates what are or are not objects of aesthetic value; and second, it determines the amount of aesthetic value that these objects have whenever the defining terms admit of quantitative differences. For instance, for the common hedonistic view that defines aesthetic value as experiences giving immediate pleasure, or as the liking of things in themselves, it would appear to follow that things liked as means to other things are not aesthetic values. Similarly, sentimental objects valued only for their associations, such as the pen used by Lincoln to sign the Emancipation Proclamation, are not aesthetic values. Such a definition is a powerful standard. It keeps many objects out of art museums and many musical scores from ever being played at a concert. But the definition also determines the amount of aesthetic value that an object has in terms of the amount of pleasure or liking that it gives.

This definition, as has been seen, is often regarded as too wide and has to be qualified in various ways—to distinguish, for instance, deep or superior beauty from mere entertainment. The justification of these added distinctions depends on the relevance of the evidence that can be brought up in support of the definitions. This is taken to show that the definitions are actually hypotheses about the place of aesthetic values in relation to other values and, indeed, to all of the other facts in the world.

If such definitions, which function as the standards of beauty and ugliness, are not firm and backed up with adequate evidence to justify their being used as they are in the arts, then there would appear to be no justifiable aesthetic standards at all. Some skeptical scholars have argued that this is in fact the case; the conclusion seems nonetheless improbable in view of the great body of constructive work that has been done to date in the field.

(S.C.P.)

ORIENTAL AESTHETICS

General seminal features of Oriental aesthetics. The distinction between the Orient and the Occident is often made on a simple, geographic basis, as in the division of the Earth into two equal parts. The terms Oriental art and Oriental aesthetics would then apply to all art and aesthetics produced in the Eastern Hemisphere. With advances in exploration and cartography, however, the inadequacy of that simple division for the study of human cultures was long ago disclosed. It was supplemented by the concept of Near Eastern, embracing the lands and cultures around the Nile, the Mediterranean, the Tigris–Euphrates, and sometimes even the Mohenjo-daro culture (about 3000 BC) of the lower Indus Valley. Islāmic influence in the arts later spread eastward through Iran, some of western India, and southern Spain. From the biogeographic point of view, the Orient is the land between the southern Himalayas and the western Malay Archipelago. The term Far Eastern usually includes the lands and cultures of India, Nepal, Ceylon, Indonesia, China, Korea, Tibet, and Japan.

Pragmatic and exotic terminology

Many unscientific terms, based on the inaccurate observations of travellers, were incorporated into discussions of the arts. In some cases, they served a useful purpose, as in classifying particular works and styles in art museums. It became a never-ending task, however, for aestheticians and historians to describe and compare the innumerable styles and movements in Far and Near Eastern art from Paleolithic Asia to the present time.

The concept of aesthetics eventually outgrew its former restriction to definitions and judgments of beauty in order to join in the larger task of discovering the nature and potency of art in a changing social world (see ART, PHILOSOPHY OF). Surmounting linguistic difficulties, modern scholars gradually appropriated foreign words from exotic languages, such as *zen*, *moribana*, *yoga*, and *rasa*, along with the ideas they signify until the foundations of an international aesthetic language had been laid. To some extent, this practice was a continuation of an earlier one in which, for example, Italian was the language of music and French that of diplomacy.

Modern civilization, with its desire to understand and control things by means of reason, had its beginnings in Egypt and Sumeria (in Mesopotamia). Early hints of a philosophy of beauty in Europe, before Pythagoras, a 6th-century-BC number philosopher, included so much mysticism in regard to mathematics as to suggest a previous influence from the Orient.

Religious foundations. As a factor in the social context of art in any moment of history, religious and philosophic thought is highly significant. It is necessary only to list the great, advanced religions of Oriental antiquity in order to realize how varied must have been their inspiration to artists of all sorts: specifically, the Brahman, Buddhist, Jain, and Hindu faiths; the Zoroastrian faith in Iran; the ascendancy of Hinduism in India as Buddhism shifted to the north; the importation of Judaism and Christianity from the Near East; new variants of Buddhism in Tibet and Japan; Shintō in Japan; Islāmic beliefs spreading from Arabia to the East in India and to the West in Spain. All of these religions were producers of art at certain times and places, with long or short florescences. Although the concept of an evolution of religion is controversial, numerous authorities have held that many of these religions went through a primitive stage of animism, magic, and realistic polytheism, each with its legacy of expressive imagery, before passing on to monotheism and pantheism. These last were increasingly devoid of vivid imagery.

Much Indian art and criticism obviously had a religious form and content, which is evident, first, in the countless myths and legends employed as represented subject matter, such as stories of the ancient hero Kṛṣṇa and the cowgirls; second, in the willingness of artists, rulers, and public to accept fantastic representations of the supernatural, as in the several arms of Śiva, a major Hindu deity; third, in the emphasis placed upon mystic symbolism in the arts, including dance and theatre.

On the whole, within an extensive culture such as the Greek or Indian, the arts tended to develop first partly for magical and religious purposes. Theoretical discussions of them—including instructions from priests and other patrons of the artist—came later. Philosophic discussions of their aims and standards of value, as in Aristotle's *Poetica*, tended to come relatively late. (Such a relative position in the evolutionary sequence is not, of course, the same as a date. At the present time, many social groups and individual writers are at relatively primitive stages in cultural development.)

Theism and naturalism

Though it has been said by some historians and denied by others that all Indian art is religious, it is true that most Indian art can be given a religious interpretation by persons of a strongly theistic attitude. This was not necessarily the intention of the artist or of his public. Religious interpretations can be applied to the most humble, utili-

tarian products as to the most magnificent palaces. A work of art can be appreciated on various levels by the educated observer, and one of these levels or attitudes may be supernaturalistic, another naturalistic. Many Indian works and their interpretations, which include ideal representations of gods and spirits, are clearly theistic. But all personifications of the gods can be regarded as avatars, or aspects, of the cosmic spirit, as explained in the Vedas (early religious poems), the *Upaniṣads*, and other sacred writings. Primitive polytheism could thus give way to what Western philosophers call Idealism, panpsychism, and pantheism.

Sensory, naturalistic, and humanistic expressions. Although many Indian theories of aesthetics disparage the realm of sensory experience as mere illusion, Indian artists continue to employ sensory materials and to display them for the enjoyment of others. (The same can be said of medieval Christian art and aesthetics.) In trying to explain the arts, supernaturalists tend to feel that they reflect a divine Creator, whereas naturalists—as in the poem *De rerum natura* ("On the Nature of Things") by Lucretius, an ancient Roman Atomist—regard them as reflecting man alone.

Neither India nor the Far East in general ever developed a strong, enduring naturalistic movement, such as might have led to empirical science. It had a brief one, the Cārvāka school of philosophy, which flourished about the time of the Greek naturalist Epicurus (died 270 BC), to whose philosophy it was similar. It was Materialistic in metaphysics (theory of the nature of Being) and hedonistic, or pleasure oriented, in ethics. But these and possibly other attempts of the sort were overwhelmed by the tide of mysticism that swept over the Far and Near East, holding sway until the 20th century. (Confucianism in China, though lacking a strong theistic basis, was consistent with a popular belief in ghosts and minor spirits.)

In the ancient military empires, a main purpose of the arts was to glorify the ruler and the noble, wealthy classes. This effort was restrained to some degree by priests and philosophers who urged that moderation, justice, and sacrifice were better ways to happiness in the next life than costly works of art and "vain oblations." In some times and places—particularly in China, which witnessed an increasing humanism and rationalism (though with ups and downs through the ages)—a clear-cut issue arose between the ethics of religious asceticism and that of humanistic art. Some kinds of art were said to be anagogic; *i.e.*, able to lead the mind upward to heavenly things—an aim similar to one of Dante's in *La divina commedia*, in which his guide, Beatrice, personifying philosophy, leads him heavenward.

Universal and mystic symbolism. One way in which art can perform this function, according to the ethics of supernaturalism, is by helping the student understand universal symbolism. Although monasticism flourished throughout most of the Orient, especially under Buddhism, it took various forms and prescribed different spiritual exercises, in which chants and rituals had a place. Mural paintings, scrolls, and stone or bronze sculptures helped to stimulate the desired attitude, as a means of reaching heaven or Nirvāṇa. Prolonged contemplation of an abstract design, such as one of interlacing triangles, could also help to induce a mystical attitude. Where the theory of reincarnation was accepted, many people thought that art—under the authority of priests and officials—could be a means of acquiring merit in this life, to assure a happier life in the next rebirth.

Symbolic religious figures

Another basic concept of Oriental aesthetics was that of mystic symbolism, also much studied in medieval Europe until well into the Renaissance. In its ordinary usage, the word symbolism refers to the practice of using a familiar image, such as a lion, to mean something very different, such as the abstract quality courage. In Christian art, the figure of a lamb is used to signify gentleness, meekness, innocence, and sacrifice—all as attributes of Christ; in India, the bronze figure of the dancing Śiva is well-known as a symbol of that god, a figure that signifies the cosmic rhythm of creation and destruction as symbolized by a drum—especially its destructive aspect through an emphasis on fire. Śiva, a favourite subject of Hindu sculpture, combines male and female attributes, such as those of Mars and Venus in Lucretius' poem.

Many early Christian artists and theologians accepted the doctrine of universal symbolism, attributed to St. Augustine and St. Clement of Alexandria, early Church Fathers. It taught that everything in the universe, especially important images in the Bible and other sacred writings, had not one meaning but many, as in the Book of Ezekiel and in Revelation. These meanings, it was thought, had been divinely established for man's guidance. Most of them are cryptic and hard to decipher. The process of doing so can be along naturalistic or mystical lines. In the Orient and in medieval Europe it was usually the latter.

Mystic symbolism refers not only to the cryptic meanings attributed to the symbol but also to the way in which the symbol-meaning combination is formed, grasped, and appreciated. In general, mystic experience is conceived as occurring not by sense perception but through a special, perhaps divine revelation, as in a trance or unsought vision. It is thought to open the observer's mind to some ineffable higher truth and level of reality. Works of art, it is thought, can be so perceived if the meaning is genuinely of divine origin.

In India and Tibet, various types of bodily and spiritual exercise called Yoga and Tantra were developed and are still used as a means of freeing the mind from unwanted desires and controlling inner experience in a positive way. The method called Tantric combines Hindu with older, prehistoric elements. It emphasizes the role of the female as a source of physical and spiritual energy. Some of the finest Indian sculpture shows Tantric influence. In Europe during the Council of Trent (1545–63), church authorities, noting the often farfetched attributed meanings with which the arts had become increasingly laden in medieval times, took steps to discourage the universal symbolism. From then on, this symbolism declined in Western art, although traces of it survive in the paintings—especially Flemish—of the early Renaissance. In the Orient, on the other hand, mystic symbolism remained a vital force in aesthetics and philosophy until the 20th century. Though it is now being gradually replaced by Rationalistic science, even those scholars who reject it as a mode of ascertaining truth may still value it as a source of aesthetic satisfaction on a high cultural level.

Specific cultural expressions of Oriental aesthetics. The principal cultures here reviewed are those of China, Japan, India, and contemporary Marxist lands.

China. There is a distinct analogy between the aesthetic theories of Confucius (died 479 BC) and those of Plato (died 347 BC). Both sages discoursed on music and ritual as means of producing desired kinds of emotion, attitude, and character. They showed how young men could be educated to strengthen and harmonize the state. Both praised music of a stately and dignified sort, which helps to produce inner harmony. Confucius emphasized the value of traditional ritual in the development of emotional self-control, and his follower Hsün-tzu stressed the need for simplicity in the arts as leading to a well-ruled state. Both Plato and Confucius denounced the seductive, exciting effects of the "modern," complex music of their day.

Confucian and Taoist aesthetics

Austere Confucian critics and officials condemned all novel writing as licentious, but popular novels, nonetheless, were written in China. In the long conflict between Confucian and Taoist schools of thought, animated arguments developed, which contributed vitality to this stage of aesthetic thinking. Lao-tzu, a founder of Taoism, disapproved all art as blinding the eye, deafening the ear, and cloying the palate. Rare treasures, he said, tempt men to do wrong. His severity toward art was extended to all attempts at changing conditions. Later Taoism became freer and encouraged the love of nature as well as the artistic depiction of deities and minor spirits. As against Confucianism, Taoism proposed a less formal, more intuitive approach to life in general. This approach was carried over into the arts and into theoretical discussions of them. On the whole, there is more humour and

lightness of touch in traditional Chinese than in Indian philosophy.

The movement called Ch'an in China and Zen in Japan denounced scientific technology for preventing a simple, direct acceptance of the world as it is. Though Zen has been interpreted in both a naturalistic and a mystic way, it has avoided the attempt to define its methods clearly.

In general, Chinese ideas on art and aesthetics were written in shorter, more epigrammatic rules and comments than those of Hindu thinkers. In India, aesthetic discussions sometimes formed parts of lengthy philosophic systems. In China around the turn of the 6th century, Liu Hsieh, a Buddhist scholar and critic, showed how literature has developed from the simple to the complex and ornate; how it expresses the spirit and socio-economic conditions of its time.

Japan. Although early Japanese art and aesthetic theory owed much to China, in later centuries more original features developed in Japan. One of the most striking is the great novel, *Genji monogatari* (*The Tale of Genji;* AD 1000), by Murasaki Shikibu, a lady-in-waiting to the Empress. It is great not only as a work of art but also for its theoretical discussions of good and bad features in various arts.

By the 18th century, many aesthetic problems were discussed in a sophisticated way, with recognition of the need for different methods. Appropriate words recorded the perceptive, emotional responses of sensitive observers to different kinds of art and nature, including human nature. They also provided a combination of descriptive and normative meanings, capable of use in characterizing a work of art and relating how an observer feels about it. The many value concepts of this sort, provided in Chinese and Japanese theory, contain a rich vocabulary for use in the criticism of various arts.

Descriptive and normative concepts

Many of these concepts were set forth by Zeami Motokiyo (died 1443), a Japanese playwright and actor-manager. He used the traditional Nō play as a symbolic theatre in which ideas and moods are suggested but not clearly represented. In the late 18th century, Moto-ori Norinaga, the leading scholar in Shintō and Japanese classics, emphasized a quality called *aware*, meaning gentle melancholy or sensitivity to the beautiful sadness of things. Other aesthetic qualities are *en* ("charming"), *okashi* ("cheerful," "amusing"), and *sabi* (having the beauty of old, faded, worn, or lovely things). These value concepts invite comparison with the Hindu *rasa* (flavour) and with such words as picturesque in European Romanticism.

The concept of *yūgen* ("mystery and depth") was developed by Zeami into a far-reaching principle, involving elements of several religions. Zeami assumed the animistic idea of a collective, cosmic mind, a kind of spiritual power underlying all parts of the Nō play and expressed in the work of art.

D.T. Suzuki, a present-day intrepreter of Zen, links these concepts to the Buddhist and Taoist absolute passivity in accepting things as they come. On the other hand, the popular art of the 18th and later centuries belongs to the so-called floating world of transitory phenomena. To the Buddhist, it symbolizes the ephemeral world of sensuous pleasures.

India. In India, the legends and myths of Kṛṣṇa, the great god who is also a mischievous child, a heroic warrior, a divine lover, and a religious teacher, fill many volumes. He is worshipped in countless rites, dramas, ballets, and folk festivals. The Kṛṣṇa legend as a whole is far more than a cycle of stories; it is a running thread through most of Indian civilization and a key to understanding that civilization at its best.

Theory of *rasa*: aesthetic flavour

In the psychology of art, an integral part of modern Western aesthetics, the medieval Orient has much to offer, especially in its theory of *rasa*, or aesthetic flavour. Recent Western psychologists have so strongly emphasized Behaviorism and externally observable phenomena that they have almost ignored the inner world of individual subjective experience. Indian philosophers call attention to the importance of *rasa* in the total experience of creating and appreciating the arts, especially the theatre arts, and to the many kinds of *rasa* and ingredients in *rasa*, which, as they are felt and cultivated, are enhanced in value.

The theory of *rasa* was first proclaimed by a mythical sage-priest, Bharata (*c.* AD 500), and developed by Abhinavagupta (*c.* AD 1000), a great rhetorician and philosopher. It was used in explaining the forms, techniques, and varieties of poetry and theatre arts. Bharata listed as the principal feelings of human nature: delight, laughter, sorrow, anger, heroism, fear, disgust, and astonishment. They are transformed by art into the *rasas*: erotic, comic, pathetic, furious, terrible, odious, marvellous, and quietistic—qualities into which ordinary feelings can be analyzed. Psychological studies of this sort were not purely mystical; but the development of aesthetic ability and sensitivity could be made a part of Rāja-Yoga, a kind of spiritual exercise, which rejected the supposed antagonism between sensuous and spiritual enjoyment. Bharata interpreted the power to feel and understand beauty (to taste *rasa*) as a reward for merit in a previous life.

On seeing the richly carved and painted ornamentation of Hindu temples, Western travellers are often puzzled by what they regard as obscenities, involving complete nudity and extremely erotic scenes carved and painted in full detail. They cannot understand the claim of Hindu scholars that such art is "religious" and "spiritual." The same may be said of erotic Indian literature, such as that of the *Kāma-sūtra*, written about AD 300 (Eng. trans., *The Kama Sutra of Vatsyayana*, 1883, by Sir Richard Burton). This distinguished treatise on love and pleasure was intended as a technical guide to sexual enjoyments and to the sensuous delights of perfume, music, and lyric poetry as auxiliary arts. It also presents a realistic account of amatory relations between the sexes in ancient India. Sensuous pleasure is not proposed as the highest good, nor is it morally disparaged. It is respected as a fitting element in the good life on a certain age level and in the institution of marriage.

Western art and aesthetics have recently tended to follow Oriental customs in approving greater freedom of sexual representation, but they seldom call it spiritual. Whatever term is used, it is a psychological and cultural fact that Orientals do not regard their erotic art as pornographic in the Western, popular sense. Sex as a means to love and joy is regarded as a basic fact of life, religious in that it carries on the divinely established rhythm of creation and destruction. It can be transformed into fine art by a capable artist.

Marxist lands. It may be asked whether the Soviet Union and China are now Oriental and whether they are productive in art and aesthetics. Briefly, it can be answered that, geographically, both nations occupy large areas of Oriental territory. Culturally, China has inherited much from ancient Oriental traditions; and Confucianism, which was a major influence until recent years, has affected all of the arts as well as philosophy and ethics.

Western influence in the U.S.S.R.

The Soviet Union, instead, has inherited much of western European civilization, especially its arts. With this aid, it has, at times, led the world in film and ballet. The Soviet Union has also inherited much of German philosophy and British economic liberalism. Indeed, the tradition of cultural evolution, deriving from G.W.F. Hegel, the foremost Idealist of the 19th century, led directly to the modern conception of aesthetics as a branch of philosophy, distinct from art and from particularized art criticism. But, whereas Hegel's philosophy was Idealistic, Karl Marx has transformed it into a kind of Materialism. In many respects, Marxism has been the opposite of Oriental mysticism, subjectivism, and Taoist laissez-faire. Hence, its influence on art and aesthetics has not tended to follow Oriental lines.

The dominant Marxisms of the Soviet Union and China, however, have had aesthetic implications in their own right. Briefly summarized, these are that one main aim and standard of value in the arts is to reinforce and strengthen the revolution along Marxist lines and that another is to enrich and improve the cultural life of the workers. Toward these ends, the arts have been frankly employed for indoctrination.

While acknowledging their debt to the past, Marxist theoreticians emphasize that they aim to create a new culture—not merely to admire and follow the old. In both Communist countries, there has been a long period of competition between the old and the new, with disagreement on the extent to which any departure from orthodox Marxism (any revisionism) is to be allowed. The newer bases for cultural advance, especially in Maoist China, tend to reflect an emphasis on local needs and the development of technical skills through practical education. Both countries encourage folk arts on a local basis, as well as urban performances on a high level of proficiency. Both use the arts to glorify the conception of the worker and other aspects of Communist ideology. Marxist aesthetics has stressed the need for art (especially for music and drama) that will appeal to the masses and inspire them to revolutionary attitudes, as opposed to artificial formalism. Socialist realism can be a means to that end.

The arts as reflecting and subserving social conditions and aims

Such issues have led to prolonged discussions among Soviet aestheticians—as, for example, in judging from a Marxist point of view the merits of the novelists Leo Tolstoy, Honoré de Balzac, and Voltaire; the composer Ludwig von Beethoven; and the painter Honoré Daumier. They have given moderate praise to past artists who were regarded as revolutionists or fellow travellers before their time. But they have also insisted that all art, including that of radical artists in other ages, is much influenced by the social conditions and tastes of the time. In the Soviet Union, much scholarly writing is devoted to working out the details of this hypothesis and to attacking contrary views. (Th.M.)

CONCLUSION AND EVALUATION

A glance cast back over the preceding survey shows that a noncommittal definition describing aesthetics as the study of beauty and ugliness was about as fair as might be found. When one comes to closer terms with the field, however, it is discovered that there is an issue blocking the desire to be more definite—the issue as to whether, or to what extent, aesthetics may be regarded as a science.

Whether aesthetics is only a science

If the narrow conception of science as restricted to the exact sciences (such as physics) is set aside, there is no question but that aesthetics is a developing science. Indeed, only two or three centuries ago, that is all that could be said for physics. If one merely considers the development of the naturalistic school of aesthetics from Santayana to Thomas Munro and includes the detailed studies of the musicologists and other aesthetic specialists, he finds a rather extensive, systematized body of conceptual knowledge about the arts and their materials and modes of composition, and about the psychological responses and perceptions of them. The issue is no longer whether aesthetics constitutes a special scientific field; it is whether it does not include something else besides. And it clearly does have to include other nonscientific factors.

An inclusive descriptive definition of the field of aesthetics today would, accordingly, be spread out something like the following: (1) On the conceptual side, aesthetics is a special science of which the subject matter is the appreciation of the fine arts and other similar acts of appreciation. For some writers aesthetic appreciation extends very far and is allowed to take in all immediate liking and disliking whatsoever, excluding only instrumental liking of things as means to an end. Not only objects of nature acquire aesthetic value in their view but all sensuous pleasures and displeasures from smell and taste and sex to games and puzzles do so. Other writers draw the line at some natural juncture, as Santayana did with the requirement that pleasure be "objectified." (2) To this body of data conceptually organized in the manner of a special natural science, the subject of aesthetics adds a number of nonscientific modes of cognition. Among these modes are the contextualistic intuition of the vivid quality of an event and the formistic intuition of the essence of things as distinct from their appearance. Also there is the organicist's awareness of an integration of feelings and the expressionist's recognition of authentic (in contrast to corrupt) emotions embodied in works of art. These are all forms of nonscientific cognition. To the writers who stress these cognitions, they are regarded as the heart of aesthetic experience; and the scientific conceptual treatment is peripheral and inferential. (3) To complete the series, one must add sheerly felt emotions devoid of any cognitive ingredient, scientific or nonscientific. These must be interposed to satisfy the exponents of the emotive judgment theories. The scientific aestheticians regard these as data, of course, for their aesthetic systems—as a zoologist could study a cat's fear or a dog's belligerence. In doing this, the emotions are designated and conceptually related to one another through conceptual correlations and laws of contrast and gradation for compositions of painting or music. But the pure emotionist wants to say that value judgments are simply emotional expressions of Ah!, Oh!, Ooo!, Ouch!, or Whoopee!, with aesthetic emotion distinguished from the moral by not being directed upon acts of persons pro and con.

This series of three sorts of items seems to cover the aesthetic field as it is being worked over today. Any further restriction would represent a bias of some sort.

The major contrast, however, is that between the scientific conceptualized treatment of group (1) and the nonscientific, nonconceptual treatments of (2) and (3). There need not be any insurmountable conflict between these two treatments. The data could be regarded as the same set of data in both cases. The scientific aestheticians treat them conceptually as by observers from the outside. The nonscientific aestheticians, on the other hand, treat them intuitively as they feel them immediately from the inside. It just happens that in the science of aesthetics (in contrast to physics) the facts being conceptually described as "out there" (in the picture or in the music) are also capable of being immediately felt and intuited by the observer. A physicist can describe an electron with great accuracy in technical conceptual units, but he cannot feel it from the inside. An aesthetician, however, cannot only describe the composition of a picture in explanatory conceptual terms from the outside with considerable precision, but he can also feel the composition intuitively and immediately from the inside. These two different modes of aesthetic cognition can, moreover, complement each other. The conceptual description can benefit from the intuited feeling and vice versa.

Objective and subjective vantage points

With this insight in mind, a condensed definition can perhaps, without too much distortion, be given of aesthetics somewhat as follows: aesthetics is the study of immediate experience and its objects when valued for its own sake, either described conceptually as a special science observing it from without or reported intuitively as it is felt from within. (S.C.P.)

BIBLIOGRAPHY

Histories of aesthetics: The classic history in English is BERNARD BOSANQUET, *A History of Aesthetic* (1892; 4th ed., 1917), an Hegelian, organistic interpretation. KATHERINE EVERETT GILBERT and HELMUT KUHN, *A History of Esthetics*, rev. ed. (1953), is the most frequently used standard text. An excellent up-to-date treatment is MONROE C. BEARDSLEY, *Aesthetics from Classical Greece to the Present* (1966). WLADYSLAW TATARKIEWICZ, *Historia estetyki*, 2 vol. (1960; Eng. trans., *History of Aesthetics*, 2 vol., 1971), is the most comprehensive scholarly history; each section is followed by salient quotations from the original language with parallel English translations.

Anthologies of contemporary aesthetics: ELISEO VIVAS and MURRAY KRIEGER, (eds.), *The Problems of Aesthetics* (1953); JOHN HOSPERS (comp.), *Introductory Readings in Aesthetics* (1969).

Critical studies: MONROE C. BEARDSLEY, *Aesthetics: Problems in the Philosophy of Criticism* (1958), is a most detailed scholarly critical book on modern men and movements in contemporary aesthetics. A unique treatment of art as a form of emotional symbolism is SUSANNE K. LANGER, *Feeling and Form* (1953), which combines features of symbolic logic and emotional expressionism (stimulating and controversial). NELSON GOODMAN, *Languages of Art* (1968), presents a critical interpretation of such concepts as "imitation" and "expression" in terms of symbolic logic (highly original and controversial). See also MORRIS WEITZ, *Philosophy of the Arts*

(1950), by a strong exponent of linguistic aesthetics and the open texture concept; and JACQUES MARITAIN, *Art et scholastique*, new ed. (1927; Eng. trans., *Art and Scholasticism, with Other Essays*, 1930), a formistic, Thomistic, interpretation.

Schools of thought in aesthetics: Many of the most important works of the various schools are cited in the text; in addition, the following may be listed. (*Hedonistic school*): CURT J. DUCASSE, *The Philosophy of Art* (1929), probably the most rigorous analytical account to be found; DEWITT H. PARKER, *The Principles of Aesthetics* (1920), describes materials of composition in the arts; STEPHEN C. PEPPER, *Principles of Art Appreciation* (1949), a systematic development consistently employing element analysis; DAVID W. PRALL, *Aesthetic Judgment* (1929) and *Aesthetic Analysis* (1936), careful analyses of aesthetic value, with sensitive illustrations of aesthetic discrimination. (*Contextualistic school*): One of the most deliberately consistent expositions is STEPHEN C. PEPPER, *Aesthetic Quality* (1937); also IRWIN EDMAN, *Arts and the Man* (1939). See also LEO STEIN, *A-B-C of Aesthetics* (1927), perhaps the earliest clear statement of contextualism; historically remarkable for its originality of insight and its reflection of *avant garde* influences. (*Organistic school*): One of the best condensed statements is BERNARD BOSANQUET, *Three Lectures on Aesthetic* (1915). Prominent recent exponents of the theory are BERTRAM MORRIS, *The Aesthetic Process* (1943); T.M. GREENE, *The Arts and the Art of Criticism* (1940); HAROLD OSBORNE, *Aesthetics and Criticism* (1955) and *Aesthetics and Art Theory* (1970). (*Formistic school*): HIPPOLYTE ADOLPHE TAINE'S two theoretical books are *Philosophie de l'art* (1865; Eng. trans., 1865) and *De l'Idéal dans l'art* (1869; Eng. trans., 1869); but his principal work is his *Histoire de la littérature anglaise*, 4 vol. (1863–64; Eng. trans., 1873), an intensive study showing how the literary art exhibits the essence of the product of its environment. From a non-Scholastic background, one of the most systematic developments is JAMES K. FEIBLEMAN, *Aesthetics* (1949) and *In Praise of Comedy* (1939).

Oriental aesthetics: WING-TSIT CHAN (comp. and trans.), *A Source Book in Chinese Philosophy* (1963; paperback, 1969), a monumental history, based on original sources, tracing the entire panorama of Chinese philosophy; A.K. COOMARASWAMY, *The Dance of Śiva*, rev. ed. (1957), authoritative essays treating Hindu and Buddhist painting and music with insight, and interpreting the religious symbolism and profound mysticism of Indian art; SHERMAN E. LEE, *A History of Far Eastern Art* (1964), a thorough, clearly-stated treatise, lavishly illustrated, explaining the social, religious, philosophical, and visual aspects; RADHAKAMAL MUKERJEE, *The Culture and Art of India* (1959), a comprehensive work that stresses the great variety of styles and focusses upon the spread of Buddhism to China, the Golden Age of Gupta culture, and the Tantric renaissance; THOMAS MUNRO, *Oriental Aesthetics* (1965), a far-reaching critical study with comparison of Eastern and Western attitudes and beliefs (with many bibliographic notes); MILTON B. SINGER (ed.), *Krishna: Myths, Rites, and Attitudes* (1966), a treatment of legends that form a thread running through Indian civilization and provide a key to the understanding of it at its best; MAKOTO UEDA, *Literary and Art Theories in Japan* (1966), essays by specialists in each of the arts bringing out its essential forms, ideas, aims, and values; VATSYAYANA, *The Kāma-sūtra*, ed. by W.G. ARCHER (1963), a distinguished guide to sexual enjoyments and the sensuous delights of perfume, music, and lyric poetry.

(S.C.P./Th.M.)

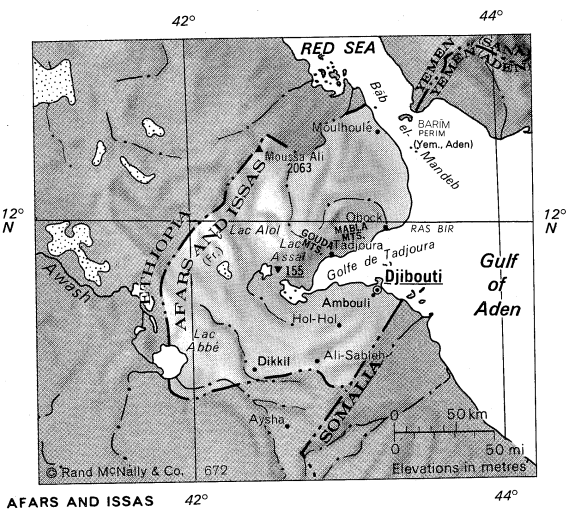
Afars and Issas

The French Territory of the Afars and Issas—officially the Territoire Français des Afars et des Issas—is an overseas territory of France, located on the east coast of Africa, facing the Strait of Bab el-Mandeb between the Red Sea and the Gulf of Aden. It is bordered by Ethiopia to the north, west, and southwest, and by Somalia to the south. It covers an area of 8,880 square miles (23,000 square kilometres), and its coastline, deeply indented by the Gulf of Tadjoura, runs some 500 miles from north to south. The population in the early 1970s was over 125,000. The capital is Djibouti; other towns are Ali-Sabieh and Dikkil (Dikhil) in the south, and Tadjoura and Obock in the northeast.

Strategic importance

The Territory of the Afars and Issas, formerly known as French Somaliland, is a desert land of nomads and traders. Its natural resources are poor, and its landscape and climate forbidding. Its location at the southern end of the Red Sea, however, lends it strategic value. Djibouti is the only seaport to which neighbouring Ethiopia is connected by rail. The nomadic tribes cross into both Ethiopia and Somalia.

The landscape. The territory consists mostly of slightly elevated, volcanic plateaus that are bordered in places by sunken plains and lakes, some of which (Lakes Assal and Alol) are below sea level. Mountain ranges north of the Gulf of Tadjoura reach 6,600 feet at Moussa Ali. Drainage runs either to the coast or the lakes, but few streams flow above ground except after rain.



The climate is extremely hot during most of the year; temperatures average 86° F (30° C) at Djibouti. A relatively cool season extends from November to April. Humidity is high near the coast but decreases notably in the interior. The scarce and irregular precipitation averages less than 15 inches annually along the coast and about 20 inches in the mountains. Except for the few permanent wooded areas on the Mabla and Gouda mountains, the vegetation consists chiefly of scrub bush. The animal life is similar to that of Ethiopia and Somalia and includes hyenas, jackals, antelopes, and gazelles.

The people. The population in 1967 numbered 125,000, of whom 40,000 were nomads. The two major groups are the Afars, or Danakil, and the Issas. Both share a common Hamitic origin and language, both are largely nomadic, and both adhere to Islām. The Afars are concentrated in the north and west and the Somali Issas in the south.

The Afars, who numbered about 48,000 in 1967, are part of a larger group, the majority of which lives in Ethiopia. Those in the territory are divided into two main groups—the Adoyammaras and the Asahyammaras—which in turn are subdivided into several tribes and subtribes. All belong to three feudal-type sultanates, of which only that of Tadjoura is wholly within the territory. The sultan, who is traditionally chosen alternately from one of two families, shares his power with a vizier and a tribal council.

The Issas numbered about 58,000 in 1967. They are divided into three main groups—the Abgals, the Dalols, and the Wardiqs; they, too, are divided into several tribes. Their society is not as rigidly hierarchical as that of the Afars, although it has pariah castes. The supreme chief, the *ougaz*, is elected for life from the Wardiqs and lives in Ethiopia.

There are also over 8,000 Arabs and more than 10,000 Europeans and *assimilés* (persons not of European or African origin), most of whom live in Djibouti. Created by the French in 1892, Djibouti accounts for about half of the territory's population. In addition to the larger Issa, Afar, and Arab communities, there is also an unstable population of Indian traders, Sudanese fishermen, and Somali and Ethiopian labourers. The Europeans are chiefly French, Greek, and Italian.

The economy. Agriculture is limited to the small-scale production of dates, other fruits, and vegetables. Most

Afars and Issas, Area and Population

	area		population	
	sq mi	sq km	1960 census	1967 estimate
Cercles				
Ali-Sabieh	...	...	...	11,000
Dikkil	...	...	...	22,000
Djibouti	...	...	...	62,000
Obock	...	...	...	9,000
Tadjoura	...	...	...	21,000
Total Afars and Issas	8,880	23,000	81,000	125,000

Source: Official government figures.

cultivation is at Ambouli and Dikkil; the produce is generally sold at Djibouti.

There are some 22,000 camels, 632,000 sheep and goats, and 16,000 cattle. In the late 1960s the administration was conducting a program of well digging and introducing new fodder crops. Limited fishing is practiced on an individual basis along the coast.

In addition to salt, iron, copper, and other ores exist in small quantities. None of the ores were exploited by the early 1970s, and the salt trade was discontinued in the mid-1950s.

Internal trade is negligible. The economy rests almost entirely on services and external trade. Imports are mainly textiles and other manufactured consumer goods, coal, and cement. Exports include local products such as hides and skins, Ethiopia's transit shipments of coffee and livestock, goods unloaded for transshipment, and shipping supplies.

Djibouti was severely affected by the closing of the Suez Canal in 1967. Traffic declined by about 75 percent, and the situation remained serious in the early 1970s.

Port facilities

Transport and communications. Djibouti is equipped with some 6,400 feet of quays, nine piers, a refrigeration plant, water reservoirs, and a fuel-storage tank. The port is connected to Addis Ababa, Ethiopia, by rail.

There are about 50 miles of paved roads and about 560 miles of tracks. The Djibouti airport serves international air routes to France, Ethiopia, East Africa, Madagascar, and Egypt. Several small airfields are mostly used by military aircraft. There are daily radio broadcasts in French, Afar, Issa, and Arabic.

Government. The territory is an overseas territory within the French Republic. It is financially autonomous and is represented in the French Parliament. The organs of administration are the Government Council and the Chamber of Deputies. There are five local administrative districts, or *cercles*.

The Government Council consists of a president and from six to eight territorial ministers, who are chosen among and elected by the members of the Chamber of Deputies. It manages the affairs of the territory and is in charge of public services. It is responsible to the Chamber of Deputies, which consists of 32 members who are elected by direct universal suffrage for five years, and which is concerned with public finance, economics, and social problems. It is subject to dissolution by the French government, in which event elections must take place within two months.

The French government is represented by a high commissioner who is appointed by decree of the French Council of Ministers.

The labour force

Social conditions. The labour force contains numerous individuals who are only temporarily employed. Most of the wage-earning population is employed at the port and railway in Djibouti. Other labour markets are the administration, the construction and public-works industries, and the trading companies.

The main hospital at Djibouti is equipped with some 600 beds. In the interior of the territory, there are several multipurpose dispensaries and medical posts.

Education, which is free but not compulsory, is provided by both religious and secular institutions. Public primary education is financed by the government and mission schools (most of which are Catholic) receive financial aid. There are about 7,600 primary school students and more than 1,100 lycée (state-run secondary school) students.

Cultural life and institutions. The territory's cultural life is centred in Djibouti and is limited to a few public libraries and the showing of educational films by the government. The Djibouti-based Société d'Études de l'Afrique Orientale (East African Studies Association) publishes a quarterly magazine, *Pount*, dealing with a variety of subjects pertaining to the territory. A weekly newspaper, *Le Réveil de Djibouti*, is issued by the Ministry of Information and Tourism.

BIBLIOGRAPHY. V.M.THOMPSON and R. ADLOFF, *Djibouti and the Horn of Africa* (1968), a well-documented study of the territory; A. TEUILIERES, *L'Outre-Mer français: hier, aujourd'hui, demain* (1970), a general work on the past, present, and future economic and human problems of French overseas territories.

(M.B.P.)

Afghanistan

Afghanistan is a country in south central Asia. It was a constitutional monarchy until proclaimed a republic in a coup d'etat of July 1973.

It has an area of 251,823 square miles (652,221 square kilometres) and an estimated population (1971) of about 17,500,000. A mountainous country, it is bounded to the north by the Soviet Union (the Turkmen, Uzbek, and Tadzhik Soviet Socialist republics), to the west by Iran, to the east and south by Pakistan (Baluchistan Province, Northwest Frontier Province, and Pakistani-occupied Jammu and Kashmir), and to the extreme northeast by the Sinkiang Autonomous Region of the People's Republic of China, with which Afghanistan shares a 50-mile (80-kilometre) border. The capital of the republic remains the city of Kābul.

The country's strategic situation

The present boundaries of Afghanistan were established in the 19th century by agreement between imperial Britain and imperial Russia, which sought to lessen the dangers of their power rivalry in the region by maintaining Afghanistan as a buffer state. This circumstance resulted in the creation of the Wākhān Corridor—a strip of Afghan territory projecting northeastward along the Wākhān Ridge of the Hindu Kush ranges for almost 200 miles; in places it is only seven miles wide.

Afghanistan is a landlocked country; the nearest sea—the Arabian Sea—lies 300 miles to the south. This relative isolation, together with the location of the country on the only practical western approach overland to the Indian subcontinent, has always exerted a decisive influence upon the country's history. During periods when communication has been overland between Europe and the Middle East, on the one hand, and India and the Far East, on the other, Afghanistan has been a crossroads of trade and cultural exchange. Whenever communications have been by sea—for example, after the Portuguese opened a sea route to India at the end of the 15th century—Afghanistan has, on the contrary, been isolated and left outside the mainstream of civilization. Its geographic location has also meant that practically all the great military conquerors who marched to India passed through Afghanistan—a fact that also explains the great diversity in the country's population.

Lacking most of the social and economic institutions necessary for the development of its economy and seriously handicapped by its geographic location, Afghanistan is now attempting to modernize through economic planning and political reforms. While it accepts assistance from both the Western and the Communist nations, the country has avoided becoming entangled in the power-bloc rivalries that predominate in the region. (For associated physical features, see HINDU KUSH [MOUNTAINS] and PAMIR MOUNTAIN AREA; for historical aspects, see AFGHANISTAN, HISTORY OF.)

THE LANDSCAPE

The natural environment. *Relief.* Afghanistan's shape has been compared to a leaf, of which the Wākhān (Vākhān) Corridor forms the stem. The Hindu Kush moun-

tain system forms a formidable and almost impenetrable barrier, reaching heights of 13,000 to 21,000 feet. It is crossed by the Wākhān Corridor and follows a generally southwesterly direction until it reaches a point about 100 miles north of Kābul. Here it opens like a fan, with the main range continuing due west toward the Iranian border—under the names of Kūh-e Bābā, Safīd Kūh, and Band-e Bāyan—and sending out spurs in different directions. One of these, a subsidiary range to the northwest, is the Band-e Torkestān; other spurs thrust southwestward in diminishing order of size until the final spur merges with the southern highlands. Along the eastern frontier is another series of important spurs that block the country from the rain-laden winds of the Indian Ocean, thus partly accounting for the dryness of the climate.

Geographic regions

The Hindu Kush and its subsidiary ranges divide the country into three distinct geographic regions. These are firstly, to the north, a region of plains and fertile foothills, sloping gently toward the Amu Darya (ancient Oxus River); secondly, to the south, a region of high plateaus and sandy deserts; and thirdly, between these two regions, a region of lofty mountains and deep and narrow valleys. Afghanistan is a country of high altitudes. Only a small part of it—the region along the Amu Darya in the north and the delta of the Helmand River in the southwest—is at an altitude of less than 2,000 feet. The greater part of the country lies at heights of between 2,000 and 10,000 feet. The remaining part consists of lofty mountains. Extensive sandy deserts—including the Rīgestān desert in the southeast—cover the south up to the Helmand River and its tributaries.

The present landforms of Afghanistan are of relatively recent origin, having developed in the late Tertiary Period (from 26,000,000 to 2,500,000 years ago), during which a generally slow but occasionally lively uplifting movement occurred, in the course of which the structural unit of the Hindu Kush evolved.

Drainage. Practically all the important rivers have their sources in the central mountainous regions and run toward the periphery of the country. Only the waters of the rivers flowing eastward reach the sea, through the Indus; all others empty their waters into inland lakes or dry up in sandy deserts. The major drainage systems are those of the Amu Darya, Helmand, Kābul, and Harīrūd rivers.

The Amu Darya, which forms, for a distance of 600 miles, the frontier between Afghanistan and the Soviet Union, starts from the glaciers of the Pamirs and drains an area of about 93,000 square miles in the northeastern (Badakhshān) and northern (Kataghan and Mazār-e Sharīf) parts of the country. Its main tributaries on the Afghan side are the Kowcheh and Qondūz. The waters of other streams from the northern slopes of the Hindu Kush, such as the Samangān, Balkh, and Almār, are completely used up for agriculture before reaching the Amu Darya. The Amu Darya is navigable from the point of its confluence with the Kowkcheh.

The Helmand, about 900 miles long, rises in the mountains of Paghmān, west of Kābul. With its main tributary, the Arghandāb, it drains all southern Afghanistan (100,000 square miles) and empties into the Hāmūn-e Ṣāberī, an inland lake lying across the Afghan-Iranian border.

The Kābul River and its tributaries—the Panjshēr, Alīshang, and Konar—drain eastern Afghanistan (32,000 square miles); the Kābul flows into the Indus River at Attock.

The Harīrūd, which derives its name from the city of Herāt, runs westward; after irrigating the valley of Herāt, it dries up in the sandy desert of the Turkmen Soviet Socialist Republic. Other important rivers are the Daryā-ye Morghāb (Murgab) in the northwest, the Farāh Rūd in the southwest, and the Gowmal in the east.

Afghanistan has few lakes of any considerable size. The two most important are the Hāmūn-e Ṣāberī, already mentioned, and the Āb-e Īstādeh-ye Moqor, which is situated 60 miles south of Ghaznī. In the central highland of Hazārajāt there are five small-sized lakes, called the Band-e Amīr; they are famed for their rare shades of colour, which, caused by the underlying bedrock, vary from milky white to dark green. These lakes constitute one of the most important tourist attractions of the country.

Climate. Although Afghanistan is situated in the so-called subtropical climatic belt, the climate is conditioned more by the altitude of its different regions than by their latitude. Lying at a minimum distance of 300 miles from the sea and cut off from the Indian monsoon (rain-bearing) winds by its eastern ranges, Afghanistan has a typical continental dry climate, characterized by seasonal and diurnal extremes of temperature and by clearly defined seasons. Cold air masses that enter the country from the north, and the Atlantic low depression, which enters from the northwest, influence the weather in winter and early spring, bringing snowfall and severe cold in the highlands and rain in the lower regions. The effects of the monsoon winds manifest themselves in summer in the form of heavy rainfall and thunderstorms, which occur in the easternmost districts of Nangarhār and Paktīā provinces and in the adjacent high mountains. Also in summer, humid air masses from the Persian Gulf enter the country from the southwest, causing sporadic showers and thunderstorms.

Temperature ranges

Afghanistan is characterized by a wide range of temperature, with extremely cold winters and hot summers. In Jalālābād, one of the hottest localities in the country, the highest temperature recorded in the month of July has been 120° F (49° C) and the lowest about 27° F (−3° C). In Kābul, which lies at an altitude of 6,000 feet, the highest recorded temperature has been 97° F (36° C), while the lowest was −24° F (−31° C); the average in July is about 76° F (24° C) and in January about 26° F (−3° C).

There is much regional diversity and uneven distribution of precipitation resulting from the differences in altitude and terrain. Annual precipitation varies from 53 inches, reported in the Sālang Pass in the Hindu Kush, to some three inches in the arid region of Farāh in the west. In Kābul, the average precipitation is about 14 inches. In the highlands, snowfalls are frequent from December to March, but the main precipitation occurs in early spring in the form of rain. In the lowlands it rains intermittently from December to April or May. The summer months are hot, dry, and cloudless all over the country, except in the monsoon region, which receives heavy and frequent showers in July and August. In spite of these variations, the climate of the country as a whole is considered to be invigorating and healthy. Spring and autumn are particularly fine, with pleasant days and nights. Even in the hottest season the temperature falls after sunset, and is generally followed by a cool night. At lower altitudes—for example, at Jalālābād and Farāh—winters are mild and fine, but the heat in summer is almost unbearable.

Vegetation and animal life. The southern part of Afghanistan, especially toward the west, where sand deserts predominate, is a region of scanty vegetation. Trees are almost nonexistent, and only in the rainy season of early spring is the soil covered with flowering grasses and herbs. Northward, the plants become more numerous; at higher altitudes they are almost luxuriant. Here they include shrubs and plants such as the camel thorn; plants such as locoweed and poison vetch, which may be poisonous to livestock; spiny restharrow, a woody herb whose fibrous root may be used as a toothbrush; mimosa plants; the common wormwood; and several species of salsola, or Russian thistle.

Still higher, the wild walnut and oak (evergreen, holly-leaved, and kermes) appear, mixed with alder, ash, khinjuk (a member of the pistachio family), juniper, and dwarf laburnum. Between 6,000 and 12,000 feet, especially in the eastern and northeastern part, there occur large forest trees, among which conifers are predominant. Yews, hazels, wild peaches, almonds, and vines also are found. Growing under the shade of these are several varieties of roses, honeysuckle, hawthorn, currants, and gooseberries.

Most of the wild animals of the subtropical temperate zone are to be found. They include wolves, foxes, hyenas, jackals, gazelles, wild dogs, wild cats, leopards, mongooses, moles, shrews, hedgehogs, bats, and several species of kangaroo rat (jerboas). Bears are found in the forests and the mountains; the Siberian tiger, however, which used to inhabit the banks of the Amu Darya, has disappeared.

Game animals

Among the game animals sought by trophy hunters are the wild goat, including the markhor (prized for its long, twisted horns) and the ibex (with long, backward-curved horns), both found in the central ranges, as is the urial, a wild sheep with tightly curved horns; and argali and Marco Polo sheep, two species of wild sheep with widely sweeping horns, found in the high altitudes of the Pamirs.

There are many varieties of birds, including several of pheasant, duck, teal, crane, pelican, snipe, partridge, and crow. Vultures, eagles, and other birds of prey are also to be found. Smaller birds include the sparrow, quail, bulbul, magpie, and swallow.

Many varieties of freshwater fish live in the rivers and streams. Brown trout are found in the watercourses running toward the Amu Darya.

Traditional regions. The Hindu Kush, a tremendous barrier to communication, divides Afghanistan into two principal regions: northern Afghanistan, sometimes called Afghan Turkistan, and Afghanistan south of the Hindu Kush. Northern Afghanistan, in its turn, is traditionally subdivided into the Badakhshān-Wākhān region in the east and the Balkh-Meymanah in the west. The first, which is mainly a region of mountains and high plateaus, is inhabited principally by peoples of Tadzhik

MAP INDEX

Political subdivisions

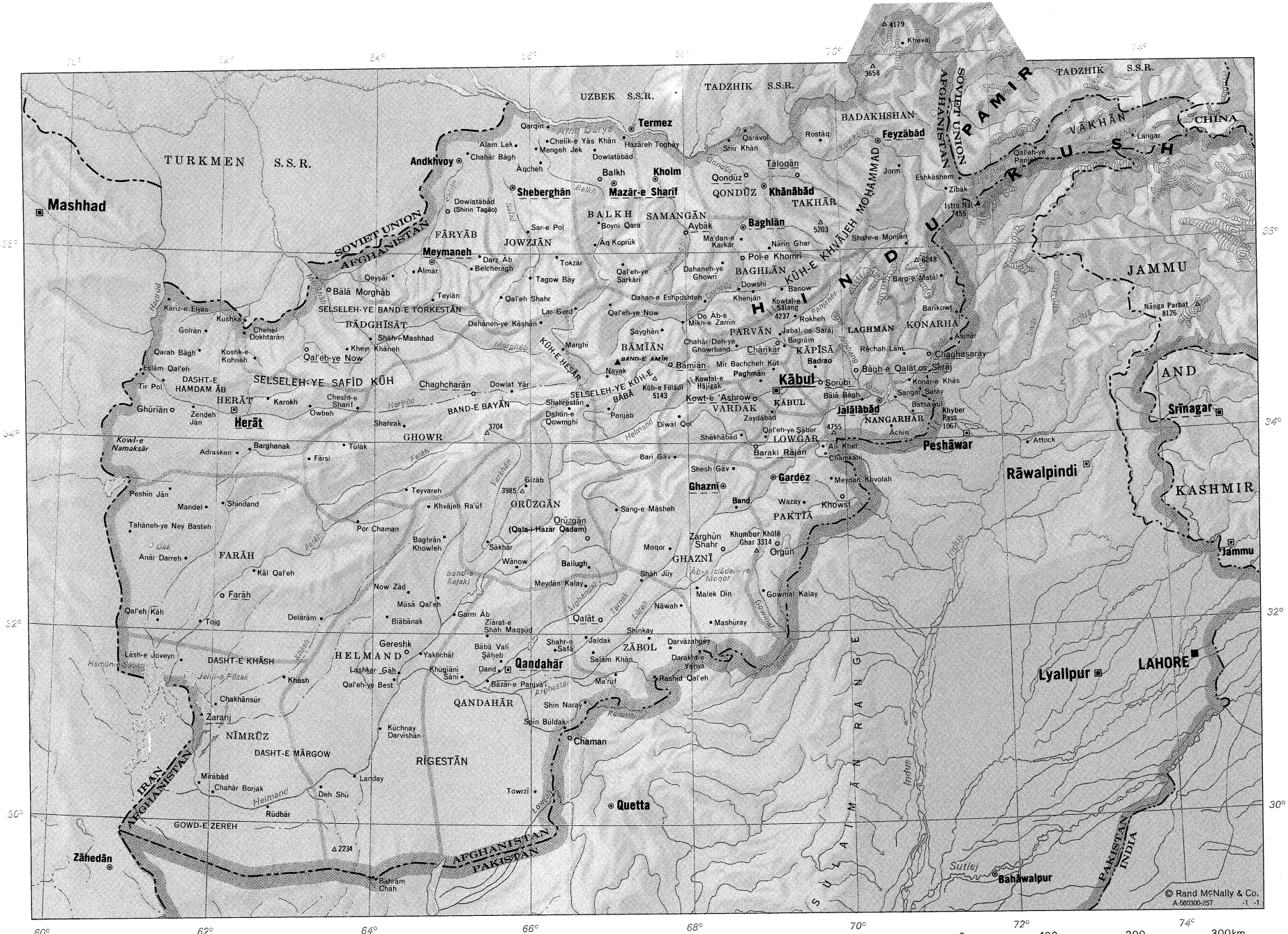
Badakhshān.....36·45n 72·00e
Bādghīsāt.....35·00n 63·45e
Baghlān.....35·45n 69·00e
Balkh.....36·30n 67·00e
Bāmiān.....34·45n 67·15e
Farāh.....33·00n 62·30e
Fāryāb.....36·00n 65·00e
Ghaznī.....33·00n 68·00e
Ghowr.....34·00n 65·00e
Helmand.....31·00n 64·00e
Herāt.....34·30n 62·00e
Jowzjān.....36·30n 66·00e
Kābul.....34·30n 69·00e
Kāpīsā.....34·45n 69·30e
Konarhā.....35·15n 71·00e
Laghmān.....35·00n 70·15e
Lowgar.....33·50n 69·00e
Nangarhār.....34·15n 70·30e
Nīmrūz.....30·30n 62·00e
Orūzgān.....33·15n 66·00e
Paktīā.....33·30n 69·30e
Parvān.....35·15n 69·30e
Qandahār.....31·00n 65·45e
Qondūz.....36·45n 68·30e
Samangān.....36·15n 67·40e
Takhār.....36·30n 69·30e
Vardak.....34·15n 68·00e
Zābol.....32·00n 67·15e

Cities and towns

Āchīn.....34·08n 70·42e
Adraskan.....33·39n 62·16e
'Alam Lek.....37·02n 65·57e
'Al ī Khēl.....33·57n 69·43e
Almār.....35·50n 64·32e
Anār Darreh.....32·46n 61·39e
Andkhvoy.....36·56n 65·08e
Āqcheh.....36·56n 66·11e
Āq Koprūk.....36·05n 66·51e
Āsmār.....35·02n 71·22e
Āybak.....36·16n 68·01e
Bābā Valī Ṣāheb.31·40n 65·40e
Badrao.....34·46n 69·40e
Bāgh-e Qalāt os Serāj.....34·40n 70·13e
Baghlān.....36·13n 68·46e
Baghrān Khowleh.....33·01n 64·58e
Bagrām.....34·58n 69·17e
Bahrām Chāh....29·26n 64·03e
Bailugh.....32·41n 66·46e
Bālā Bāgh.....34·24n 70·14e
Bālā Morghāb....35·35n 63·20e
Balcherāgh.....35·50n 65·14e
Balkh.....36·44n 66·47e
Bāmīān.....34·50n 67·50e
Band.....33·17n 68·39e
Banow.....35·38n 69·15e
Barakī Rājān.....33·56n 68·55e
Barg-e Maṭāl....35·40n 71·21e
Barghanak.....33·56n 62·26e
Barī Gāv.....33·52n 67·49e
Barikowṭ.....35·18n 71·32e
Bātsawul.....34·15n 70·52e
Bāzār-e Panjvā'ī..31·32n 65·28e
Bībānak.....32·11n 64·11e
Boyni Qara.....36·19n 66·53e
Chaghasarāy.....34·52n 71·09e
Chaghcharān....34·32n 65·15e
Chahār Bāgh....37·00n 65·14e
Chahār Borjak...30·17n 62·03e
Chahār Deh-ye Ghowrband.....34·59n 68·44e
Chakhānsūr.....31·10n 62·04e
Chamkanī.....33·48n 69·49e
Chārīkār.....35·01n 69·11e
Chehel Dokhtarān.....35·06n 62·19e
Chelīk-e Yās Khān.....37·05n 66·14e
Chesht-e Sharīf..34·21n 63·44e
Dahan-e Eshposhteh.....35·18n 68·06e
Dahaneh-ye Ghowrī.....35·54n 68·30e
Dahaneh-ye Kāshān.....35·09n 66·14e
Dahan-e Qowmghī.....34·28n 66·31e
Dand.....31·37n 65·42e
Darakht-e Yahyá.31·50n 68·08e
Darvāzahgêy.....31·48n 67·44e
Darz Āb.....35·58n 65·22e
Deh Shū.....30·26n 63·19e
Delārām.....32·11n 63·25e
Dīwāl Qol.....34·19n 67·54e
Do Āb-e Mīkh-e Zarrīn.....35·16n 68·00e
Dowlatābād (Shīrīn Tagāo)..36·26n 64·55e
Dowlatābād.....36·29n 66·50e
Dowlat Yār.....34·33n 65·47e
Dowshī.....35·37n 68·41e
Eshkāshem.....36·42n 71·34e
Eshposhteh, see Dahan-e Eshposhteh
Eslām Qal'eh....34·40n 61·04e
Farāh.....32·22n 62·07e
Fārsī.....33·47n 63·15e
Feyẓābād.....37·06n 70·34e
Gardēz.....33·37n 69·07e
Garm Āb.....32·14n 65·01e
Gereshk.....31·48n 64·34e
Ghaznī.....33·33n 68·26e
Ghūrīān.....34·21n 61·30e
Gīzāb.....33·23n 66·16e
Golrān.....35·06n 61·41e
Gowmal Kalay...32·29n 68·55e
Hazāreh Toghāy.37·12n 67·13e
Herāt.....34·20n 62·12e
Jabal Os Sarāj...35·07n 69·14e
Jalālābād.....34·26n 70·28e
Jaldak.....31·58n 66·43e
Jorm.....36·52n 70·51e
Kābul.....34·30n 69·11e
Kāl Qal'eh.....32·38n 62·32e
Kārīz-e Elyās.....35·25n 61·20e
Karkar, see Ma'dan-e Karkar
Karokh.....34·28n 62·35e
Khānābād.....36·41n 69·07e
Khāsh.....31·31n 62·52e
Khenjān.....35·36n 68·54e
Khevāj.....38·13n 70·59e
Kheyr Khāneh...34·57n 63·37e
Kholm.....36·42n 67·41e
Khowst.....33·22n 69·57e
Khūgīanī Ṣānī...31·31n 66·12e
Khvājeh Ra'ūf...33·19n 64·43e
Konar-e Khāṣ....34·39n 70·54e
Koshk-e Kohneh.34·52n 62·31e
Kowt-e 'Ashrow..34·27n 68·48e
Kūchnay Darvīshān.....30·59n 64·11e
Landay.....30·31n 63·47e
Langar.....37·02n 73·47e
Lar Gerd.....35·29n 66·40e
Lāsh-e Joveyn...31·43n 61·37e
Lashkar Gāh.....31·35n 64·21e
Ma'dan-e Karkar.36·02n 68·46e
Malek Dīn.....32·25n 68·04e
Mandel.....33·17n 61·52e
Marghī.....34·58n 66·31e
Ma'rūf.....31·34n 67·03e
Mashūray.....32·12n 68·21e
Mazār-e Sharīf...36·42n 67·06e
Mengeh Jek.....37·02n 66·07e
Meydān Kalay...32·25n 66·44e
Meydān Khvolah..... 33·36n 69·51e
Meymaneh.....35·55n 64·47e
Mīrābād.....30·25n 61·50e
Mīr Bachcheh Kūt.....34·45n 69·08e
Moqor.....32·52n 67·47e
Mūsá Qal'eh.....32·22n 64·46e
Nārīn Ghar.....36·01n 69·06e
Nāwah.....32·19n 67·53e
Nayak.....34·44n 66·57e
Now Zād.....32·24n 64·28e
Orgūn.....32·51n 69·07e
Orūzgān (Qala-i-Hazār Qadam)..32·56n 66·38e
Owbeh.....34·22n 63·10e
Paghmān.....34·36n 68·57e
Panjāb.....34·22n 67·01e
Peshīn Jān.....33·25n 61·28e
Pol-e Khomrī....35·56n 68·43e
Por Chaman.....33·08n 63·51e
Qalāt.....32·07n 66·54e
Qal'eh Shahr....35·33n 65·34e
Qal'eh-ye Best...31·30n 64·21e
Qal'eh-ye Now...34·59n 63·08e
Qal'eh-ye Now...35·27n 67·08e
Qal'eh-ye Panjeh.37·00n 72·36e
Qal'eh-ye Ṣāber..34·02n 69·01e
Qal'eh-ye Sarkārī.....35·54n 67·17e
Qandahār.....31·35n 65·45e
Qarah Bāgh.....34·56n 61·46e
Qarāvol.....37·14n 68·46e
Qarqīn.....37·25n 66·03e
Qeyṣār.....35·41n 64·17e
Qondūz.....36·45n 68·51e
Rashīd Qal'eh...31·31n 67·31e
Rēchah Lām.....34·58n 70·51e
Rokheh.....35·16n 69·28e
Rostāq.....37·07n 69·49e
Rūdbār.....30·09n 62·36e
Sākhar.....32·57n 65·32e
Salām Khān.....31·47n 66·45e
Sangar Sarāy....34·24n 70·38e
Sang-e Māsheh..33·08n 67·27e
Sar-e Pol.....36·14n 65·55e
Sayghān.....35·11n 67·42e
Shah-i-Mashhad.35·03n 63·58e
Shahrak.....34·06n 64·18e
Shahr-e Monjān..36·02n 70·46e
Shahr-e Ṣafā....31·50n 66·22e
Shahrestān.....34·22n 66·47e
Sharīf, see Mazār-e Sharīf
Sheberghān.....36·41n 65·45e
Shēkhābād.....34·05n 68·45e
Shesh Gāv.....33·45n 68·33e
Shīndand.....33·18n 62·08e
Shīnkay.....31·57n 67·26e
Shīn Naray.....31·19n 66·43e
Shīr Khān.....37·11n 68·36e
Sorūbī.....34·36n 69·43e
Spīn Būldak.....31·01n 66·24e
Tagow Bāy.....35·42n 66·03e
Tahāneh-ye Ney Basteh.....32·59n 60·53e
Tāloqān.....36·44n 69·33e
Teylān.....35·32n 64·47e
Teyvareh.....33·21n 64·25e
Tīr Pol.....34·36n 61·15e
Tojg.....32·04n 61·48e
Tokzār.....35·52n 66·26e
Towrzī.....30·11n 65·59e
Tūlak.....33·58n 63·44e
Wānow.....32·38n 65·54e
Wazay.....33·22n 69·26e
Yakhchāl.....31·47n 64·41e
Zaranj.....31·06n 61·53e
Zarghūn Shahr..32·51n 68·25e
Zaydābād.....34·17n 69·07e
Zendeh Jān.....34·21n 61·45e
Zīārat-e Shāh Maqṣūd.....31·59n 65·30e
Zībāk.....36·34n 71·22e

Physical features

Āb-e Īstādeh-ye Moqor, *salt lake*.32·30n 67·55e
Āb-e Panj, *river*..37·06n 68·20e
Alīshang, *river*...34·38n 70·14e
Amīr, Band-e, *lakes*.....34·49n 67·12e
Amu Darya, *river*.....37·27n 67·00e
Arghandāb, *river*.....31·27n 64·23e
Arghastān, *river* .31·23n 65·45e
Bābā, Selseleh-ye Kūh-e, *mountains*.....34·30n 67·30e
Balkh, *river*.....36·39n 66·56e
Band-e Bāyan, *mountains*.....34·20n 65·30e
Dak, *lake*.....32·48n 61·14e
Farāh, *river*.....31·29n 61·24e
Fūlādī, Kūh-e, *mountain*.....34·38n 67·32e
Gowd-e Zereh, *salt lake*.....29·50n 62·00e
Gowmal, *river*....31·56n 69·10e
Hājīgak, Kowtal-e, *pass*..34·39n 68·05e
Hamdam Āb, Dasht-e, *plain*...34·45n 61·30e
Harīrūd, *river*....35·35n 61·12e
Helmand, *river*...30·55n 61·45e
Ḥeṣār, Kūh-e, *mountains*.....34·50n 66·30e
Hindu Kush, *mountain range*.36·00n 71·30e
Istro Nāl, *mountain*.....36·23n 71·54e
Kābul, *river*.....33·55n 72·14e
Kajakī, Band-e, *reservoir*.....32·22n 65·16e
Khash, *river*.....31·15n 62·05e
Khāsh, Dasht-e, *desert*.....31·40n 62·25e
Khumbur Khūlē Ghar, *mountain*.....32·49n 68·47e
Khvājeh Moḥammad, Kūh-e, *mountains*.....36·22n 70·17e
Khyber Pass.....34·05n 71·10e
Konar, *river*.....34·25n 70·32e
Kowkcheh, *river*..37·10n 69·25e
Lowṛah, *river*.....29·40n 65·45e
Lūṛah, *river*.....31·32n 66·32e
Mārgow, Dasht-e, *plain*.....30·45n 63·10e
Morghāb, *river*...35·55n 63·05e
Namaksār, Kowl-e, *salt lake*.....34·00n 60·30e
Panjshēr, *river*...34·38n 69·42e
Pūzak, Jehīl-e, *lake*.....31·30n 61·45e
Qondūz, *river*....37·02n 68·15e
Rīgestān, *physical region*..31·00n 65·00e
Ṣāberī, Hāmūn-e, *lake*.....31·30n 61·20e
Safīd, *river*.....36·44n 65·38e
Safīd Kūh, Selseleh-ye, *mountains*.....34·30n 63·30e
Sālang, Kowtal-e, *pass*.....35·21n 69·05e
Samangān, *river*.36·42n 67·42e
Shīrīn, *river*.....36·49n 65·01e
Tarnak, *river*.....31·25n 65·30e
Torkestān, Selseleh-ye Band-e, *mountains*.....35·25n 64·15e
Vākhān, *physical region*.....37·00n 73·00e
Vākhān, Āb-e, *river*.....37·02n 72·40e
Varkhān, *river*....32·54n 65·30e
Wākhān, see Vākhān


AFGHANISTAN
Size of symbol indicates relative size of town • ○ ◉ ▣ ▣ ■
Elevations in metres
0 100 200 300km
0 100 200 mi
© Rand McNally & Co.
A-560300-257

(Tajik) stock, while in the latter, which consists of plains of comparatively low altitude, dwell a mixture of different races in which the dominant element is the Uzbek-Turkmen group, of Turkic origin.

Afghanistan south of the Hindu Kush, which covers by far the largest part of the country, is more diversified; it may, nevertheless, be divided into four main regions —the Kābul, Qandahār, Herāt, and Hazārajāt regions.

The Kābul, Qandahār, Herāt, and Hazārajāt regions

The Kābul region comprises, besides the area drained by the Kābul River, the high plateau of eastern Afghanistan as far south as the Gowmal River. It is inhabited by Pashtuns (Pathans), Tadzhiks, and Nuristanis (formerly known as Kafirs); historically, the region forms the main connecting link between the other regions and the various peoples of Afghanistan.

The region of Qandahār (Kandahār), consisting of the sparsely populated southern part of Afghanistan, is inhabited mainly by the Durrani branch of the Pashtuns and by a small number of Baluchi and Brahui peoples. The city of Qandahār, former capital of the country, which is situated in a fertile oasis near the Arghandāb River, is the economic and cultural centre of the region.

The region of Herāt, or western Afghanistan, is inhabited by a mixture of Tadzhiks, Pashtuns, and Chahar Aimaks. The life of the region revolves around the city of Herāt, an important trade centre, which is equidistant from the borders of the Soviet Union and Iran.

The mountainous region of Hazārajāt occupies the central part of the country and is inhabited by Mongoloid Ḥazāras. Although Hazārajāt is situated in the heart of the country, its high mountains and poor communication facilities make it the most isolated part of Afghanistan.

Human settlement. About 8 percent of the people of Afghanistan live in towns and cities. The rest are either farmers, semi-sedentary farmers, or nomads.

Sedentary farmers live in small villages, scattered mainly near irrigated land in the vicinity of the principal rivers. These villages are built usually in the form of small forts, each containing a number of mud houses, inhabited by closely connected families who form a defensive community.

The countryside and the cities

In the high alpine valleys, where land is scarce, farmers live mostly in isolated hamlets. Each household owns a few head of livestock, which are moved in summer to the highland pastures and are returned in the winter. These are the semi-sedentary farmers, half agriculturalists and half livestock breeders.

The nomads are, for the most part, Pashtun herdsmen, who seasonally move in groups (tribes or clans) from summer to winter pasturages and back again. They live in tents and, while on the move, pack their simple belongings, together with their small children, on the backs of camels, donkeys, and cattle. According to a sample survey, Afghanistan has about 15,000 villages, some of them of not more than a dozen households, others with populations reaching 2,000 people. Every village and every subdivision of a nomadic clan possesses a *malik*, or headman, who is nominated by the elders and serves as a liaison officer between the inhabitants of the village and the administration.

Kābul, the capital and principal city of Afghanistan, is situated south of the Hindu Kush, at the crossroads of the trade routes from the Indian subcontinent and from Central Asia. It is a city of about 500,000 inhabitants, built on two sides of the Kābul River. The old city, situated south of the river, is composed of crooked, narrow streets and overcrowded mud houses, two or three stories high. North of the river are a number of new quarters that have small gardens or compounds attached to each house. The existence of these compounds accounts mainly for the relatively large area of the city (26 square miles) in proportion to its population (about 500,000). Besides being the administrative capital, Kābul is also the main centre of economic and cultural activity in the country. All banking institutions and most of the important trade houses are situated there, and an industrial area is developing east of the city. The same is true of cultural institutions, such as theatres, concert halls, and libraries, which are usually sponsored by the government.

Other important cities are Herāt (about 104,000 inhabitants) in western Afghanistan; Qandahār (134,000), 300 miles southwest of Kābul; and Mazār-e Sharīf (60,000), north of the Hindu Kush. These are all important road junctions, situated near the frontiers of Iran, Pakistan, and the Soviet Union, respectively.

PEOPLE AND POPULATION

The people. *Linguistic groups.* The population of Afghanistan is formed of a mosaic of ethnic and linguistic groups. The two major language families are the Indo-European and the Ural-Altaic. To the Indo-European group belong the two official languages of the country: Pashto (Pushtu) and Dari (a form of Persian). Pashto is spoken by about half the population, whose members are mainly concentrated in the southern and eastern part of the country but are also well represented in western and northern Afghanistan. Dari (or Farsi) is spoken by about a third of the population—mainly the Tadzhik, Ḥazāras, Chahar Aimak, and Kizilbash peoples. It is commonly used in administration and education and may be considered as the lingua franca of Afghanistan. Other languages of this family are Nuristani (Dardic branch), Baluchi (Iranian branch), and a number of languages of Indic and Pamiri branches spoken in isolated valleys in the northeastern part of the country.

The two major language families

The Ural-Altaic family is represented mainly by Uzbek and Turkoman. Uzbek is spoken by about 1,000,000 people in northern Afghanistan and Turkoman by some 300,000 in the same general area. These languages are closely connected with each other, as well as with Kirgiz, which is spoken by a small community in the extreme northeast.

Ethnic groups. The present population of Afghanistan contains a number of elements, which, in the course of history and as a result of large scale migration and conquests, have been superimposed upon each other. Dravidians, Indo-Aryans, Greeks, Scythians, Arabs, Turks, and Mongols have, at different times, inhabited the country and have influenced its culture and ethnography. The main ethnic groups currently found in the country are the Pashtuns, the Ḥazāras, and the Turkic groups.

The Pashtuns, inhabiting principally the southern and eastern part of the country (with strong contingents also in the western and northern parts), are divided into a number of tribes, some sedentary and others nomadic. The two most important tribal confederations are the Durranis and the Ghilzais. The first is based on the city of Qandahār, while the second inhabits the region between Kābul and Qandahār. Tadzhiks, mostly farmers and artisans, live in northeastern Afghanistan (Kābul and Badakshān areas), around Herāt, and in strong pockets in other regions. They are sedentary in the plains and semi-sedentary in the higher valleys and are not divided into clear-cut tribal groups.

The Ḥazāras constitute another distinctive ethnic group. Their traditional location is the central mountainous region of Hazārajāt, but, as a result of scarcity of land and pressure of population, they migrate in great numbers to other parts of the country and can be found in all towns and cities as labourers and small traders. They speak a dialect of Dari in which words of Mongol origin may still be traced.

To the north of the Hindu Kush, Turkic groups are found. They are mostly Uzbeks and Turkmen, but a small number of Kirgiz dwell in the Wākhān Corridor, or panhandle, in the extreme northeast. The Uzbeks are for the great part farmers, while the Turkmen and Kirgiz are mainly semi-nomadic herdsmen. In the west of the country lives a group of people of mixed origin, known under the general term Chahar Aimak.

Religion. About 99 percent of the people are Muslims, most members of the Sunnī sect (Ḥanafī branch). Ḥazāras and Kizilbash Turkmen are Shīʿites; a few Ismāʿīlīs live in the Hindu Kush and Badakshān. There are also a few thousand Hindus and Sikhs; a few hundred Jews live mainly in the cities and are active in trade.

Sunnī Muslims

Demography. No comprehensive census has so far been taken in Afghanistan. On the basis of sample surveys carried out, however, for the purpose of planning, a rate of growth of 2 percent per year for the population has been accepted.

Out of an estimated population of 17,500,000 in 1971, about 2,500,000 are nomads. Some of these have been settled in the last few years in the plains north of the Hindu Kush or in the Helmand Valley Project. A number of nomads (*pāvendehs*) cross the eastern frontier with their herds toward their summer pastures in Pakistan. Others go to Pakistan and India as labourers or moneylenders. The average density of population is 70 per square mile, although there is considerable variation related to the differences in the economic potential of different regions. Generally speaking, the river valleys, with fertile land and irrigated plains, are densely populated, while the unproductive mountains and sandy deserts have only a sparse population.

Afghanistan, Area and Population

	area		population	
	sq mi	sq km	census*	1970 estimate
Provinces (*velāyāt*)				
Badakhshān	17,374	44,998	—	344,000
Bādghisāt	9,844	25,497	—	320,000
Bāghlān	5,383	13,943	—	733,000
Balkh	6,033	15,626	—	354,000
Bāmiān	7,815	20,242	—	346,000
Farāh	23,008	59,590	—	314,000
Fāryāb	8,122	21,036	—	435,000
Ghaznī	12,096	31,329	—	1,205,000
Ghor	13,808	35,762	—	323,000
Helmand	20,536	53,188	—	318,000
Herāt	16,107	41,718	—	686,000
Jowzjān	8,388	21,726	—	413,000
Kābul	1,266	3,280	—	1,293,000
Kāpīsā	1,798	4,658	—	587,000
Konarhā	3,742	9,693	—	327,000
Laghmān	3,218	8,334	—	223,000
Lowgar	1,673	4,334	—	309,000
Nangarhār	2,771	7,176	--	787,000
Nīmrūz	20,979	54,336	—	122,000
Oruzgān	12,563	32,537	—	528,000
Paktīā	6,086	15,762	—	778,000
Parvān	3,456	8,951	—	887,000
Qandahār	11,696	45,833	—	742,000
Qondūz	3,122	8,085	—	452,000
Samangān	6,090	15,772	—	207,000
Takhar	6,745	17,469	—	494,000
Vardak	4,122	10,675	—	416,000
Zābol	7,981	20,671	—	358,000
Total Afghanistan	251,823†	652,221	—	17,125,000‡§

*No census has ever been taken. †Converted areas do not add to total given due to rounding. ‡Total includes 2,825,000 nomads for whom no provincial breakdown was available. §Figures do not add to total given because of rounding.
Source: Official government figures; Kabul Times.

THE NATIONAL ECONOMY

Resources. When Afghanistan began to plan the development of its economy in the middle 1950s, it lacked not only the necessary social organization and institutions for modern economic activities but also the essential managerial and technical skills. On the whole, the country was at a much lower stage of economic development than most of its neighbours. By the early 1970s the same situation prevailed, but not to the same extent.

Mineral resources. Extensive surveys have revealed the existence of a number of minerals of economic importance. Reserves of natural gas (10,600,000,000,000 cubic feet) have been located in Jowzjān (Shibarghān, northern Afghanistan), and high-grade iron ore is found near Ḥājīgak, northwest of Kābul. Many coal-bearing strata have been found in the northern slopes of the Hindu Kush. Some, like the Qal'eh-ye Sarkārī deposits, are of coking quality, but presently only those of lower quality, at Karkar and Eshposhteh, are exploited because of their favourable location relative to Kābul, the main centre of consumption. There are also deposits of copper, lead, and zinc near Qondūz, rock salt in Tālogān, beryllium in Konar, the semiprecious stone lapis lazuli in Badakhshān, and some oil near the gas field in Jowzjār.

Biological resources. Afghanistan is essentially a pastoral country. Only 0.2 percent of the total area is under cultivation. The remaining parts consist of steppes or mountains that are seasonally covered by flowering grasses and herbs and serve as pastureland for nomads and other livestock breeders. Dominance of pastoralism

Forests cover about 7,700 square miles; they are found mainly in the eastern part of the country and on the northern slopes of the Hindu Kush. Those in the east consist mostly of conifers, providing timbers for the building industry as well as some wild nuts, such as pine kernels, for export. Others, especially oaks, are used as fuel. North of the Hindu Kush is found the pistachio tree, the nuts of which are exported. The gum raisin obtained from asafetida, a herb of the carrot or parsley family, is exported to India, and rhubarb from the mountains around Kābul is used in cooking.

Power resources. Afghanistan is potentially rich in hydroelectric resources. The main rivers rise in high mountains and rush down toward the plains and could produce plenty of power if harnessed. But the high cost of construction in remote areas and the negligible demand render such projects unprofitable except near the large cities and industrial centres.

The potential of hydroelectricity has so far been tapped substantially only in the Kābul-Jalālābād region. Coal has been used in small quantities for power production, and the first thermal station, based on natural gas, was due to be completed at Mazār-e Sharīf in 1972.

Sources of national income. The annual gross national product of Afghanistan is about 77,831,000,000 afghanis (89.15 afghanis = $1 U.S.; 213.96 afghanis = £1 sterling, on December 1, 1970). Per capita national income in the year 1968 to 1969 was estimated at 7,132 afghanis.

Agriculture and animal husbandry. Agriculture and animal husbandry are by far the most important items of the national product, accounting for about 45 percent of its total value. The total arable area in the country has been calculated to be 19,400,000 acres, of which 340,000 are under cultivation. The area under irrigation is being enlarged by a number of new projects implemented with foreign assistance. The most important of these are: the Helmand Valley Project (United States), the Nangarhār Canal (Soviet Union), and the Tarwan Canal (the People's Republic of China).

The chief cultivated crop is wheat, the staple food of the population. Other crops are barley, rice, corn (maize), millet, pulses, clover, lucernes, vegetables, sesame, cotton, tobacco, sugarcane, sugar beets, and a rich variety of fruit and nuts. Annual production of wheat is estimated at about 2,300,000 tons, that of fruit, which is exported in sizable quantity to various Asian and European countries, at about 890,000 tons. Chief crops and animal products

Next to farming, animal husbandry provides the main component of national income. It produces meat and dairy products for local consumption; skins, especially the famous karakul (incorrectly known as Persian lambskin), and wool (both for export and for carpet weaving) are also important products. There are about 21,500,000 sheep—6,500,000 are of karakul quality—3,600,000 cattle, and a number of goats, donkeys, and camels.

Minerals. Natural gas is obtained from the fields in northern Afghanistan and is exported to the Soviet Union through a pipeline begun in 1968. Coal production is about 190,000 tons per year and rock-salt production about 40,000 tons. Other minerals exploited are lapis lazuli, limestone, gypsum, and talc.

Industry. Modern industry is just beginning and is based mainly on agricultural and pastoral raw materials. The lead has been taken by the cotton-textile industry, which has a yearly production of about 187,000,000 feet; next come cement, sugar, vegetable oil, woollen and artificial silk textiles, and fruit-processing industries. An important nitrogenous fertilizer factory, based on natural gas, is under construction in Mazār-e Sharīf with Soviet assistance, is due for completion in 1974. Besides organized industries, Afghanistan has a number of traditional handicrafts, which account for a fair proportion of the Traditional crafts

country's foreign-exchange earnings. The most important craft is carpet making; sheepskin coats (*pūstīns*) and embroidery are also produced.

Tourism. Active support, which includes the ready granting of tourist visas, is being given by the government for the development of the tourist industry. There are two luxury hotels, one in Kābul and another on the road between Kābul and Paghmān; other hotel facilities are being expanded throughout the country. As a result, the number of foreign tourists increased tenfold in the decade beginning in 1960, reaching the figure of 100,000 per year in the early 1970s. The best known tourist attractions of the country are at Bāmīān, where the tallest Buddha in the world (175 feet) has been cut into a cliff, and Band-e Amīr, the lakes famous for the multicoloured hue of their waters.

Management of the economy. *The private and public sectors.* The government budget is divided into two parts, ordinary and development. The former covers administrative activities and the latter investment expenditure, incorporated in the national plans of development. Total domestic revenue is usually exceeded by expenditures; the difference is covered through deficit financing. (The amount of deficit financing, represented by government borrowing from the Central Bank, reached an average of 10 percent of total revenue between 1966 and 1971). In addition, large amounts are invested in the economy by foreign loans and grants. The main contributors of foreign assistance have been the Soviet Union, the United States, West Germany, and the People's Republic of China.

Deficit financing and foreign assistance

Total annual imports from foreign countries consistently exceed exports, resulting in an adverse balance of trade that is covered partly by invisibles, such as tourism, but mainly by foreign assistance. Roughly one-third of Afghanistan's exports go to the Soviet Union, and 26 percent go to India and Pakistan. Next come the United Kingdom, Switzerland, and the United States. The Soviet Union is the leading source of imports, followed by Japan, India, and the People's Republic of China.

The private sector primarily includes agriculture and livestock breeding. There is a mixed pattern of small, medium, and large holdings. Landlordism, involving large holdings, is more pronounced in the less populated areas of southern and western Afghanistan, while in the more populated eastern and central parts individual medium- and small-sized holdings are more common. North of the Hindu Kush both large and small holdings are found side-by-side.

The bulk of the trade and transport and most of the manufacturing industries are in the hands of private entrepreneurs. Public enterprise is confined to a section of the foreign trade, to mining, and to some industries. Although for some years a liberal foreign investment law has been in operation, so far no substantial foreign capital has been invested.

Taxation. Indirect taxes (mainly import duties) form the largest single source of revenue, accounting for about 50 percent of total domestic revenue. Income from government enterprises is the next largest source of income, followed by taxes on commercial transactions, and direct taxes, such as individual and corporation income taxes. Land revenue at one time was the main source of government income, but by the early 1970s it represented less than 2 percent of total revenue; livestock taxes have been abolished by Parliament.

Trade unions. Although some 50,000 workers are employed in the organized industries, trade unions have not yet made their appearance. After the promulgation of the 1964 constitution, which guaranteed the right of association, some agitation occurred among the workers for better wage and work conditions. The only employer association is the semi-official Chamber of Commerce and Industries.

Problems and prospects. Theoretically, the government is committed to developing a mixed, guided economy by means of a series of five-year plans. In practice, the effectiveness of such a policy is limited by the paucity of government resources and by a cumbersome bureaucracy. So far, three five-year development plans have been implemented, and the fourth (for 1972 to 1977) is being prepared.

The five-year plans

Although Afghanistan has made a major developmental effort, it is still faced by many difficult problems. The rate of growth of the population is increasing, and is estimated to be around 2 percent per year. Because of this, the country, which formerly was self-sufficient in food, by 1972 was importing 5 to 10 percent of the food it consumed. The economy is almost stagnant, and unemployment is widespread. On the other hand, an important change can be discerned in the long-dormant agricultural sector. It appears that the farmers have at last accepted modern methods, such as the use of improved seeds, fertilizers, and mechanized equipment. In the case of wheat, yields are already on the increase, and the country may soon be able to dispense with food imports. It is hoped that an upsurge in agriculture may also help to reactivate other sectors of the economy as well.

Transportation. Being a landlocked country, Afghanistan is entirely dependent on transit facilities from its neighbours for its international trade. Lacking both railways and navigable rivers, it has to rely on roads as the mainstay of its transport system. These factors put a heavy burden on the country's economy in the form of higher transport bills; they also add to the difficulty of integrating the transport system of the country with those of its neighbours. During the first two five-year plans, covering the period from 1956 to 1967, efforts were directed toward upgrading the highway system and to connecting the main trading centres of the country with each other, as well as with the railheads or road networks of neighbouring countries. At the same time, the road network was integrated with the plan of the United Nations Economic Commission for Asia and the Far East (ECAFE) for the construction of the Asian Highway, running from Istanbul to Saigon. The new road network of Afghanistan, which connects the Soviet railheads in Kushka and Termez with those of Pakistan, at Chaman and Peshāwar, respectively, opens the way for the first time for direct overland transit between the Soviet Union and the Indo-Pakistani subcontinent. The most important Afghan highways are those connecting Kābul with Shīr Khān, on the Soviet border in the north, and with Peshāwar in Pakistan. The first, which cuts across the Hindu Kush through the Salang tunnel, which is about two miles long, has shortened the distance between Kābul and the rich northern provinces by about 120 miles. Other paved roads link Kābul with Qandahār, Herāt, and Mazār-e Sharīf as well as those cities with the frontier towns of Pakistan, Iran, and the Soviet Union. The total length of modern, paved roads is 1,550 miles. Provincial and district centres in the country are connected with the national highways by a series of gravel roads; these provide some facilities for the movement of goods and people in fair weather but become almost impassable in the rainy and snowy seasons.

The road system

In spite of the rapid development of motor transport, camels and donkeys are still used as draft animals in the remote parts of the country as well as in farm areas. In the countryside, many have not abandoned their cherished horses, which are important for prestige.

Civil aviation has been increasing in importance. Almost all provincial centres possess at least a seasonally operable airport, while there are major international airports at Kābul and Qandahār. There are daily or weekly services from Kābul to different parts of the country, as well as to neighbouring countries, to the Middle East, and to Europe. Ariana Afghan Airlines, owned jointly by the Afghan government and Pan American World Airways, operates international flights. Internal service is provided by Bakhtar Alwotana Company, a completely Afghan venture.

ADMINISTRATION AND SOCIAL CONDITIONS

The structure of government. *The constitutional framework.* Although constitutions were promulgated by the monarchy in 1923 and in 1931, Afghanistan remained virtually an absolute monarchy until 1964. Un-

der the 1931 constitution, the government was not responsible to Parliament nor did it have to receive a vote of confidence. Under the 1964 constitution, freedom of press, speech, and association was guaranteed, and, until the 1973 coup, members of the royal family were excluded from membership in the Cabinet, Parliament, and the High Court.

The 1964 constitution

The 1964 constitution was based on the separation of the executive, legislative, and judiciary authorities. The legislative power was vested in shūrā (Parliament), which was composed of two houses—the House of the People (the lower house) with 216 members, and the House of the Elders (the upper house) with 87 members. Members of the House of the People were elected by universal, secret, and direct suffrage for a period of four years. One-third of the membership of the House of the Elders was elected in the same way (with one member from each province); one-third was nominated by the king; and the remaining third was elected indirectly by the provincial assemblies, when they were formed, from among their members. All legislation was passed by both houses; in the case of difference of opinion, a joint committee, composed of an equal number of members from each of the two houses, was appointed to settle the difference. The final authority for financial bills was vested in the House of the People.

The ministers were appointed by the prime minister, who was nominated by the king, but they had to obtain a vote of confidence from the House of the People, both on his Cabinet and on his policies. The House of the People also had the power to dismiss the Cabinet or to impeach any of its members.

Regional and local government. Afghanistan has a centralized system of government. For administrative purposes the country is divided into 28 provinces (*velāyāt*), each administered by a governor appointed by the king on the advice of the Ministry of Interior. The constitution provided for the future establishment of provincial assemblies, with advisory power only, to be elected by universal suffrage in each province. Local administration in the towns and cities was vested in elected city councils, headed by a mayor who was also elected. The constitution also provided for the gradual extension of local *jirgahs* (councils) down to village level.

The political process. Elections are held every four years on the basis of adult suffrage. After 1964 women acquired the right to vote, but outside of the chief cities this right has seldom been exercised. In the first Parliament after the promulgation of the constitution, four women were elected. The second returned none, but a certain number of women were nominated by the king as members of the upper house. By the early 1970s, no elections for provincial assemblies or municipalities had been held. Although a party law has been enacted by both houses, it has not been promulgated, and consequently no recognized legal parties exist in the country. Several groups that function somewhat like political organizations, however, have been formed around newspapers of different shades of ideology, and may, in due course, serve as nuclei for political parties.

Women's suffrage

After the promulgation of the constitution, less than 20 percent of the eligible voters exercised their franchise. This was partly because of the poor organization of the voting and the limited number of polling stations opened. Parliament was dominated by conservative elements, such as landlords and religious leaders, who were opposed to reforms and were reluctant to bring about change. At the other end of the ideological spectrum were the students who had been agitating for drastic reforms. No substantial middle-of-the-road movement has emerged, and in its absence both Parliament and government were groping in the dark for a new balance.

The judiciary. Under the constitution, the judiciary, as mentioned, was an independent organ of the state; it was composed of a Supreme Court and of a number of subordinate courts. The Supreme Court consisted of a supreme judge and eight members—all appointed by the king—whose terms of office were renewable only after a period of ten years. The courts applied the provisions of the constitution and the laws. In the absence of such provisions they were guided by Sharīʿah (Islāmic law).

Armed forces. Until 1973 the king was the supreme commander of the army. The minister of defense, who is a member of the Cabinet, assures liaison with the civil authorities. Every male Afghan is required to serve two years in the armed forces. This period, however, has been reduced to one year in the case of college graduates, who are trained as reserve officers. A military academy in Kābul trains cadets. There was a royal bodyguard of brigade strength. A corps of gendarmes, under the Ministry of Interior, was responsible for internal security and for the supervision of the frontiers. The small air force is equipped mostly with modern Soviet aircraft and has bases at Bagrām and Shīndand. A parachute corps is attached to the air force. The police force, attached to the Ministry of Interior, remains responsible for the maintenance of law and order and for the protection of life and property of the people.

Education. According to the constitution, education is free at all levels. Elementary education is compulsory wherever it is provided by the state. Foreigners are not allowed to operate education institutions without previous authorization, and even then cannot enroll Afghan nationals in them. There are primary schools throughout the country, but secondary schools exist only in the provincial and sometimes in the district centres. The ratio of boys to girls attending schools is about seven to one. Kābul University, known as Kābul Pohantoon, was founded in 1946. It incorporated a number of previously existing faculties, the oldest of which was the faculty of medicine, established in 1932. In 1963 a second university, Nangarhār, was formed in Jalālābād but by the early 1970s consisted only of a medical college.

The universities

Public health and welfare. Malaria has been eradicated through systematic campaigns supported by the World Health Organization of the United Nations. Progress has also been made against smallpox and cholera, and good headway has been made against typhoid fever and typhus. Thanks mainly to a favourable climate, Afghans are on the whole healthy and tough. The child mortality rate is high, however, especially among the nomads and in the poor mountainous regions. Tuberculosis and other diseases that are caused by an insufficient diet are widespread in the country. Syphilis and trachoma, an infectious eye disease that results in blindness, occur especially in the western regions. Hazārajāt is still plagued by leprosy. Where doctors are lacking, people still use medicinal herbs, sometimes with good effect. The country has almost 70 hospitals.

A woman's welfare society centred in Kābul and with branches in important cities provides guidance and adult education to women. Its activities have been enhanced since the emancipation of the women in 1959; since that date the wearing of the veil is no longer obligatory, women have found employment in offices and shops, and some Afghan women have received university education. The Sera Myasht, or Red Crescent, functions under the auspices of the crown prince. Besides providing ad hoc assistance in case of calamity, it runs a number of welfare houses for the elderly and the crippled.

Social conditions. *Wages and cost of living.* There are about 50,000 workers employed in industry. The law does not provide for a minimum wage, but protection is extended to wage earners against arbitrary dismissal or in the event of work accidents.

The cost of living tends continually to rise. The price index of the main consumer goods in 1969–70, for example, stood at about 200, compared with 100 in 1961–62.

Social and economic divisions. The bulk of the population in the rural areas consists of small farmers exploiting their tiny plots of land or working as tenants for a local landowner. In southern and western Afghanistan, peasants tend to be exploited by the feudal landlords. The majority of the city and town dwellers are artisans, small traders, or government employees. There is also a business community of merchants and industrialists. A new community of entrepreneurs is being formed.

Cultural life and institutions. *The cultural milieu.* Outside the main cities, most Afghans live in what may best be described as a peasant tribal society. The main characteristics of such a society are its conservatism and the fact that it is inward looking rather than outward looking. Kinship rather than family is the basis of social life, and determines the patriarchal character of the community. Religion also plays a significant role. Among the Pashtun tribes a system of traditions, known as *pashtūnwalī*, and administered by the *jirgha*, or council of elders, is used in place of the penal code.

Characteristics of social life

In the cities and to a lesser degree in the rural areas, however, the emergence of a new educated class and the introduction of new technologies in agriculture and industry have precipitated a shift away from the traditional inward-looking attitudes.

Cultural activities and institutions. Archaeological research carried out since 1922 has uncovered many fine works of art both in the pre-Islāmic and Islāmic periods. A revival of these arts, together with interest in new forms of expression, has given a new dynamism to artistic creation. Of the new painters, some draw direct inspiration from the Herāt school of the 15th-century Timurid period; others are influenced by Western style. Some of the old monuments of architectural value are being restored and redecorated, through government initiatives, by modern artists and craftsmen.

The theatre, as known in the West, has flourished only since about 1960. Adaptations of European classics were introduced at first, but the present trend favours the didactic treatment of themes from everyday life. A revival of traditional folk singing has gone hand in hand with imitation of modern Western and Indian music.

The performance of the *attan* dance in the open air has long been a feature of Afghan life. The *attan* has now been institutionalized as the national dance. To the accompaniment of drums and pipes, the dancers—mostly men but at times including women—form a circle after which, taking each other by the hand, they revolve in circles of their own. The dance is continued at an increasingly fast tempo until the dancers are exhausted.

The Afghan Historical Society publishes a bimonthly, *Āryāna*, in Pashto and Dari, and a quarterly, *Afghanistan*, in French and English with occasional articles in other Western languages. The society also arranges for the publication of works by writers doing research on Afghan history, on newly found manuscripts, and on the historical sites and monuments of the country.

The Pashto Ṭolana, or Pashto Society, is working for the development of Pashto literature and for its promotion among the non-Pashto-speaking populations. It publishes the monthly *Kābul* and arranges for the publication of works on Pashto linguistics and literature, as well as of old Pashto manuscripts.

Press and radio. Before the promulgation of the 1964 constitution, no independent press existed in Afghanistan except for a short period in the early 1950s. A number of dailies, weeklies, and magazines were, however, published by the Press Department and other government agencies. In the early 1970s the daily *Ḳslāh*, published in Kābul under the aegis of the Press Department, was considered the official organ of the government. Since 1965, several independent periodicals based on different ideologies have been started.

Radio Afghanistan broadcasts from Kābul on short and medium wavelengths. In addition to local and foreign music, it broadcasts programs on subjects of such general interest as hygiene, agriculture, and the popularization of science and technology.

Prospects for the future. As stated above, the destiny of Afghanistan has, in the course of history, been closely connected with its geographic location and with the general pattern of communication in the area. For a long time the country was kept isolated from the mainstream of prevailing civilization by the policies of its powerful neighbours, British India and tsarist Russia, as well as by a lack of communication facilities. After gaining its independence in the early 1920s, Afghanistan made determined efforts to emerge from its isolation, but it was only after the British left the Indian subcontinent that any measure of success could be achieved. Now, with the advent of air communication and the construction of a new road system, the country hopes for a new era in which relations will be further developed with its neighbours and with distant regions.

Jealous of its strict neutrality, Afghanistan has not adhered to the Regional Cooperation for Development (RCD) formed by Turkey, Iran, and Pakistan. The increasing volume of trade between the Soviet Union (especially the Soviet Central Asian republics) on the one hand and Pakistan and India on the other, however, impels these countries to take advantage of the Afghan road system for the transit of their goods, and the day may not be far off when the country regains its former position as a crossroads of trade and communication.

The overthrow of the monarchy in 1973, however, casts some doubt on the nation's future; at the same time, the fact that the country has been able to combine a measure of democracy with political stability, indicates that it may also find, in due course, peaceful solutions to its economic and social problems.

BIBLIOGRAPHY. JAMAL-UD-DIN AHMAD and MUHAMMAD ABDUL AZIZ, *Afghanistan: A Brief Survey* (1934), contains important information on the geography, history, and government of the country. JOHANNES HUMLUM, *La Géographie de l'Afghanistan* (1959), is a comprehensive geography. Social life and customs are discussed in MOHAMMED ALI, *The Afghans* (1965); MARY LOUISE CLIFFORD, *The Land and People of Afghanistan* (1962); and DONALD N. WILBER *et al.*, *Afghanistan: Its People, Its Society, Its Culture* (1962). See also MOHAMMED ALI, *A Cultural History of Afghanistan* (1964); and BENJAMIN ROWLAND, *Ancient Art from Afghanistan: Treasures from the Kabul Museum* (1966). MOUNTSTUART ELPHINSTONE, *An Account of the Kingdom of Caubul and Its Dependencies in Persia, Tartary, and India* (1842), is the first detailed account of Afghanistan by an English observor. More recent travel books and guides, which contain contemporary descriptions, include: MARY BRADLEY WATKINS, *Afghanistan: Land in Transition* (1963); SOHRAB K. KATRAK, *Through Amanullah's Afghanistan*, 2nd ed. (1963); ROSANNE KLASS, *Land of the High Flags: A Travel-Memoir of Afghanistan* (1964); CAROLA BURKNER, *Afghanistan*, 2nd ed. (1970), in German; and ROLAND MICHAUD, *Afghanistan* (1970), in French. ARNOLD FLETCHER, *Afghanistan: Highway of Conquest* (1965), is a well-documented account based on collections of the India Office library and the British Museum, manuscripts and materials available in Afghanistan, and discussions with Afghan leaders. For statistical data, see the publications of the Afghanistan Information Bureau. Also of interest are EDWARD STENZ, *The Climate of Afghanistan, Its Aridity, Dryness and Divisions* (1946); and the GOVERNMENT OF AFGHANISTAN, MINISTRY OF PLANNING, *Survey of Progress* (1957–). Additional sources of information may be found in DONALD N. WILBER, *Annotated Bibliography of Afghanistan*, 3rd ed. (1968).

(M.S.F.)

Afghanistan, History of

The earliest recorded use of a form of the name Afghan (Avagāṇa) was by an Indian astronomer Varāhamihira (6th century AD). The Afghans proper comprise a group of tribes speaking an Iranian language, Pashto, and centred, ever since they were first recorded, in the Sulaimān Mountains in the east of what is now Afghanistan. Their legends survive in traditional forms, of which none is more than 500 years old. There is no evidence that the ancient Afghans occupied any areas of modern Afghanistan other than those in which Pashto is spoken, and only in modern times have they settled in the parts occupied by such peoples as Tadzhik and Ḥazāras, over whom they hold feudal domination. The distribution of Pashto as the predominating local language is from a little north of Jalālābād in the northeast, southward to Qandahār (Kandahār), and thence westward to Farāh and Sabzevār. Pashto also extends into the Northwest Frontier Province of Pakistan and thence southward throughout the Sulaimān ranges as far as Quetta.

HISTORICAL BEGINNINGS (TO THE 7 TH CENTURY AD)

The routes that, in ancient times, linked India with the Middle East and both of these regions with China and

Early cultural influences

Central Asia, passed through Afghanistan. Thus, while the main cultural influences were those of Iran, others from India and Central Asia were constantly being experienced. Notice of the area historically begins with the Achaemenian rulers of Iran in the 6th century BC; and, although it is not easy to locate all the names derived from their inscriptions or from Herodotus (whose account the inscriptions remarkably confirm) and other sources, the principal regions are recognizable. Thus Bactria must have included the Afghan plain of the Amu-Darya, Arachosia the neighbourhood of Qandahār, and Aria the country of the Harīrūd Valley around Herāt; while the Drangiana of the inscriptions and later Greek sources must have been a region of southwestern Afghanistan, including, perhaps, parts of Sīstān. Sindhu and Gandhāra may be taken as the Sind and Northwest Frontier regions of West Pakistan and probably did not include lands of Afghanistan. Problematic names remain: along with the Gandhārians in the seventh *satrapy* (province) of Darius, Herodotus mentions the Sattagydae, Dadicae, and the Aparytae (possibly the same as the Afrīdī, a later Afghan tribe), and one of these may indicate tribes occupying the country around Kābul and the ancient Kapisa, which was included from the time of Alexander the Great in the 4th century BC with the Hindu Kush (mountains) in Paropamisus, whence arose the satrapy of the Paropamisadae. Then there are the Pactyes of Herodotus, whom some have sought to connect with the Pathans, speakers of Pashto with the modern dialectic variant Pakhtūn.

The Achaemenians and Greeks. Achaemenid rule was first established in the regions of Afghanistan by Cyrus and consolidated by Darius I. It survived until Alexander the Great overthrew the dynasty. After Alexander's death his empire, in turn, broke up and the eastern satrapies passed to the Seleucid dynasty. By 322 BC Candragupta Maurya, who founded the Maurya dynasty of India, had gained control of northern India, and, after his defeat of Seleucus (*c.* 305 BC), he won also the eastern provinces of Afghanistan, the Paropamisadae, and Arachosia. Confirmation of the Indian hold over these two provinces half a century later is provided by the so-called Aśokan edicts discovered at Qandahār and Pol-e Darūntah. By *c.* 250 BC Diodotus, the local Greek ruler in Bactria, broke away from the Seleucids and established an independent kingdom. At the same time the Parthians also broke away. Probably Sīstān (Sakastan) and Aria were in dispute between Bactria and Parthia, finally falling to Parthia. The Bactrian Greek dynasty, famed for its coins, flourished for about a century. Soon after 175 BC its ruler Demetrius moved southward through the mountains to capture the region of Kābul and thence to invade northwestern India and establish the Indo-Greek dynasty there. Parthian power, however, was increasing along the western frontiers of Bactria; and to the north beyond the Oxus, some time before 125 BC, the barbarous Indo-Scyths, the Yüeh-chi, appeared in Sogdiana. This tribal group had been driven from the Chinese frontier region by a non-Iranian group, the nomadic pastoral Hsiung-nu. Before or with them came Iranian tribes of nomadic Śaka, who seem, before 130 BC, to have made a pact with the Parthians and to have settled in Sīstān whence they spread eastward through Arachosia and, in the 1st century BC, into India. Meanwhile, the Yüeh-chi had wrested Bactria from the Greeks and settled there.

The Kuṣāṇas

Nomadic rulers. At the beginning of the Christian era one group, henceforth known as the Kuṣāṇas, followed the Greeks into the mountains and soon emerged as conquerors of northern India. The zenith of Kuṣāṇa power was reached under Kaniṣka, whose rule stretched from beyond Mathurā to Kābul and thence from Bactria to the frontiers of China. During this time (2nd century AD) Afghanistan may have been momentarily united under one rule. The Kuṣāṇas were patrons of the arts and of religion and under Kaniṣka Buddhism flourished.

The empire did not long survive Kaniṣka, though for centuries Kuṣāṇa princes continued to rule in the various provinces. During the 4th century the Sāsānids offered a growing challenge from Iran, while in the latter half of the century fresh incursions of northern tribes took place —the Little Kuṣāṇas under their leader Kidara, closely followed by or associated with the White Huns, or Chionites. These people are generally associated with the Hsiung-nu, the Mongolian tribes who, centuries before, had driven out the Yüeh-chi. Two branches were discerned—the northern settled in Bactria and the Hindu Kush and were known eponymously as the Hephtalites (after their hero Hephtha), and the second (Zavula) branch moved southward through the mountains to Kābul and Arachosia. They, too, came under Sāsānid influence and formed the centre for their raids into India between 455 and 510. In India they were long remembered for their barbarity; and Hūṇas, as they were called, became a name regarded with horror. At about the same time, their victory (484) over the Sāsānid king Fīrūz marked the beginning of a period in which they were more the masters than the vassals of Iran. Hephtalite domination lasted for a century, until in 565 the Western Turks, another newly arrived tribal group from central Asia, made an alliance with the Sāsānid Khosrow Anūshīrvān and destroyed their power. The Turks ruled Afghanistan until their defeat by the Chinese in 658, by which time Arab armies had already reached the area.

MEDIEVAL PERIOD (7TH–18TH CENTURIES)

The advance of Islām

The first Muslim dynasties. The advance of Islām into Iran and Central Asia was rapid; but in the mountains of the Hindu Kush it was much slower, and not until the 9th and 10th centuries did the last non-Muslim dynasties disappear from Kābul. The initial control of the caliphate soon declined in these remote areas, however, and the 9th and 10th centuries also witnessed the rise of numerous local dynasties, both Turkish and Iranian. One of the earliest of these was the Arab Ṭāhirid dynasty, which included Balkh and Herāt in its kingdom of Khorāsān. This dynasty was succeeded by a native Iranian dynasty, the Ṣaffārids, one of whom, Yaʿqūb ebn Leyś aṣ-Ṣaffār made a series of whirlwind campaigns throughout Afghanistan. In the north, soon after, the local princes became feudatories of the powerful Sāmānids, also an Iranian house, under whose rule Bukhara, Samarkand, and Balkh enjoyed a golden age.

The Ghaznavids. In the middle of the 10th century a former Turkish slave Alptigin seized Ghazni from its former ruler and built himself a principality. He was succeeded by another former slave, Subüktigin, who extended the conquests to Kābul and the Indus. His son was the great Maḥmūd of Ghazna, who came to the throne in 998. Maḥmūd not only made extensive conquests in the north and west, thus bringing all Afghanistan under his sway, but also conquered the Punjāb and Multan and carried his raids into the heart of India. The hitherto obscure town of Ghazni became a splendid city, as did the second capital at Bust (Lashkar Gāh).

The Ghūrids. Maḥmūd's descendants continued to rule from Ghazni over a gradually diminishing empire until 1152, when ʿAlā-ʾud-Dīn Ḥusayn of Ghūr, a mountain-locked region in northwestern Afghanistan, descended upon the city, sacked it, and drove the last Ghaznavid out to India. The Ghūrids appear to have been of Iranian origin. ʿAlā'-ud-Dīn's nephew, Muʿizz ud-Dīn Muḥammad, commonly known as Muḥammad of Ghūr, drove Oğuz Turkmen nomads from Ghazni and in 1175 invaded India. After his death in 1206, his general, Quṭb-ud-Dīn Aybak, became the first sultan of Delhi.

The Khwārezm-Shāhs. Shortly after Muḥammad of Ghur's death, the Ghurī empire fell apart and Afghanistan was occupied by Sultan ʿAlā' ad-Dīn Muḥammad, the Khwārezm-Shāh, or *shāh* of Khwārezm (Khiva). The Khwārezm-Shāh dynasty was at this time at the zenith of its power, its territories extending from Chinese Turkistan in the east to the borders of Iraq in the west. ʿAlā' ad-Dīn Muḥammad, early in 1219, marched with a large army against Baghdad. (F.R.Al.)

The Mongol invasion. ʿAlā' ad-Dīn had hardly reached his destination when he learned that Mongol hordes

under Genghis Khan had invaded the eastern part of his empire, seizing some of the most important cities of Transoxiana and everywhere spreading death and destruction. He turned to repel them, but though he had 400,000 horsemen under his command, they were not considered enough. Avoiding a battle, he retreated to a small, deserted island in the Caspian Sea, where he died shortly afterward (1220).

War of liberation

Soon after 'Alā' ad-Dīn's death, his energetic son Jalāl ad-Dīn Munkabirni decided to wage a war of liberation against the Mongols. With Ghazni as his military headquarters, he rallied the Afghan highlanders and at Parwandarah near Kābul inflicted a crushing defeat on the Mongols under Kutikonian. Genghis Khan, who was then at Herāt, hastened to avenge the defeat and laid siege to Bāmiān, a place of great strategical importance in the Hindu Kush. There Mutugen, the Khān's grandson, was killed, an event so infuriating to Genghis Khan that when he captured the citadel he ordered that no living being be spared. The beautiful city of Bāmiān was utterly destroyed.

Advancing on Ghazni, Genghis won a great victory over Jalāl ad-Dīn, who then fell back toward the Indus (1221), where he made a final but unsuccessful stand. Genghis Khan permitted Jalāl ad-Dīn to escape into India, and then returned to Afghanistan through Peshāwar and brought a dreadful retribution to the Afghans' gallant resistance. He turned the whole country from Herāt and Balkh to Qandahār and Ghazni into a wilderness, and butchered many thousands of people.

Later medieval dynasties. After his death in 1227, Genghis Khan's vast empire fell to pieces. In Afghanistan some local chiefs succeeded in establishing independent principalities and others acknowledged Mongol princes as suzerains. This state of affairs continued until the end of the 14th century, when Timur (Tamerlane) conquered a large part of the country and repeated the horrors of the Mongol invasion, though on a lesser scale.

The Timurids of Herāt. Timur's successors, the Timurids, were great patrons of learning and the arts who enriched their capital city Herāt with fine buildings. Under their rule (1404–1507) Afghanistan enjoyed peace and prosperity. The first Timurid king, Shāh Rokh Mīrza, Timur's fourth son, was an enlightened monarch of literary taste; during his reign Herāt became an important centre of Islāmic civilization. Sultan Ḥoseyn Beyqara, another Tīmūrid king, was also a patron of scholars and artists; Jāmī, a great poet, and Behzād, a painter, flourished at his court.

The Mughals. Early in the 16th century the Uzbeks, a Turkic people, rose to power in Central Asia under Muḥammad Khan Shaibani, who triumphantly entered Herāt in 1507. Bābur, a direct descendant of Genghis and Timur and the founder of the Mughal dynasty in India, was at this time in possession of Kābul, which, after being driven from Fergana by the Uzbeks, he had made the capital of an independent principality in central Afghanistan. In December 1510, the Ṣafavid *shāh* Esmā'īl, to avenge an insult, attacked Shaibani, besieged him in Merv, and slew him in an ambush. Bābur thought the time opportune to re-establish his authority in Central Asia, but he was defeated and forced to retreat. Falling back on Kābul, he made that city a base for conquests in the east and south, taking Qandahār in 1522 and then turning his attention toward India. A series of expeditions enabled him to seize the Punjab, and in 1526 he marched on Delhi. He defeated Ibrāhīm, the last of the Lōdī Afghan kings of India, at Pānīpat, 80 miles (129 kilometres) north of Delhi, and established the Mughal empire, which lasted until the middle of the 19th century. Bābur made Āgra his capital, and all eastern Afghanistan up to the Hindu Kush became a part of the empire. When he died in 1530 his body, as he had wished, was taken from Āgra to Kābul to be buried in his favourite garden.

Foundations of the Mughal empire

During the next 200 years Afghanistan had no political identity and was parcelled between the Mughals of India and the Ṣafavids of Persia—the former holding the regions south of the Hindu Kush and the latter Herāt and Farāh. Qandahār was for many years in dispute between them.

MODERN HISTORY (SINCE THE EARLY 19TH CENTURY)

The national awakening. Although it was not until later in the 18th century that Afghanistan emerged as a united nation under Afghan rule, the national awakening may be said to have been started earlier by Bayazid Ansar, popularly known as Pīr-e Rowshan ("The Enlightened Leader") and Khushḥal Khān Khaṭak, a soldier-poet. Later in 1709, Mīr Veys Khān, an influential leader of the Hotaki Ghilzay tribe, led a rising against Gorgīn Khān, the Persian governor of Qandahār. The Persians, caught unawares, offered no effective resistance and were slaughtered.

The Hotakis. Mīr Veys Khān governed Qandahār until his death in 1715. In 1716 the Abdālīs of Herāt, encouraged by his example, took up arms against the Persians and under their leader, Asadullāh Khān, succeeded in liberating their province.

Maḥmūd, Mīr Veys's young son and successor, was not content with holding Qandahār, and in 1722 he led about 20,000 men against Isfahan; the Ṣafavid government could offer no effective resistance and surrendered after a six-month siege.

Maḥmūd immediately set out to reorganize the Persian administration and economy; but he died in 1725. His young cousin Ashraf, who succeeded him, came to power at a critical time: the Afghans in Persia were threatened on all sides; the Russians were advancing from the north, and the Ottoman Turks seized part of western Persia. Internal strife and jealousy among the Afghan chiefs made the situation worse. Shāh Ashraf met the dangers with courage and determination. He halted the Russian advance at Darband and in 1726 inflicted a crushing defeat on the Turks, who were marching on the capital. Meanwhile news was brought to him that a brigand chief, Nāder Qolī Beg, had attacked Ṭūs and was on his way to take Meshed and Herāt. Collecting all available troops Ashraf hastened northward to meet him but was defeated at Dāmghān on October 2, 1729. The Afghan position in Persia was thereafter untenable, and during the retreat Ashraf himself was murdered by a Baluchi chief in Sīstān.

Nāder Shāh. After consolidating his position in Persia, Nāder Qolī Beg turned his attention against Herāt to the east. The Abdālī chief Zūlfeqār Khān appealed to Qandahār for help, and Saydal Khān, a Ghilzay chief of repute, was sent to his aid. In an engagement at Kāfar Qal'eh, near the Afghan border, the Afghans inflicted a crushing defeat on the Persians (1731). Nāder himself, with only 8 men, was cut off from the rest of the army but managed to escape; in 1732, on receiving reinforcements, he marched again against Herāt. The besieged fought desperately, but dissension forced them to sue for peace. Impressed by their valour, Nāder treated them with courtesy and recruited many of them into his own army. He was elected *shāh* of Persia, with the name Nāder Shāh, in 1736.

In 1737 Nāder Shāh advanced with an army of 80,000 men against Qandahār. The city fell after a year's siege and Nāder then came into conflict with the Mughal empire of India. He seized Ghazni and Kābul and, advancing into India, defeated the Mughals at Karnāl, north of Delhi, in 1739. Then, having seized the jewels and treasure of the capital, including the Kūh-e-Nūr diamond and the famous peacock throne, he returned through Afghanistan to Persia. He was assassinated at Khabushan in 1747.

The Durrānī dynasty. After the death of Nāder Shāh the empire disintegrated. In Afghanistan, Aḥmad Khān Abdālī (later called Aḥmad Shāh Durrānī), with a contingent of 4,000 Afghans, entered Qandahār and in 1747 was elected king of the Afghans by a tribal assembly. Aḥmad Shāh changed the tribal name from Abdālī to Durrānī and soon proved himself a capable ruler, securing the affection of his people and welding them into a strong nation. In 25 years he was able to liberate Afghanistan from foreign rule and to lead conquering armies

Liberation from foreign rule

from Meshed to Kashmir and Delhi and from the Amu Darya to the Arabian Sea. His popular manners and his indomitable courage and other martial virtues won him the title of *bābā* ("father of the nation"). When he died, about 1773, he was succeeded by his son Tīmūr, who received but nominal homage from the feudal chieftains. Much of Tīmūr's reign was spent in quelling their rebellions. Observing how the people of Qandahār hated him, Tīmūr shifted his capital to Kābul.

Zamān. After the death of Tīmūr in 1793, his fifth son, Zamān, seized the throne with the help of Sardār Pāyenda Khān, an influential Barakzai chief. Zamān's elder brothers, Homayūn and Maḥmūd, governors of Qandahār and Herāt, respectively, took up arms against him. At first Zamān was successful, Homayūn being defeated and blinded and forced to take refuge in Persia. Zamān then turned to India with the object of repeating the exploits of Aḥmad Shāh. The move alarmed the British, who induced Fatḥ ʿAlī Shāh of Persia to bring pressure upon the Afghan king and to divert his attention from India. The *shāh* went a step further, helping Maḥmūd with men and money and encouraging him to advance on Qandahār. Maḥmūd, assisted by his vizier, Fatḥ Khān, eldest son of Sardar Pāyenda Khān, and by Fatḥ ʿAlī Shāh, took Qandahār and then pushed on toward Kābul. Zamān, in India, hurried back to Afghanistan. There, after various vicissitudes, he was handed over to Maḥmūd, blinded, and imprisoned in the Bala Hissar fort at Kābul (1800).

Maḥmūd. Maḥmūd, who took no interest in the welfare of his people, squandered the nation's treasure on his personal pleasures, and left the affairs of state in the hands of Fatḥ Khān. Some of the chiefs who had grievances against the king or his ministers joined forces and invited Shāh Shojāʿ, a full brother of Zamān, to Kābul. The intrigue was successful. Shāh Shojāʿ occupied the capital, and Maḥmūd sued for peace.

Shāh Shojāʿ. The new king, Shāh Shojāʿ, ascended the throne in 1803. The chiefs had become powerful and unruly, and the outlying provinces were asserting their independence. The Sikhs of the Punjāb were encroaching upon Afghan territories from the east, while the Persians were threatening from the west; and the *shāh* could not conciliate Fatḥ Khān.

At this time Napoleon was at the zenith of his power in Europe. He proposed to Alexander I of Russia a combined invasion of India. A British mission, headed by Mountstuart Elphinstone, met Shāh Shojāʿ at Peshāwar to discuss mutual defense against this threat, which never developed. A treaty of friendship was concluded (June 7, 1809), the *shāh* promising to oppose the passage of foreign troops through his dominions. Hardly had the mission left Peshāwar, however, when news was received that Kābul had been occupied by the forces of Maḥmūd and Fatḥ Khān. Shāh Shojāʿ's troops were routed, and Shāh Shojāʿ withdrew from Afghanistan. In 1810, after making some ineffective efforts to recover his throne, he was captured and imprisoned in Kashmir. He escaped in 1812 and went to Lahore, hoping to receive help from Ranjit Singh, the ruler of the Punjāb; but instead the Sikh ruler robbed him of the famous Kūh-e-Nūr diamond. Finally, Shāh Shojāʿ found asylum with the British at Ludhiāna.

Treaty with the British

The rise of the Barakzai. The Barakzai were now dominant throughout Afghanistan. This situation incited the jealousy of Kāmrān, Maḥmūd's eldest son, who had a grudge against Fatḥ Khān. His chance came in 1816 when Fatḥ Khān and his younger brother Dōst Moḥammad Khān, who had been sent to defend Herāt against a Persian army, insulted Kāmrān's sister, the daughter-in-law of the governor of Herāt. When Kāmrān seized and blinded Fatḥ Khān, the vizier's brothers, 21 in all including Dōst Moḥammad, took up arms. Shāh Maḥmūd asked Fatḥ Khān to write to his brothers and order them to submit. Fatḥ Khān refused, and the *shāh* had him cut to pieces.

Dōst Moḥammad. Advancing from Kashmir in 1818, Dōst Moḥammad took Peshāwar and Kābul and drove the *shāh* and Kāmrān from all their possessions except Herāt, where they maintained a precarious footing for a few years. The rest of Afghanistan was parcelled out among the Barakzai brothers. In the confusion that followed, Afghanistan suffered heavily—Balkh was seized by the ruler of Bukhara; the trans-Indus Afghan districts were occupied by the Sikhs; the outlying provinces of Sind and Baluchistan assumed independence. Ghazni, Kābul, and Jalālābād fell to Dōst Moḥammad, who thus became the most powerful of the Barakzai *sardars*.

Dōst Moḥammad established the Barakzai (or Moḥammadzai) dynasty. His position secure at Kābul, he decided to recover Peshāwar by force from the Sikhs. Declaring a *jihād*, or holy war, in 1836, he collected an army of Muslims and, confident of victory, advanced on Peshāwar. The shrewd Sikh leader, however, sowed dissension in Dōst Moḥammad's camp, the invading army melted away, and Peshāwar was permanently lost to the Afghans.

In November 1837 Moḥammad Shāh of Persia laid siege to Herāt, which was looked upon by the British as the key to India. The Russians supported the Persians; and the British, fearful that Persia was falling completely under Russian influence, determined to stem the tide by entering into alliances with the rulers of Herāt, Kābul, and Qandahār. A British mission to Kābul under Capt. (later Sir) Alexander Burnes in 1837 was welcomed by Dōst Moḥammad, who hoped the British would help him recover Peshāwar. Burnes, however, would not give him the required assurances; and when a Russian agent appeared in Kābul, the British broke off negotiations and left for India.

With the failure of Burnes's mission, the governor general of India, Lord Auckland, ordered an invasion of Afghanistan, with the object of restoring Shāh Shojāʿ to the throne. In April 1839, after suffering great privations and guerrilla attacks on the way, a British army entered Qandahār; Shāh Shojāʿ was then crowned in the mosque next to the mausoleum of Aḥmad Shāh. Ghaznī and Kābul were captured in the following July and August, Shāh Shojāʿ being crowned again at Kābul. Dōst Moḥammad escaped first to Balkh, then to Bukhara, where he was arrested. The Afghans, however, could tolerate neither a foreign occupation nor a king imposed on them by a foreign power; and insurrections broke out. Dōst Moḥammad escaped from prison and returned to Afghanistan to lead his partisans against the British. In a battle at Parwandarah on November 2, 1840, Dōst Moḥammad had the upper hand; but the next day he surrendered to the British in Kābul. He was honourably treated by the British and deported to India with the greater part of his family.

First Afghan War 1838–42

Outbreaks continued all over the country; and when winter set in, the British in Kābul found their position untenable. Terms for their withdrawal were discussed with Akbar Khān, Dōst Moḥammad's son. Negotiations were delayed by Sir William Hay Macnaghten, the British political agent, whose attempts to make headway against Akbar Khān failed, and Macnaghten was killed by Akbar Khān himself. On January 6, 1842, about 4,500 British and Indian troops, with 12,000 camp followers, marched out of Kābul. Bands of Afghans swarmed around them and the retreat ended in a massacre that few survived.

Though in the summer of the same year British forces reoccupied Kābul, the new governor general, Lord Ellenborough, decided on the immediate evacuation of Afghanistan. In 1843 Dōst Moḥammad was allowed to return to his country, which was now without a ruler (Shāh Shojāʿ had been killed by his own subjects after the British had left the capital). Dōst Moḥammad reigned until 1863. During the last 20 years of his reign he occupied Qandahār, Mazār-e Sharīf, Kataghan, and Herāt.

Shīr ʿAlī. Shīr ʿAlī Khān, Dōst Moḥammad's third son, then became *amīr*, but he spent five years establishing his authority over his two elder brothers. Shīr ʿAlī's reception of a Russian mission at Kābul and his refusal to receive a British one, on British terms, led directly to the war of 1878–81. On November 21, 1878, British armies

Second Afghan War, 1878–80

entered the three great passes leading into Afghanistan; by the end of the year they had captured Jalālābād and Qandahār. Shīr ʿAlī, leaving his son, Yaʿqūb Khān, as his regent in Kābul, sought help from the Russians, but they advised him to make peace. Shīr ʿAlī died on February 21, 1879.

Yaʿqūb Khān. The Treaty of Gandamak (May 26, 1879) recognized Yaʿqūb Khān as *amīr*. He agreed to receive a permanent British embassy at Kābul; to conduct his foreign relations with other states in accordance "with the wishes and advice" of the British government; and to allow rectification of the frontier. In return, he was promised help against foreign aggression. This British triumph, however, was short-lived. On September 3, 1879, the British envoy and his escort were murdered in Kābul. British forces were again dispatched, before the end of October they occupied Kābul. Yaʿqūb, forced to abdicate, was sent to India, where he died in 1923.

ʿAbdorraḥmān Khān. The British remained in Kābul during the winter of 1879–80. In the meantime Lord Ripon, the viceroy, set out to achieve a speedy but peaceful settlement of the Afghan question. Accordingly, in July 1880, the British recognized ʿAbdorraḥmān Khān, a grandson of Dōst Moḥammad, as *amīr* of Kābul and agreed not to require the admission of a British resident anywhere in Afghanistan.

As the British prepared to leave Kābul, they learned that at Maywand, near Qandahār, a British force had been annihilated by Ayūb Khān, Yaʿqūb Khān's brother. On August 31, 1880, a force of Indian and British troops attacked and defeated Ayūb at Baba Wali, three miles west of Qandahār. The British then handed over Qandahār to a governor deputed by ʿAbdorraḥmān and evacuated Afghanistan. Ayūb, from his base at Herāt, once more seized Qandahār; but he was decisively defeated there by ʿAbdorraḥmān, who thus won possession of both Qandahār and Herāt. Ayūb fled to Persia.

ʿAbdorraḥmān established peace and a strong central government in Afghanistan. During his reign (1893) the Durand line was laid down to fix the boundary between Afghanistan and the Northwest Frontier Province of India. ʿAbdorraḥmān died in October 1901 and was succeeded by his eldest son Ḥabībollāh Khān.

Ḥabībollāh Khān. In Ḥabībollāh's reign the foundations of the Habibia School were laid, roads were constructed, and steps were taken to introduce machinery and other modern devices. A weekly Persian paper, *Serāj ol-Akhbār*, was started, and infused new life into the country, playing a great part in the national awakening. In February 1919, having observed strict neutrality during World War I, Ḥabībollāh wrote to the viceroy of India demanding recognition by the Paris Peace Conference of Afghanistan's "absolute liberty, freedom of action and perpetual independence." On the night of February 20, 1919, however, he was assassinated in his camp at Kalagosh, near Jalālābād.

Amānollāh Khān. Ḥabībollāh's third son, Amānollāh Khān, seized the treasury and the throne and proclaimed himself king, displacing Ḥabībollāh's younger brother, Naṣrollāh Khān. Amānollāh proclaimed Afghanistan's complete independence in both internal and external affairs. The reluctance of the government of India to recognize this independence led to a third war between Great Britain and Afghanistan. Hostilities began on May 3, 1919; but the fighting was inconclusive, and when peace was signed at Rāwalpindi on August 8 (supplemented by further agreements signed November 22, 1921), the independence of Afghanistan was recognized. Amānollāh then established diplomatic relations with the outside world.

Third Afghan War, 1919, and independence

Amānollāh introduced internal reforms on European lines. This innovation brought him into conflict with reactionaries who rigidly opposed modernization. Civil war broke out; and in January 1929 a brigand chief, Ḥabībollāh, siezed Kābul. Amānollāh abdicated and left for Qandahār. That spring, after an unsuccessful attempt to recover the capital, he went into exile. Ḥabībollāh declared himself king and ushered in a reign of terror that lasted for nine months.

Nāder Khān. Ḥabībollāh was finally driven from the capital in October 1929 by tribal forces led by Amānollāh's cousin Muḥammad Nāder Khān and his brothers. In a tribal assembly held shortly afterward, Muḥammad Nāder Khān was unanimously elected king. Ḥabībollāh and his accomplices were arrested, court-martialed, and executed. Thereafter Nāder Khān ruled without challenge and devoted his attention to consolidating the country. New schools, a literary academy, a military college, and a medical college were established; some Afghan graduates were sent abroad for higher education; and a cautious constitution was introduced in 1931 providing for a bicameral legislature.

Zahir Shah. The king was assassinated, while distributing prizes to schoolchildren on November 8, 1933. Nāder was succeeded by Muḥammad Zahir, his only surviving son, who continued his father's policy of national consolidation and cautious progress. Neutrality was observed in World War II.

Following Pakistan's independence (1947), friction arose between Afghanistan and Pakistan over the establishment of an independent Pathan state within Pakistan, and relations remained strained for some years. In 1956 a five-year plan for better roads and better means of communication and for the development of mines, industry, education, and agriculture was drawn up. To finance it the Afghan government signed agreements with certain foreign governments, notably the Soviet Union and the United States. Afghanistan also imported foreign experts and construction firms and welcomed specialists and educationalists to help advance its modernization program. (M.Al.)

Afghanistan in the modern world

A second five-year plan, launched in 1961, stressed the development of communications and heavy industry, with notable reliance on foreign aid. Soviet and United States engineers executed several important projects to open up parts of the country hitherto inaccessible.

Afghanistan soon became sensitive to the danger that massive aid, whether from the United States or the U.S.S.R., might create pressure threatening the neutrality enforced by the national passion for independence. It therefore made persistent efforts to broaden the basis of its international friendships. In 1963 a more flexible foreign policy was adopted. Diplomatic relations with Pakistan, broken off in 1961, were resumed; communications and trade, mutually advantageous, substantially increased. Relations with the U.S.S.R. continued cordial although Afghanistan had signed a border agreement with China in 1963 and had accepted aid from that country as well as from the United States, Great Britain, West Germany, France, Sweden, and also the World Bank.

It soon became clear that Afghanistan's international reputation was bound up with its adaptation to the modern world. Economically this entailed reducing dependence on foreign aid by financial reform, by the encouragement of domestic capital investment, and by the consolidation of developments already achieved. The government provided massive inducements—an income tax moratorium, free imports of capital goods, and a protective tariff—to enlist private capital in light industry in order to free its own resources for the completion of major projects already begun. A new constitution, drafted in 1964 and approved by a grand council of all the chiefs and notables, converted Afghanistan into a true constitutional monarchy, excluding members of the royal family from political office, with a complete division of power between the executive, the legislative, and the judiciary.

In 1965, Muḥammad Hashim Maiwandwal became prime minister and proved to be the most progressive first minister the country had yet known. Under the prime minister's lead, Afghanistan made steady advances in every sphere of domestic and foreign policy; growing prosperity and contentment at home was matched by an enhanced reputation abroad. When ill health obliged him to resign in October 1967, he was succeeded by his foreign minister, Nur Ahmed Etemadi, who continued his policies at home and abroad. In 1969, the second parliamentary election brought in conservative elements

who had formerly held aloof; possibly by way of reaction to this, student unrest broke out in Kābul and other centres. Problems with the parliament forced Etemadi to resign in 1971; he was succeeded by the ambassador to Rome, Abdul Zahir. In July 1973 Afghanistan was proclaimed a republic in a bloodless coup led by a former premier, Mohammed Daud Khan, while the king was in Italy. (L.F.R.-W.)

BIBLIOGRAPHY

General works: PERCY SYKES, *A History of Afghanistan*, 2 vol. (1940), is still most readable although much subsequent research has been done; W. KERR-FRASER-TYTLER, *Afghanistan: A Study of Political Developments in Central and Southern Asia*, 3rd ed. rev. (1967), is slightly more technical but very good. B.M. МАССОН and B.A. РОМОДИН, *История Афганистана*, 2 vol. (1964–65), is a thorough and comprehensive account with a very extensive bibliography; and ARNOLD FLETCHER, *Afghanistan, Highway of Conquest* (1965).

Pre-Muslim period: W.W. TARN, *The Greeks in Bactria and India*, 2nd ed. (1951), offers a lively account of the aftermath of Alexander the Great's campaigns in the east. See also A.K. NARAIN, *The Indo-Greeks* (1957); ROMAN GHIRSHMAN, *Bégram: Recherches archéologiques et historiques sur les Kouchans* (1946), which describes the important excavations at Bégram; and *Chionites-Hephtalites* (1948).

Early Muslim period: Particularly recommended are C.E. BOSWORTH, *Sīstān Under the Arabs, from the Islamic Conquest to the Rise of the Saffārids* (1968) and *The Ghaznavids* (1963); and *Ḥudūd al-'Ālam: The Regions of the World*, 2nd ed. (1970), English translation by V. MINORSKY with notes, additional materials, and introduction by C.E. BOSWORTH. *The Cambridge History of Iran*, vol. 5, *The Saljuq and Mongol Periods*, ed. by J.A. BOYLE (1968), has material for the pre-Mongol period. For the early modern period, see relevant chapters of *The Oxford History of India*, 3rd ed. (1958).

Modern Afghanistan: LUDWIG W. ADAMEC, *Afghanistan, 1900–1923: A Diplomatic History* (1967), is essentially a diplomatic history of Afghanistan in the early 20th century that deals primarily with the activities and policies of the Afghan rulers and explains how Afghanistan survived the confrontation of Russia and England. It is a well-documented and reliable source of information on the period. JAMAL-UD-DIN AHMAD and MUHAMMAD ABDUL AZIZ, *Afghanistan: A Brief Survey* (1934), is a fine book that deals with the history of the country from the earliest times to 1934. GEORGE GRASSMUCK *et al.*, *Afghanistan: Some New Approaches* (1969), contains articles by various writers on different Afghan ethnography, literature, political modernization, and foreign relations and has an excellent bibliography. VARTAN GREGORIAN, *The Emergence of Modern Afghanistan* (1969), is an excellent study of modern Afghan history to 1946, with special emphasis on social-economic development.

(L.F.R.-W./M.Al./F.R.Al.)

Africa

Africa is the second largest continent in the world, being smaller only than Asia. Its population numbers about 350,000,000, and its area—11,700,000 square miles (30,300,000 square kilometres), about three times the size of Europe—covers about one-fifth of the total land surface of the world.

The continent is bounded on the north by the Mediterranean Sea, on the west by the Atlantic Ocean, on the east by the Red Sea and the Indian Ocean, and on the south by the mingling waters of the Atlantic and Indian oceans. Africa measures about 5,000 miles from north to south and about 4,600 miles from east to west. Its northern extremity is Ras ben Sekka, near ar-Ra's al-Abyaḍ (Cap Blanc), Tunisia; its southern extremity is Cape Agulhas, South Africa; its farthest point east is Ras Hafun, near Cape Guardafui in Somalia; its western extremity is Pointe des Almadies, on Cap Vert (Cape Verde), Senegal. In the northeast, Africa was joined to Asia by the Sinai Peninsula until the construction of the Suez Canal, which is now generally deemed to constitute the boundary between Africa and Asia.

Paradoxically, the coastline of Africa—18,950 miles in length—is shorter than that of Europe, because there are few inlets and few large bays or gulfs.

Africa's islands

Off the coasts of Africa a number of islands are associated with the continent. These include Madagascar, one of the largest islands in the world, the Comoro Islands, Mauritius, Réunion, and other islands to the southeast; the Seychelles, Socotra, and other islands to the east; the Azores, Madeira, and the Canary Islands to the northwest; the Cape Verde Islands, the Bijagós Archipelago, Fernando Po, and São Tomé and Príncipe to the west; and Ascension, Saint Helena, and Tristan da Cunha to the southwest.

From north to south, the continent is cut almost equally in two by the Equator, so that most of Africa lies within the tropical region, bounded by the Tropic of Cancer (23°28′ N) and the Tropic of Capricorn (23°28′ S). Because of the bulge formed by West Africa, the greater part of Africa's territory lies north of the Equator. Africa is crossed from north to south by the prime meridian of Greenwich (0° longitude), which passes a short distance to the east of Accra in Ghana.

In antiquity, the Greeks are said to have called the continent Libya and the Romans Africa, perhaps from the Latin *aprica* (sunny), or the Greek *aphrike* (without cold). The name Africa, however, was chiefly applied to the northern coast of the continent, which was, in effect, regarded as a southern extension of Europe. The Romans, who for a time ruled the North African coast, are also said to have called the area south of their settlements Afriga, or the Land of the Afrigs—the name of a Berber community south of Carthage. Another explanation occasionally offered is that the name applied to a productive region of what is now Tunisia and meant Ears of Corn. The word Ifriqiyah is apparently the Arabic transliteration of Africa.

The whole of Africa can be considered as a vast plateau rising steeply from narrow coastal strips and consisting of very ancient crystalline rocks. The plateau's surface is higher in the southeast and tilts downward toward the northeast. In general, the plateau may be divided into a southeastern portion and a northwestern portion. The northwestern part, which includes the Sahara (desert) and that part of North Africa known as the Maghreb, has two mountainous regions—the Atlas mountain system in North Africa, believed to be part of a system that extends into southern Europe, and the Ahaggar (Hoggar) plateau in the Sahara. The southeastern part includes the Ethiopian Highlands, the East African Plateau, and—in the eastern part of South Africa, where the plateau edge falls downward in a scarp—the Drakensberg Mountains.

According to a widely accepted theory, Africa, South America, peninsular India, Australia, and Antarctica were joined together millions of years ago to form a continent called Gondwanaland; after the breakup of Gondwanaland, the component fragments drifted apart. A part of the evidence for the theory is that the African plateau and the Brazilian plateau possess similar geological formations.

The continent has a number of volcanoes, some of which are now extinct; they include the cones of Kilimanjaro, Mt. Kenya, and Elgon, as well as the mountains of the Virunga range, in the east, and Mt. Cameroon, in the west.

The rift valley

One of the most remarkable features in the geological structure of Africa is the great East African Rift Valley, which lies between 30° and 40° E. The rift itself begins beyond the continent's limits, in Syria to the north, from where it travels down the Jordan Valley, runs along the length of the Red Sea and enters the African plateau near Ethiopia. It continues southward through Kenya to Tanzania, where it virtually disappears below the surface. This constitutes the eastern branch of the rift valley; it contains a series of lakes, including Lake Rudolf and Lake Nyasa. The western branch of the rift, which describes an arc along the eastern border of Zaire (formerly Congo [Kinshasa]), contains Lakes Albert, Edward, Kivu, and Tanganyika. A short branch of the rift, extending southwest from the northern end of Lake Tanganyika, includes the valley of the Luangwa and part of the Zambezi Valley.

The African plateau includes a number of basins, surrounded by divides. These include the basin surrounding Lake Chad, to the northeast of Nigeria; the Congo Basin; and the Baḥr al-Ghazāl Basin in The Sudan.

In tropical Africa, where rainfall is heavy, many rivers empty into the sea. The four longest are the Nile (which flows northward beyond tropical latitudes), the Congo, the Niger, and the Zambezi. In the course of their lengthy journeys to the sea, they encounter cataracts as they break through the plateau region down to the narrow coastal plain, incising deep gorges in the process. Although often navigable for sections of their courses, African rivers in general cannot be compared for navigability to their counterparts on other continents—a circumstance that helped to delay the economic development of the continent.

The climates of Africa are dominated by its position astride the Equator. Temperatures are high for most of the year in the northern and southern tropical zones. The northern tropical zone is hottest between March or April and September, that in the south between September and March. At high altitudes in parts of the interior, however, temperatures are modified by the height of the mountains. This is especially so in the mountains of East Africa, Ethiopia, and parts of Central and southern Africa. The northern zone and the southern coastal zone, which lie outside the tropics, experience more moderate temperature and a milder climate. The climate of the continent is further modified by the influence of ocean currents, such as the cooling Benguela Current, which flows north from the Cape of Good Hope toward the Congo, and the warming Mozambique Current, which passes through the channel between Madagascar and the southeast coast of Africa.

Despite the fact that it is the second largest continent, Africa contains only about 10 percent of the total world population and can therefore be described as underpopulated. This fact is due, in part, to the large expanses of desert. The Sahara alone—the largest desert in the world—covers one-quarter of the entire continent.

The greatest concentrations of population are found in Nigeria and around the shores of the East African lakes. Nigeria, with a population of about 55,000,000, is the most populous of the African states, followed by Egypt (about 34,000,000), and then by Ethiopia (25,000,000).

Africa's historical legacy, formed of a variety of influences of varying strengths, has shaped its culture and its patterns of religious affiliation. Herodotus, the Greek historian of the 5th century BC, reported that in about the year 600 BC the Phoenicians sailed around the African continent. Even earlier, more than 1,000 years before Christ, they had established trading posts in northern Africa. Later influences included those of the Greeks and of the Romans. Climatic and other factors have exerted considerable influence on the patterns of human settlement in Africa. While the greater part of the continent has always been inhabited by black peoples, there have also occurred major immigrations from both Asia and Europe. The extensive movement of Arabs into North Africa began in the 7th century AD, while a more limited settlement along the East African coast began in the following century. The extreme northern and southern regions of the continent, and certain highland parts of East Africa, because of their moderate climatic conditions, were later to attract settlers from Europe, such as the Dutch in South Africa from the 17th century onward and the French in Algeria as well as the British in Central and South Africa in the 19th century. In the colonial period the demand for labour, for such enterprises as the railways in East Africa, and the sugar plantations in Natal, also led to the importation on a significant scale of workers from the Indian subcontinent. Because of the humid nature of its climate, West Africa alone remained largely unaffected by migrations from outside the continent. Of all such foreign settlements in Africa, that of the Arabs made the greatest impact. The Islāmic religion, which the Arabs carried with them, spread from North Africa into many areas south of the Sahara, so that West African peoples such as the Hausa, Bornu, Songhai, and Malinke (Mandingo) are today largely Islāmized. Although most of North Africa and much of West Africa is thus Muslim in religion, the one area that is an exception is that of Ethiopia in the northeast, where many follow the Christian Coptic religion. There are also some Coptic communities in Egypt.

While the northern third of Africa is populated for the most part by people of Arab stock superimposed upon an older Berber substratum, people of Negro stock dominate the greater part of the remainder of the continent. Apart from the Islāmized peoples of West Africa and the growing Christian communities to be found especially near the coast, traditional animistic forms of religion survive. In black Africa important kingdoms arose, such as those of Ghana, Kanem, Mali, and Songhai—all of which flourished between the 8th and the 16th centuries—or the more recent states such as Dahomey, Ashanti, and Buganda, which survived until their conquest by the colonial powers at the end of the 19th century. It was also black Africa that became centrally involved in the slave trade with the Americas and the Caribbean. While Arab and European writers had often been impressed by the degree of social and political advancement in the old kingdoms of black Africa, the institution of slavery gradually led to the categorization of black peoples everywhere as "backward" and "inferior." This concept survived the abolition of the slave trade between Africa and Europe and the Americas in the 19th century and is only now disappearing through black activism, the influence of liberal thinkers, and the political influence of independent black African states.

The southern third of Africa, inhabited by Bantus and other peoples, was used by Europeans as a way station on the route from Europe to Asia round the Cape. This part of Africa still contains the majority of the people of European descent living on the continent, although they are still far outnumbered by the black population of the region.

African agriculture

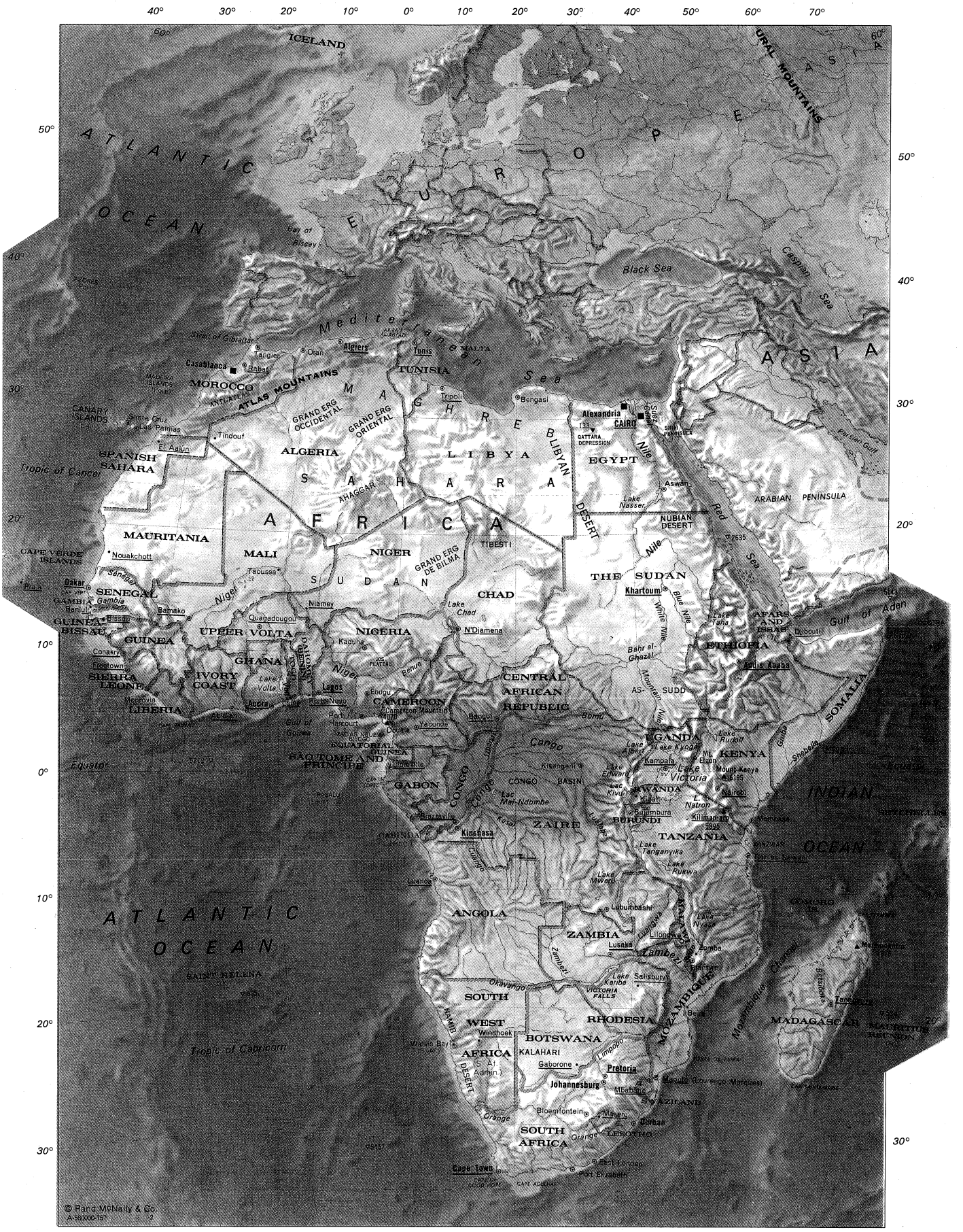
The agricultural produce of tropical Africa, which includes such commodities as palm oil, peanuts (groundnuts), cocoa, cotton, and rubber, has long been in demand among the industrial nations of Europe and the Americas. During the era of colonization, which extended over much of the 19th century and lasted until about 1960, this led to the division of most of the continent into European-ruled colonies or areas of influence. Since then, many African territories have emerged into independent nationhood. Their exports, however, still remain largely agricultural, and the prices that they obtain for their goods are still determined to a large extent by the more developed and industrialized nations.

The mineral resources of Africa are equally important. Among the most important mineral regions are the gold- and diamond-mining areas of South Africa, the Zambia copper belt, and the mineral-rich Shaba (Katanga) area of Zaire. Other mineral reserves are to be found in Ghana, Sierra Leone, Liberia, Mauritania, and elsewhere. Since the 1950s, oil reserves have been located in many parts of Africa, the most important so far being found in Algeria, Libya, Angola, and Nigeria.

The great rivers of Africa are being harnessed to provide power for industrial development or to irrigate large areas. Thus, large dams, such as the Kariba Dam on the Zambezi River, the new Aswān High Dam on the Nile, the Akosombo Dam on the Volta River in Ghana, and the Owen Falls Dam where the Nile issues from Lake Victoria, have been constructed, and more are envisaged.

The continent is well known for its wild animals. Big game animals are found roaming the savanna (grassy parkland) regions of southern and eastern Africa and, to a lesser extent, in the savanna areas north of the West African coastal belt. Game reserves have been established in such countries as Kenya, Uganda, Tanzania, Zambia, and South Africa; apart from helping to maintain the ecological balance, they also constitute tourist attractions.

Profound political changes occurred in Africa during the decade of the 1960s, at the conclusion of which most of Africa had attained political independence. Territories still remaining in a state of dependency included the Portuguese-held territories, Rhodesia (though it had unilaterally proclaimed its independence from Britain), South West Africa (Namibia), and some smaller territories, such



AFRICA

Size of symbol indicates relative size of town • ◉ ■

Elevations and depths in metres

0 400 800km

0 200 400 600 800mi

as the French-administered territory of the Afars and Issas, the French-administered Comoro Islands, and the Spanish Sahara. In the early 1970s, however, it was considered to be only a matter of time before the territories of the entire continent would have achieved independence.

Africa's trade, which in the past, because of geographical proximity, had primarily been in Europe, is now also being channelled toward Asia and the Americas as well. While different degrees of political stability prevailed in different African states, a trend toward democratic government can often be discerned. African states, in their foreign policies, are largely nonaligned, avoiding involvement in East–West conflicts, although occasionally receiving aid from either Western countries, from the Soviet Union, or from China. Meanwhile, it seems clear that the remaining years of the 20th century will be among the most important for the destiny of Africa.

This article covers, in general terms, the physical and human geography of the African continent. There are separate articles on the major geographical features of Africa, as well as on contemporary African countries and cities. Linguistic patterns are described in the article AFRICAN LANGUAGES, and ethnic and cultural patterns in the article AFRICAN PEOPLES AND CULTURES. For historical aspects, see CENTRAL AFRICA, HISTORY OF; COLONIALISM; EAST AFRICA, HISTORY OF; EGYPT, HISTORY OF; ETHIOPIA, HISTORY OF; MADAGASCAR, HISTORY OF; NILOTIC SUDAN, HISTORY OF THE; NORTH AFRICA, HISTORY OF; SOUTHERN AFRICA, HISTORY OF; TIMBUKTU; WEST AFRICA, HISTORY OF. For cultural aspects, see AFRICAN PEOPLES, ARTS OF. See also CONTINENTS, DEVELOPMENT OF. (D.S.H.W.N.)

This article is divided into the following sections:

I. Geological history
  Stratigraphy and structure
    The Precambrian
    The Paleozoic Era
    The Mesozoic Era
    The Tertiary Period
    The Quaternary Period
  Present geological structures and relief
II. Physical geography
  Relief
  Climate
    Factors influencing the African climate
    Climatic regions
  Drainage
  Soils
    Soil types
    Soil problems
III. Vegetation and animal life
  Vegetation
    Vegetational zones
    Influences on vegetation: past and present
  Animal life
    Genera and distribution
    The effects of man
    Ecology
IV. Natural resources
  Mineral resources
    Mineral fuels
    Metallic deposits
    Nonmetallic deposits
  Water resources
    Surface water
    Groundwater
    Availability for human use
  Biological resources
    Botanical resources
    Animal resources
V. Human resources
  The human ecology
    The pattern of settlement
    Migrations and population movements
    Population distribution
  Forms of government
VI. Political geography
  Exploration and settlement
  Current political geography
  Major language patterns
VII. Resource development
  Industries
  Power
  Agriculture, irrigation, and land use
    Cereals
    Fruits and vegetables
    Vineyards
    Dairy products
    Miscellaneous crops
VIII. Commerce
  Transportation
  Trade
    Internal trade
    External trade
    Problems and prospects
IX. Demographic patterns
  Current patterns
  Future trends

I. Geological history

The African continent essentially consists of an ancient Precambrian Shield (formed between 570,000,000 and 4,000,000,000 years ago) that has since been folded, metamorphosed (altered in composition or structure by heat and pressure), and granitized. As a result of later erosion, Precambrian rock outcrops on 57 percent of the continent's surface. The Precambrian Shield

An examination of the shield leads to three preliminary observations. Firstly, geologically the continent of Africa includes the Arabian Peninsula, which was closely linked to it until the appearance of the Red Sea Trough. Secondly, North Africa (contemporary Morocco, Algeria, and Tunisia) belongs to the Mediterranean and Alpine area; it emerged and became joined to Africa only quite recently. Thirdly, the Precambrian Shield shows numerous traces of life. These are found, for example, in the micro-organisms of the Fig Tree System of rocks in South Africa, which are 3,200,000,000 years old; in the Bulawayo series, which are 2,700,000,000 years old; in the Ivory Coast series, which are 2,500,000,000 years old; in the Gabon series, which are 1,700,000,000 years old; and in all of Africa in formations that are between 600,000,000 and 1,000,000,000 years old.

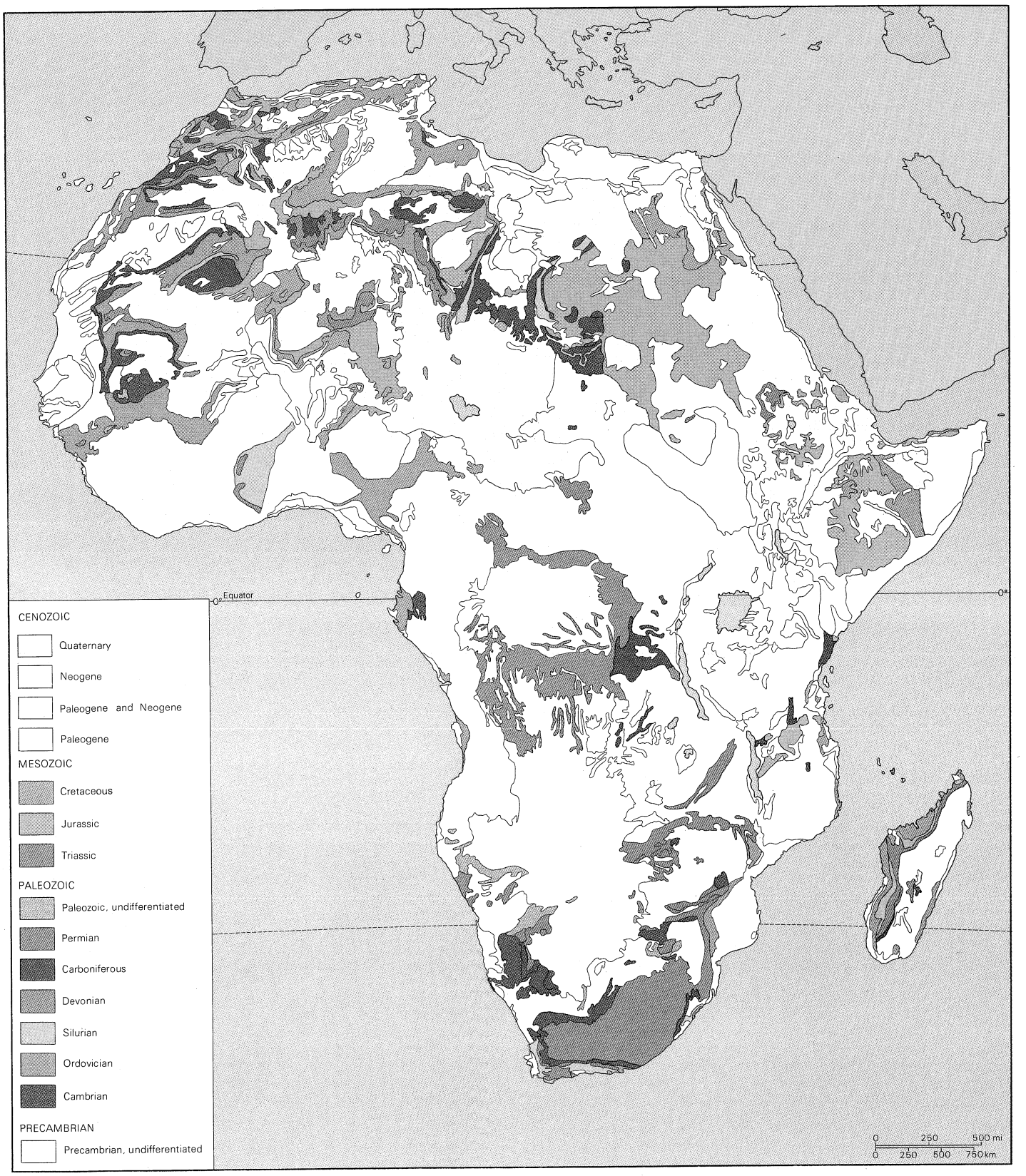
Geological analysis is simpler after the end of the Precambrian due to the very fact of the existence of this enormous shield. There are no traces of geosynclines (large linear troughs, in which sediments accumulated over geological time cause the trough itself to sag (deeper) or of orogenic (mountain-building) processes comparable to those of other continents. On the western rim of the continent, however, there are very eroded traces of ancient mountains that have been called the Mauritanids; these thrust from the west toward the east. If, however, they were associated with a geosynclinal trough, this was situated further west in the Atlantic Ocean.

The relative geological stability that characterized the continent did not, however, hold true in the north. The Tethys Sea, a sea ancestral to the Mediterranean, ran along the northern edge of the Sahara, and a geosynclinal depression covered the region; from this depression the Alpine chain of the Atlas Tellien (Tell Atlas) and Atlas Saharien (Saharan Atlas) later arose during the Tertiary Period (65,000,000 to 2,500,000 years ago).

Seas of the Paleozoic Era (from 570,000,000 to 225,000,000 years ago) covered northwestern Africa, extending first from the Mediterranean to the Gulf of Guinea, and then to South Africa. At the end of the era, during the Permian Period (from 280,000,000 to 225,000,000 years ago), the Permian Sea advanced on the coasts of the Indian Ocean as far as the Mozambique Channel.

The Mesozoic Era (from 225,000,000 to 65,000,000 years ago) is represented by marine deposits of the Jurassic Period (from 190,000,000 to 136,000,000 years ago) in the Spanish Sahara, Senegal, and on the Indian Ocean coast. Marine deposits of the Lower Cretaceous Period (from about 136,000,000 to 100,000,000 years ago) occur on all African coasts, while deposits of the Upper Cretaceous Period (from 100,000,000 to 65,000,000 years ago) are located in the Sahara and on the coasts of the Atlantic and Indian oceans. During the Tertiary Period, North Africa emerged, and the rift valley of the Red Sea subsided, isolating Arabia from the continent.

Finally, a glance at any geological map of Africa reveals firstly the distinctive effects of the great rifts,



Geological structure of Africa.

running northwest to southeast and northeast to southwest, and secondly the volcanic activity that took place both in the Tertiary and in the Quaternary periods (up to 2,500,000 years ago).

STRATIGRAPHY AND STRUCTURE

The Precambrian. The Precambrian landmass in Africa comprises four groups of interconnected cratons (immobile parts of the Earth, usually of large size). These are firstly, the West African Craton, which became stable about 2,000,000,000 years ago and which today comprises the states of Mauritania, Mali, Guinea, Liberia, the Ivory Coast, Ghana, and the western Ahaggar (Hoggar) mountain region of southern Algeria; secondly, the Arabo-Nubian Shield, which includes the eastern Ahaggar mountains of Algeria, the Tibesti moun-

The African cratons

tain region (extending into Chad, Niger, and Libya), Egypt, and the Arabian Peninsula; thirdly, the Congolese and Kalahari cratons; and fourthly, Madagascar, the eastward extension of which is unknown, as it subsided into the Indian Ocean during the Upper Cretaceous Period.

The final part of the Lower Precambrian (which ended about 3,000,000,000 years ago) was marked by important orogenic processes, traces of which are clearly defined in the West Nile System on the borders of Zaire and The Sudan, and in the Swaziland System of southern Africa. Still later mountain building at the end of this era is visible in West Africa (in Liberia), in East Africa (in the Bulawayan, Dodomian, and Sebakwian systems, found in Tanzania, Kenya, and Uganda respectively), and in Madagascar (in the Androyan System). It was during this period that the first discoverable traces of life are found in the fossilized unicellular algae of the Fig Tree Series in Swaziland and in the Orange Free State (South Africa).

The Middle Precambrian, which lasted from about 3,000,000,000 to 2,000,000,000 years ago, was characterized by orogenic activity that was manifested in such series as the Amsaga and Ghallaman of Mauritania, the Kambui of Sierra Leone, the Zadinian of the Congo Basin, the Dibayan of South West Africa, the Kavirondian of East Africa, the Great Dyke of Rhodesia, the Transvaal System of South Africa, and the Vohibory of Madagascar. The period ended with a new mountain-building episode, the widespread continental effects of which may be seen in such formations as the Berberids of the Anti-Atlas Mountains in North Africa, the West African Birrimian Formation, the Buganda–Kibali–Toro Formation of East Africa, and the Witwatersrand, Bushveld, and Limpopo formations in southern Africa.

The Upper Precambrian, which lasted from about 2,000,000,000 to 1,000,000,000 years ago, may be divided into two orogenic cycles. The first cycle is marked by the uplifting of the Anti-Atlas Mountains of Morocco and of such formations as the Tarkwaian (Ghana), Francevillian (Gabon), and Mayumbian (Congo). The second cycle consisted of the rise of the western Kibarian mountain chain (Zaire), the Irumian chain in Central Africa, and the Waterberg and Matsap chains of South Africa.

The final, or terminal, stage of the Precambrian, lasting from about 1,000,000,000 to 570,000,000 years ago, is represented throughout Africa by several thousand feet of sandstone, schist (medium- or coarse-grained metamorphic rock containing micaceous materials), and dolomite limestone (limestone rich in magnesium carbonate) with stromatoliths (fossil calcareous algae). Among these are such formations as the Adoundounian Series of Morocco, the Adrar Series of Mauritania and Mali, the Western Congo System, the Katangan Series of Zaire and Zambia, and the Damara–Otavi System of South Africa. The end of the Precambrian, at which time the Lufilian and Katangan orogenies occurred in Central Africa, is separated from the Cambrian Period by a glacial period.

The Paleozoic Era. The Paleozoic Era, which comprises the Cambrian, Ordovician, Silurian, Devonian, Carboniferous, and the Permian periods, includes two principal mountain-building episodes—the Caledonian and the Hercynian. During this era the continent may be said to have taken shape—the cratons fused and the landmass emerged. A glacial period of this era was marked by deposition tillites (sedimentary rocks composed of cemented clay and rock), which may be seen in southern Morocco, throughout West Africa and sub-Equatorial Africa as far as South West Africa. The tillite is considered as marking the break between the end of the Precambrian and the beginning of the Cambrian Period (which lasted from 570,000,000 to 500,000,000 years ago).

Paleozoic mountain-building episodes

Marine fossils of the Cambrian Period are found in Southern Morocco, the Spanish and Mauritanian Sahara, and South West Africa. In Egypt and Arabia their presence has been revealed by drilling. Elsewhere they remain unknown.

During the Ordovician Period (from 500,000,000 to 430,000,000 years ago) fossiliferous marine sandstone completely covered North and West Africa, including the Sahara. The table sandstones of South Africa constitute its only other trace. This period is, in addition, remarkable for the tectonics (forces or conditions causing movements of the Earth's crust) of the Caledonian mountain-building episode. Although no great folds resembling those of Europe and North America occurred, the episode was marked by at least two important epeirogenic (continent-forming) movements, which raised the continental table of the central and western Sahara by about 5,000 feet. Each emergence resulted in the creation of valleys that became flooded when the continent subsided. It is known that toward the end of the period the Sahara became glaciated, and tillites and sandstones filled the valleys.

Although no major folding took place, except in local instances, granites were rejuvenated. Many authentic Precambrian granites show ages of between 45,000,000 and 550,000,000 years old. These are not new granites, for nowhere do they pierce the covering of Cambrian or Ordovician sediments. They constitute a phenomenon that is found from the central Sahara to as far south as South Africa and Madagascar.

A complete change of sedimentation characterized the Silurian Period, which lasted from 430,000,000 to 395,000,000 years ago; this is indicated by the deposits of graptolitic schists (containing fossils of colonies of extinct animals of uncertain zoological affinity) in Arabia and in northwestern Africa.

Marine fossils of the Devonian Period, which lasted from 345,000,000 to 395,000,000 years ago, are found in North Africa and in the Sahara. Traces also remain in parts of Guinea, Ghana, and Arabia, as well as in Gabon; they also occur in the Bokkeveld Series of South Africa. Fossils of vegetation that includes *Archaeosigillaria* (fossils of ancient club mosses) may be traced in formations of the earlier Devonian Period in the Sahara and in South Africa (Witteberg Series).

The Carboniferous Period

The Carboniferous Period lasted from 345,000,000 to 280,000,000 years ago. Evidence of marine life in the earlier part of this period comes from fossils found in North Africa, the central and western Sahara, and in Egypt. During the middle and later parts of this period, the Hercynian mountain-building episodes occurred. The Mauritanid mountain chain appeared, running from north to south along the edge of northwestern Africa from Morocco to Guinea. Ancient sediments were metamorphosed and pushed from west to east, locally covering the tabular Paleozoic formations. Elsewhere, great swellings or subsidences took place, continuing until the end of the Triassic Period (from 225,000,000 to 190,000,000 years ago). These movements were synclinal (*i.e.*, folded with the strata dipping inward toward a central axis) in the Tindouf area of western Algeria, dorsal (*i.e.*, forming a mountainous spine) at Reguibat in eastern Spanish Sahara, basinal (forming a basin) at Taoudeni in Mali, dorsal at Kinshasa, formerly Léopoldville (in Zaire [formerly the Democratic Republic of the Congo]), and synclinal in Ghana.

The Upper Carboniferous Period is represented throughout the Sahara by layers of fossilized vegetation, and sometimes—as in Morocco and Algeria—by seams of coal. Different phenomena may be observed, however, in sub-equatorial Africa, including the important Dwyka Tillite, which covers South Africa, Madagascar, a large part of the Congo Basin, and Gabon. At several places in South Africa, the Dwyka strata are covered by thin marine layers that demarcate the transition from the Carboniferous to the Permian periods and that form the beginning of the great Karroo System.

Marine fossils of the Permian period are visible in southern Tunisia (in the Jabal aṭ Ṭabaqah region), in Egypt (at the Abū Daraj Wadī), in the Arabian Peninsula (southwest of Riyadh), on the coasts of Tanzania, and in the Mozambique Channel.

Everywhere else, the traces of the Permian Period are of continental rather than marine origin and are in-

cluded in the Karroo System in South Africa. There, the Lower Permian strata are known as the Ecca Series and are divided into three groups—the Lower Ecca, containing almost 1,000 feet of clay-bearing schists; the Middle Ecca, containing 1,650 feet of sandstone, seams of coal, and fossilized vegetation; and the Upper Ecca, containing 660 feet of clay-bearing schists.

The Upper Permian is represented by the lower part of the Beaufort Series (which was also formed during the Lower Triassic period), which is almost 10,000 feet thick and which is famous for its amphibian and reptile fossils, a similar series of which are also found in the southern U.S.S.R. Other Permian formations, less rich in coal, are found in Zaire, Tanzania, Kenya, Uganda, Zambia, Rhodesia, Mozambique, and Madagascar.

The absence of primary marine formations throughout southern Africa should be emphasized; it is not yet known whether this absence is due to a hiatus in development or to erosion.

The Mesozoic Era. Comprising three periods—the Triassic, Jurassic, and Cretaceous—the Mesozoic Era is remarkable for the overflowing of ancient seas and for the emergence of massive land formations containing interesting fossil remains.

Marine formations. During the Triassic period, ancient seas left deposits in North Africa, the southern Sahara, Egypt, Arabia, and parts of Tanzania and northern Madagascar. Jurassic deposits extended to the Atlantic basins of Rio de Oro and Senegal. In the middle of the Jurassic Period a great transgression of the Indian Ocean extended over Somalia and Ethiopia as far as Eritrea. In the Cretaceous Period (from 136,000,000 to 65,000,000 years ago) this was followed by a series of transgressions including one on the coast of Equatorial Africa at the time that the present Atlantic and Indian Ocean coasts took shape; one that totally invaded the Sahara and Egypt as far as The Sudan; and one that covered the same places once more, as well as western Arabia and the east coast of Madagascar.

Continental formations. In north Equatorial Africa, including Arabia, the continental formations took the shape of an interlocked landmass. During the Triassic Period, the Saharan Zarzaïtine Series, containing dinosaur and reptilian fossil remains, was laid down. During the Jurassic Period, the Saharan Taouratine Series, containing fossils of vegetation and great reptilian fossils, was formed. In the Upper Karroo System of sub-Equatorial Africa, formed during the Lower Triassic Period, the Beaufort Series contains fossils of fish, amphibians, and reptiles. The final stages of the Triassic and Lower Jurassic periods were characterized by the folding of the Cape mountain chain, by the rise of the Karroo Basin, by fracturing, and by widespread upwelling of basaltic lavas through fissures, creating formations from 3,000 to 6,000 feet thick, such as the Drakensberg range in South Africa.

During the Jurassic and the Cretaceous periods, layers similar to the interlocked landmasses of the north were formed; they contained fossil vegetation, dinosaurs, and reptiles. During the Cretaceous Period, certain eruptions of a particular type occurred, resulting in the creation of the Kimberlite diamond pipes (a pipe being a cylindrical rock body, usually approximately vertical).

The Tertiary Period. The Tertiary Period of the Cenozoic Era (which has lasted from 65,000,000 years ago to the present) is remarkable for its great tectonic movements, which resulted in the Alpine mountain-building episode during which the North African mountains were uplifted. It is also remarkable for the creation of the Red Sea Rift Valley and for the volcanism and rifting that took place during its later stages.

Marine formations. The Paleocene Epoch (which occurred from 65,000,000 to 54,000,000 years ago) is important for its animal fossils, which include nummulites (a large kind of foraminifera, which are unicellular animals of macroscopic size), nautiloids (shelled cephalopods, which are mollusks with tentacles attached to their heads), and echinoids (sea urchins); all these are found in North and West Africa and in the Sahara. With the exception of the Sahara, nummulites of the Eocene Epoch (from 54,000,000 to 38,000,000 years ago) are found in the same places, as well as on the African coasts of the Indian Ocean, as also are lepidocyclines (foraminiferas) of the Oligocene Epoch (from 38,000,000 to 26,000,000 years ago) and the Miocene Epoch (from 26,000,000 to 7,000,000 years ago).

Continental formations. Several levels may sometimes be distinguished in the continental formations of the Tertiary Period. They include lower Eocene levels containing *Pseudoceratodes* (gastropod genus) and *Dryosaurus* (type of reptile), as well as upper Eocene and Oligocene levels containing silicated woods and fossilized fish, turtles, crocodiles, snakes, and mammals. In Egypt, the Oligocene deposits found in the Al Fayyūm area contain mammals, birds, turtles, and crocodiles. The levels of the lower Miocene, which are found on the banks of Lakes Rudolf and Victoria in East Africa, contain mastodon (a large elephant-like mammal) and *Proconsul africanus* (a large ape). Central Asian hipparions (three-toed ancestors of the horse), which simultaneously entered Africa and Europe during the upper Miocene Epoch, also left their fossilized remains in this region, as did the *Kenyapithecus* of Kenya (possibly the oldest hominid, or manlike primate) at about the same time.

Mammalian fossils of the lower Miocene

Tectonic movements. The first great foldings of the Atlas Tellien took place in the Oligocene Epoch. In the Miocene, North African flyschs (thick and extensive deposits, largely of sandstone) formed layers that, from the Rif to northern Tunisia, pushed from the north toward the south. The movements of the Tellian layers continued into the upper Miocene Epoch and the Pliocene Epoch (from 7,000,000 to 2,500,000 years ago).

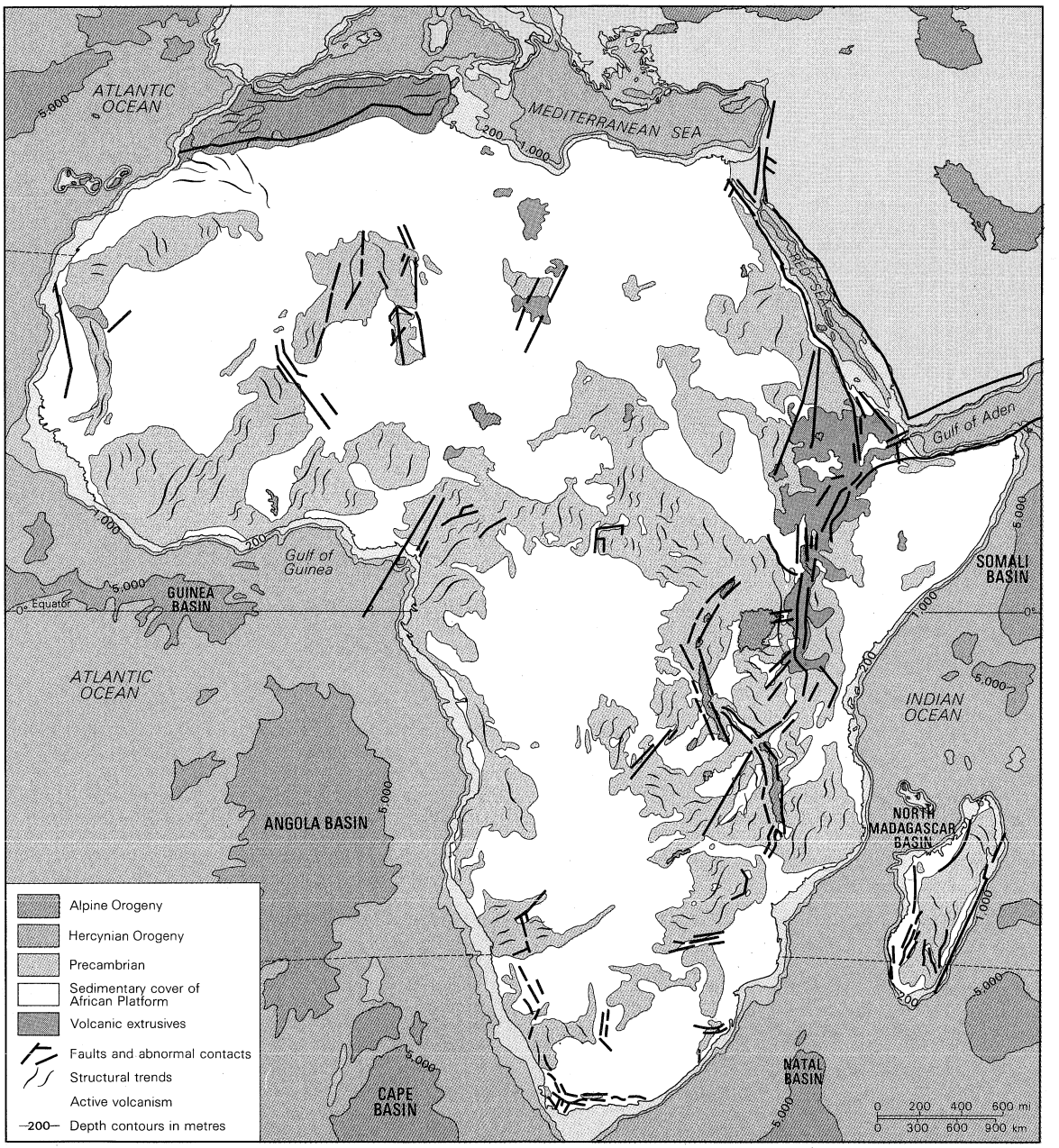
The High Plains area, further south, which as a whole experienced few tectonic movements, was bounded on the south by the northern Atlas, which intervened between it and the Atlas Saharien. Continental movements lifted the Aurès Massif (mountain mass) to a height of about 3,300 feet (1,000 metres) during the middle of the Miocene; the Aurès is bounded on the south by the northern Sahara structural line, which extends from Agadir in Morocco in the west to the Gulf of Gabes in Tunisia in the east, dividing the African Shield from the folded Mediterranean or Alpine zone.

Formation of the Red Sea. Tectonic movements in the region of the Arabo-Nubian Shield that took place at the end of the Oligocene and the beginning of the Miocene epochs almost separated Arabia from Africa. A trough sank down, and the Mediterranean Sea swept into the rift valley, forming a gulf that extended to Yemen. The gulf was only prevented from joining the Indian Ocean by an isthmus stretching from Djibouti to Aden.

At the end of the Miocene Epoch, the Isthmus of Suez was formed, and the gulf became a saline lake at the bottom of which thick evaporites (sediments formed as a result of evaporation) were laid down. The isthmus permitted Asian animal life to pass into Africa during part of the Pliocene epoch. The sinking of the Djibouti–Aden isthmus, also during the Pliocene Epoch, allowed the Indian Ocean to flow into the Red Sea as far as the Isthmus of Suez.

Volcanism and rifting. Tectonic movements in the Miocene and Pliocene scored the continent with a network of northeast to southwest and northwest to southeast faults. Volcanic eruptions and basaltic upwellings accompanied fracturing in the Ahaggar area of southern Algeria, in the Tibesti area of Libya and Chad, in Ethiopia, throughout East Africa, and in Cameroon, as well as in the islands of the Gulf of Guinea (Fernando Po and São Tomé and Príncipe).

The Quaternary Period. The Quaternary Period is divided into the Pleistocene Epoch (from 2,500,000 to 10,000 years ago) and the Recent Epoch (up to 10,000 years ago). It represented a phase of continuing volcanic activity that caused the basement rocks of the Ahaggar and Tibesti mountains of the Sahara to rise. The activity manifested itself in eruptions and in the deepening of the Saharan valleys—sometimes to a depth of about 660 feet—as well as in the welling up of basalt. Though



Structural features of Africa.
Adapted from *Carte tectonique de l'Afrique* (1966); UNESCO/ASGA

the marine deposits of the Quaternary Period as a whole are of no particular interest, continental deposits of this period bear witness to a human chronology and to human activities that parallel those of Europe.

During the cold humid periods called pluvials, which correspond to the glacial phases of the Northern Hemisphere, the glaciers that covered the high mountains of East Africa were from 3,000 to 5,000 feet thicker than they are today. Elsewhere, the desert zones of the Sahara and of the Kalahari were alternately subjected first to humid and then to dry and arid phases that expanded the desert surface at the expense of the adjacent forested zones.

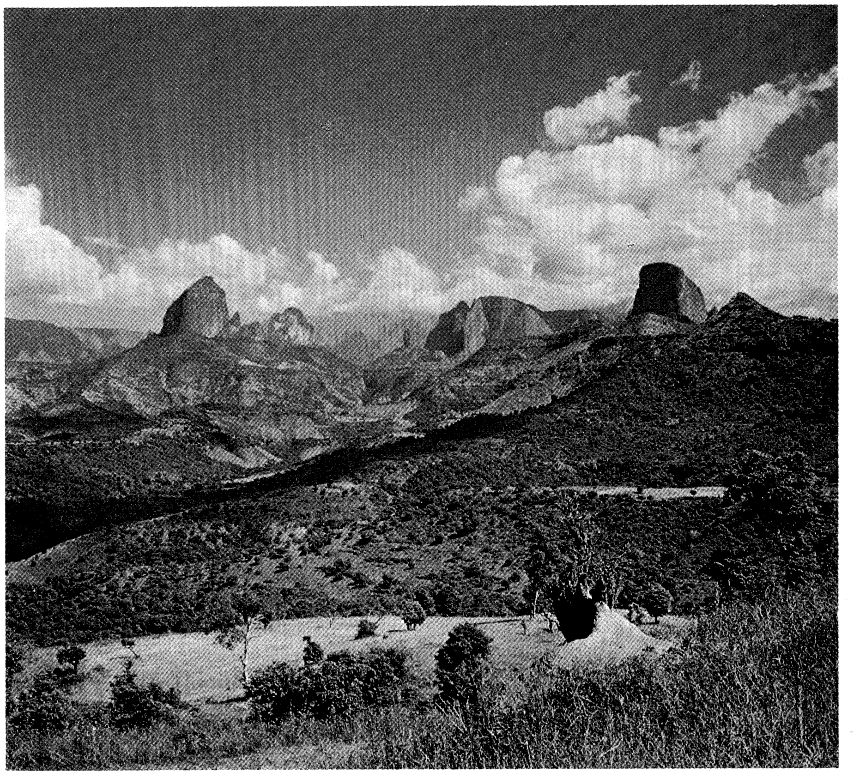
The oldest levels at which hominids are found, known as the Villafranchian–Kaguerian Series and recognized in Ethiopia and Kenya, date back some 3,000,000 to 4,000,000 years and are characterized by ancient forms of animal life, including in particular some ancient hominids.

Primitive hominids

These primitive hominids, called Australopithecines, may be classified in two main groups—*Australopithecus africanus* and *Australopithecus robustus*. Both groups left traces of stone and bone implements.

The Kaguerian–Kamasian Interpluvial levels, which date back to 500,000 years ago, contain the Pithecanthropoids (a new primitive kind of men) of Olduvai Gorge (Tanzania), Ternifine (Algeria), and Chad. The Kamasian, or Second, Pluvial of the middle Pleistocene Epoch corresponds to the Mindel in Europe and also contains *Dinotherium* (a large mammal with tusks on its lower jaw). A dry but not a desert climate is implied by the Kamasian–Kanjerian Interpluvial levels at Olduvai Gorge. The Kanjerian, or Third, Pluvial, which occurred during the middle Pleistocene Epoch and which corresponds to the Riss Pluvial in Europe, bears traces of such animals as the elephant, giraffe, and hippopotamus. An arid phase, which greatly reduced the area of forest, is revealed in the Kanjerian–Gamblian Interpluvial levels, which were laid down between about 55,000 and 60,000 BC. This period corresponds to an important tectonic phase marked by undulations in North Africa and activity along all the faults, in particular those in East Africa; it was at this time that East Africa assumed its present topographical aspect.

During the Gamblian, or Fourth, Pluvial, which occurred from about 15,000 to 30,000 BC, three humid



Simen Massif, which rises above the Ethiopian Plateau in the western region of Ethiopia, near Adi Arkay.
Archives Almasy

phases are separated by drier intervals. During these phases the dimensions of Lake Chad and of the glaciers of Mt. Kenya and Mt. Kilimanjaro diminished rapidly. The postpluvial phase that followed, equivalent to the postglacial phase of the Northern Hemisphere, was marked by a succession of alternately dry and humid stages, in which traces of Mesolithic (a transitional period of the Stone Age) and Neolithic (New Stone Age) civilizations occur, and by the desertization of the Sahara and the Kalahari, which began about 3000 BC.

PRESENT GEOLOGICAL STRUCTURES AND RELIEF

Africa represents a magnificent example of the relation of geological structure to relief. In the north, the chains of the Atlas Mountains, extending from west to east, evoke the orogenesis of the Alpine episode and are separated from the Sahara basement rock by the great interrupted line of the southern Atlas. Egypt, by contrast, exposes the magnitude of the Cretaceous and Tertiary marine transgressions. Between the southern Atlas and the Ahaggar Massif, the lower Sahara Basin, up to 30,000 feet in depth, contains a wide, water-retaining fossiliferous substratum and is partly covered by the great sand dunes of the Western and Eastern Ergs (Grand Ergs Oriental and Occidental). Encircling the Ahaggar, the Tassili plateaus are bounded by cliffs that form natural receptacles for Paleozoic deposits.

The West African land mass

The West African landmass comprises the Taoudeni syncline, the average diameter of which is more than 620 miles (1,000 kilometres). The syncline is situated between the Ahaggar mountains in the east and the Mauritania–Sénégal Basin in the west and between Reguibat in the north and Kinshasa in the south. The east-to-west crystalline dorsal formations running from Cameroon through the Ubangi (Oubangui) region to The Sudan completely separate northern from sub-Equatorial Africa. Karroo and continental deposits fill the Congo Basin, which is surrounded by Precambrian rock formations dating from all periods of that era.

In southern Africa, the Kalahari Basin and the plateaus of the Karroo provide 23,000 feet of undifferentiated Carboniferous, Permian, and Triassic continental deposits.

East Africa is isolated from the rest of the continent by huge tectonic rift valleys that have been volcanized (subjected to volcanic heat) and that extended from Syria southward through the Red Sea and the Rift Valley lakes to Mozambique. The Mozambique Channel, which dates from the Permian Period, bounds Madagascar on the west, while on its eastern periphery a great fault dates from the Upper Cretaceous Period.

Along the coasts of Africa a certain number of basins disappear beneath the sea—off Spanish Sahara; Senegal; the Ivory Coast and Ghana; Dahomey, Nigeria, and Cameroon; Gabon; the Congo; Angola; Mozambique; Tanzania; and Somalia.

Africa's geological structure as a whole may thus be said to express its geomorphology. (R.L.C.F.)

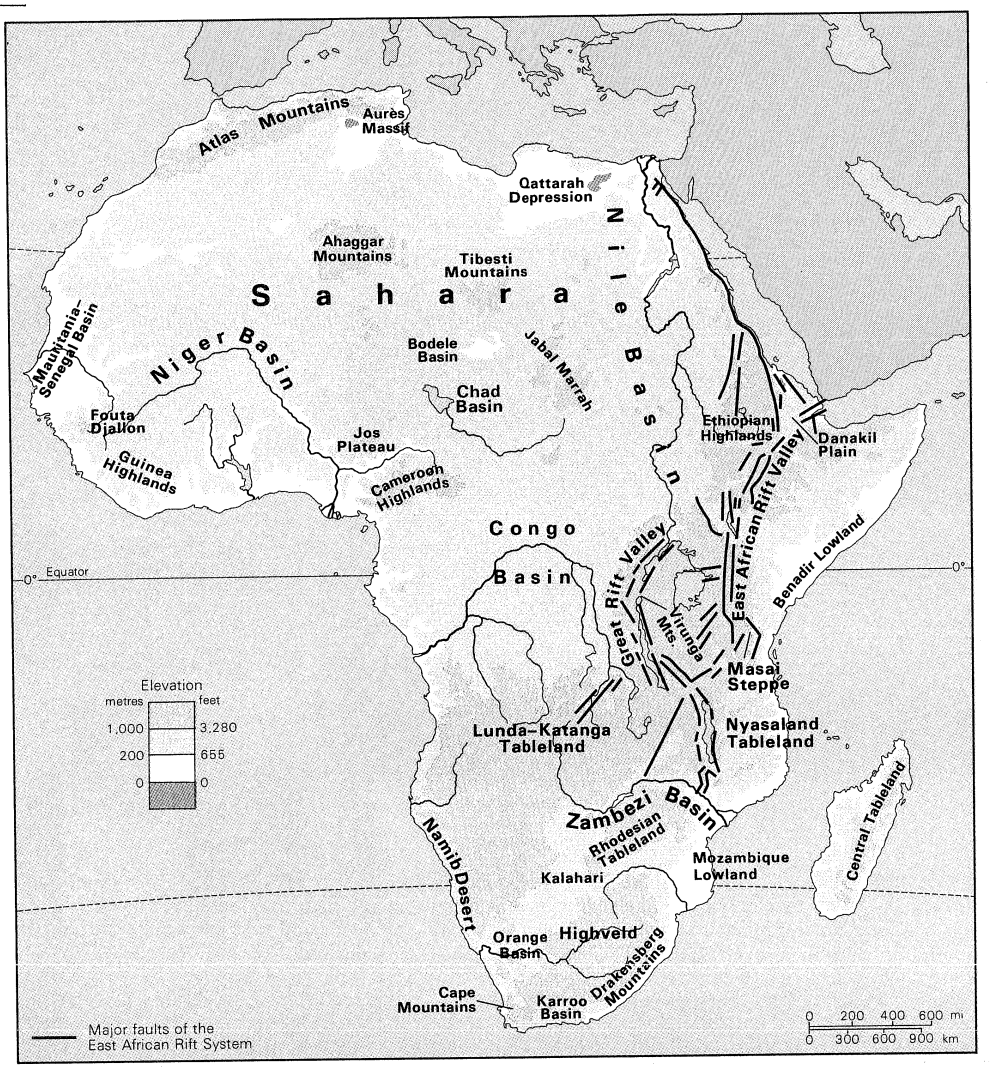
II. Physical geography

RELIEF

The physiography of Africa is largely a reflection of the geological history and geology described in the previous section. The continent, largely composed of a vast rigid block of ancient rocks, has geologically young mountains at its extremities in the highlands of the Atlas Mountains in the north and Cape ranges in the south. Between them is a series of plateau surfaces, with huge areas that are level or slightly undulating, above which stand occasional harder and more resistant rock masses. Surrounding these plateaus is a zone of plateau slopes below which are narrow coastal belts widening along the Mediterranean coast, the coastlands of Tanzania and Mozambique, a narrow belt between the Niger and Cunene (Kunene) rivers, and an area northward of the Gambia and Sénégal rivers.

Mt. Kilimanjaro (19,340 feet, or 5,895 metres) is the highest point on the continent; the lowest is Lake Assal (509 feet, or 155 metres below sea level) in Afars and Issas. In proportion to its size, Africa has fewer high mountains and fewer lowland plains than any other continent. The limited areas above 8,000 feet are either volcanic peaks or resistant massifs. All the land below 500 feet occurs within 500 miles of the coast, except for two small basins in the Sahara.

The higher areas of the south and east are in marked



Physiographic regions of Africa.

contrast to the considerably lower elevation of the western and northern parts of the continents. South of a line drawn from near the mouth of the Congo River to the Gulf of Aden, most of the land lies 1,000 feet or more above sea level, and much of it exceeds 3,000 and even 4,000 feet. North of the line there is relatively little land above 3,000 feet, most of the area being between 500 and 1,000 feet above sea level; there are also broad coastal lowlands, except in the northwest and in the east beyond the Nile.

The highest extensive areas are in Ethiopia, parts of which exceed 15,000 feet. Southward, the East African Plateau is highest in Kenya, where it is often 8,000 feet or more above sea level; there are occasional volcanic peaks that are much higher, such as Kilimanjaro, Mt. Kenya (17,058 feet [5,199 metres]), Meru (14,979 feet [4,566 metres]), and Elgon (14,178 feet [4,321 metres]). The Ruwenzori Range, or the Mountains of the Moon, which reach their highest elevation at Margherita Peak (16,763 feet [5,109 metres]), on the borders of Zaire and Uganda, is not volcanic in origin. From East Africa, the plateau extends southward, often with a well-defined, though not continuous, escarpment particularly noticeable in the Drakensberg of southern Africa, where Thabana Ntlenyana, or Thabantshonyana, is 11,425 feet and Mont aux Sources is 10,822 feet. There, the plateau edges are especially marked, because the rock formations are hard and horizontal, whereas in Ethiopia they are conspicuous because of recent faulting. Where the rocks are softer and less resistant, the escarpment is not so pronounced and so forms less of a barrier to climatic influences and to human movement.

Northward and westward of the plateau area of the southern parts of the continent there is a general descent to the lower areas of the basins of the Congo, Niger, and Nile rivers. The only large areas above 3,000 feet are in the folded ranges of the Atlas Mountains and in the central Sahara, where resistant granites form the massifs of Ahaggar and Tibesti. The interior uplands of West Africa and of Cameroon consist of ancient crystalline rocks, reaching considerable heights only in the Fouta Djallon plateau in Guinea; in the Guinea Highlands, which also extend over the borders of Sierra Leone and Liberia; in the Jos Plateau in Nigeria; in the Adamawa region of Nigeria and Cameroon; and in the Cameroon Highlands. There are extensive low-lying areas near the coast and in the basins of the Sénégal, Gambia, Volta, and Niger–Benue rivers. The high areas of Darfur in The Sudan (over 10,000 feet) and of Mt. Cameroon (13,353 feet [4,070 metres]) are volcanic in origin and are evidence of the same tensions that have resulted in rifting and volcanism in East Africa.

Volcanic activity

The Great Rift Valley, or East African Rift Valley, constitutes the most striking and distinctive relief feature of the whole continent. Associated with its formation was the volcanic activity responsible for most of the higher peaks of East Africa, including Kilimanjaro, which is always snowcapped, despite its nearness to the Equator. Seismic and volcanic disturbances are still recorded in the western portions of the rift valley system. In the Virunga Mountains, northeast of Lake Kivu, there are periodic outbursts (about every 10 or 12 years) that have created a series of lava flows. One of these volcanoes dammed the rift valley and converted a large area, formerly drained by a tributary of the Nile, into Lake Kivu.



Ouzirhlmt Valley in the Haut Atlas, Morocco.
Victor Englebert—Photo Researchers

The rift valley extends for about 4,000 miles, its course being clearly marked out by many of the lakes of East Africa as well as by the adjacent volcanic peaks. From the Gulf of Aqaba it can be traced southward along the Red Sea and into the Ethiopian Highlands to Lakes Rudolf, Naivasha, and Magadi in Kenya. Further south, through Tanzania, the line is less obvious. The walls that constitute the eastern rim have been more easily eroded, while the lakes are generally smaller and not in line, and some of them are only waterless salt beds. The largest of these lakes are Natron and Manyara, with Eyasi in a side branch of the main rift. The edges are obvious enough to the south in Malawi, where a huge crusted block collapsed along the parallel faults that constitute the steeply rising slopes of Lake Nyasa (Malawi). The lake is 360 miles long but never more than 50 miles wide; it has a maximum depth of 2,316 feet (706 metres). The rift then follows the line of the Shire Valley to reach the coast of the Indian Ocean near Beira, Mozambique.

The western branch, or Western Rift Valley, extends from the northern end of Lake Nyasa in a great arc, taking in Lakes Rukwa, Tanganyika (after Lake Baikal in Siberia the deepest lake in the world, with a sounding of 5,315 feet [1,620 metres]), Kivu, Edward, and Albert. Subsidiary branches of this valley include the basins in which lie lakes Mweru and Upemba.

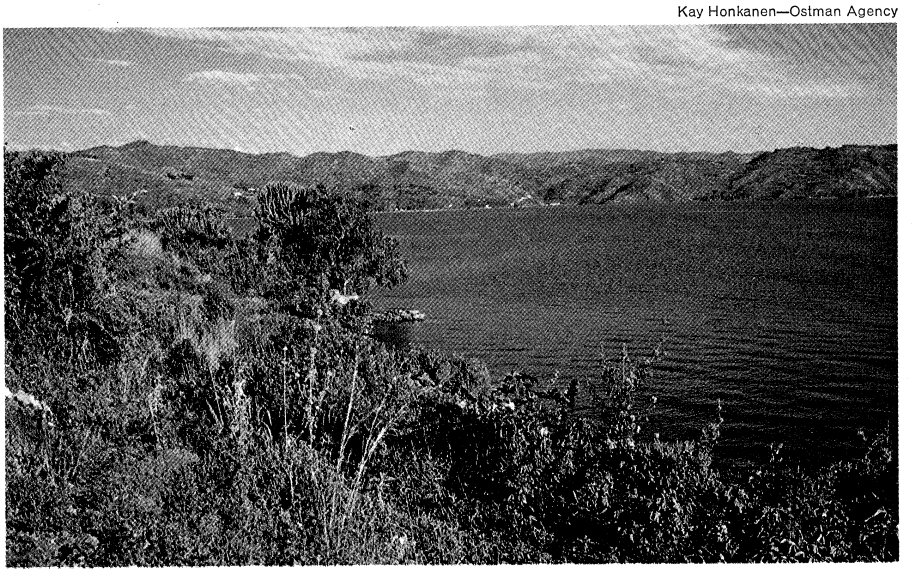
Most of the rift-valley lakes lie well below the general level of the plateau, ranging from 1,300 to 3,000 feet above sea level. They are generally very deep and fjord-like; some have floors below sea level, even though their surfaces are hundreds of feet above sea level.

In complete contrast is Lake Victoria, the largest of all African lakes, which occupies a shallow depression on a plateau 3,720 feet (1,134 metres) above sea level between the major branches of the rift valley. Its greatest depth is only 270 feet (82 metres), but its area (26,828 square miles) makes it the third largest of the world's great lakes, after the Caspian Sea and Lake Superior. (Ro.W.St.)

CLIMATE

Factors influencing the African climate. A number of factors influence the climate of the African continent.

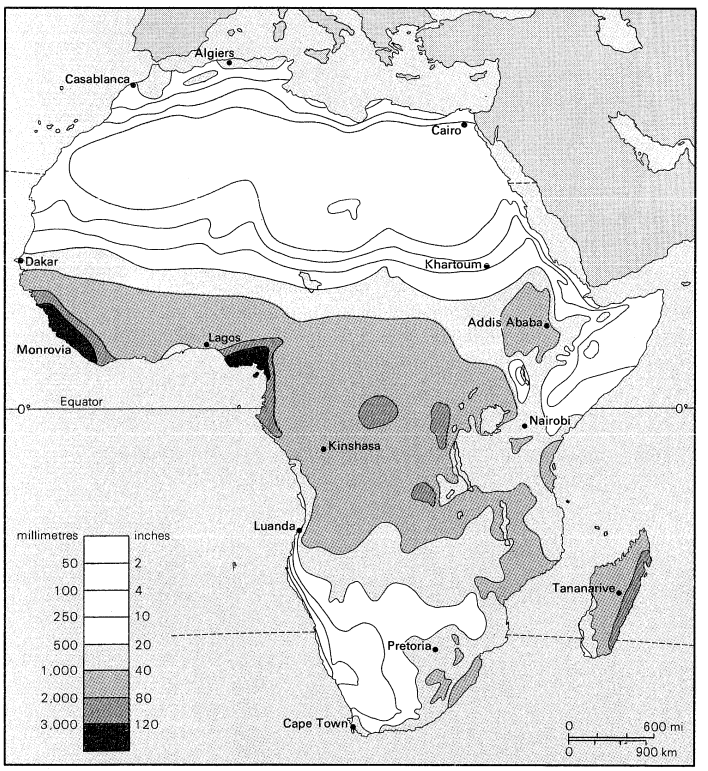
Kay Honkanen—Ostman Agency



Lake Kivu in eastern Africa between Rwanda and Zaire. The western escarpment of the Great Rift Valley rises on the far side of the lake.

Firstly, practically the whole of the continent—which extends from latitude 35° S to about latitude 37° N—lies within the tropical zone. Secondly, the near bisection of the continent by the Equator results in a largely symmetrical arrangement of climatic zones on either side. This symmetry is, however, imperfect because of a third factor—the great east-to-west extent of the continent north of the Equator, which contrasts with its narrow width to the south. In consequence, the influence of the sea extends further inland in southern Africa. Moreover, true low pressure zones develop in the heart of northern Africa, while in southern Africa the belt of high pressure on land weakens during the time of high sun (*i.e.*, at the season when the sun is overhead—in December and January). A fourth influence is constituted by the cool ocean currents, which chill the winds that blow over them and thereby influence the climate of the neighbouring shores. Fifthly, because of the extensive plateau surfaces of the continent and the absence of high and long mountain ranges comparable to, for example, the Andes of South America or the Himalayas in Asia, climatic zones in Africa tend to shade into one another, rather than to change abruptly from place to place. Finally, the high mountains have climatic zones of their own that vary with altitude.

While these factors help to account for broad climatic patterns, there are, nevertheless, numerous local variations from place to place within the same climatic zone. Towns, for example, have their own climates, although in Africa the relevant data are not yet available to permit a detailed study of urban climates.



Average annual precipitation for Africa.

The air masses

The most important differentiating climatic element is rainfall; this, together with other climatic elements, depends upon the characteristics and incidence of air masses. The air masses of relevance to the African climate may be broadly classified as tropical maritime, equatorial maritime, tropical continental, polar maritime, and polar continental. Of these, the least important are the polar continental air mass, which may occasionally bring intense cold to lower Egypt in December and January, and the polar maritime air masses, which are associated with rain-bringing depressions over the northern and southern extremities of the continent during the winter. With the exception of these, the continent is affected, on the one hand, by the tropical continental air mass to the north and, on the other, by the tropical maritime and equatorial maritime air masses to the south. These northern and southern air masses meet at the Intertropical Convergence Zone. The hot, dry tropical continental air mass, present the year around in the upper levels of the atmosphere, descends to the ground only at the convergence zone. Less hot than the tropical continental are the tropical maritime and equatorial maritime air masses, which originate from the Indian and South Atlantic Oceans, respectively; they differ from each other only in that the equatorial maritime air mass is unstable and brings rain, while the tropical maritime air mass, when fully developed, is stable and will not normally bring rain unless it is forced to rise by a high mountain.

In July, the Intertropical Convergence Zone moves northward toward the area of low pressure over the Sahara; here, the maritime and tropical continental air masses converge, with the maritime air masses swinging inland from the sea. There is no rainfall on the northern side of the Convergence Zone, since the region is completely under the dry tropical continental air mass. Conversely, on the southern side rainfall increases with distance away from the Convergence Zone as the maritime air masses become thicker. South of the Equator, at yet greater distance from the Convergence Zone, the maritime air masses are so thick and stable that they bring hardly any rainfall, except over some of the East African highlands. Only the southern tip of South Africa receives rainfall at this time, from winter cyclones.

During the period of low sun in the Northern Hemisphere (*i.e.*, from December to January, when the sun has moved to its southern limit), the situation described above is reversed. The Convergence Zone moves southward, dipping into southern Africa, while its western section hovers along the West African coast. At this season the whole of northern Africa (except the Atlas region) is under the dry tropical continental air mass, while southern Africa receives rainfall except in the Cape region and on the southwestern coast, where the maritime air mass remains stable offshore over the cool Benguela Current.

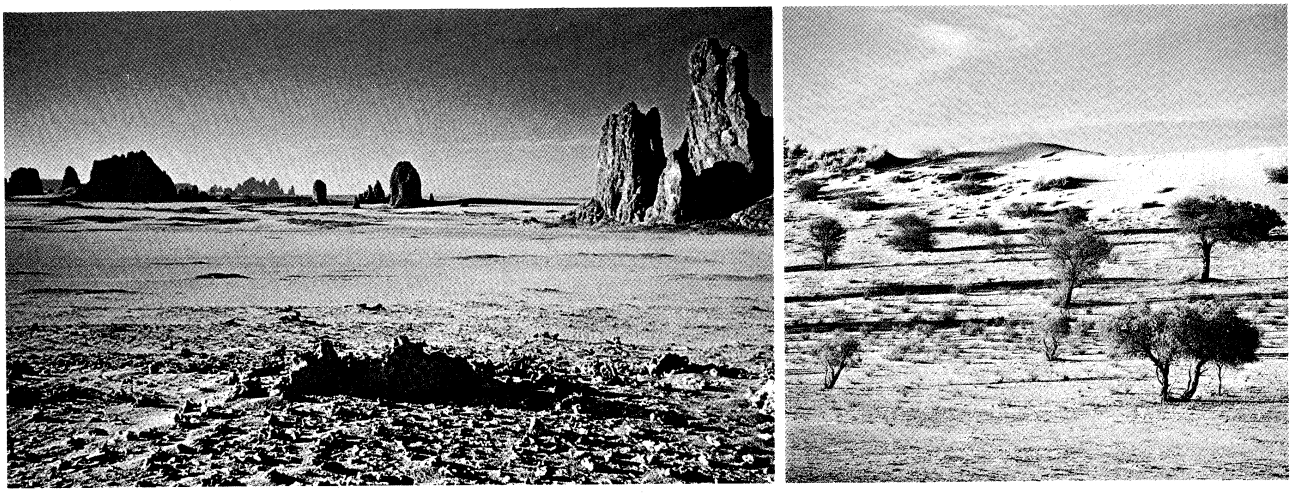
Climatic regions. When considered in detail, the movement of air masses and their effects provide the basis for a division of the continent into eight climatic regions. These are the hot desert, semi-arid, tropical wet-and-dry, equatorial (tropical wet), Mediterranean, humid subtropical marine, warm temperate upland, and mountain regions.

The eight climatic regions

The hot desert region comprises the Sahara and Kalahari, which are always under the influence of dry tropical continental air masses, and the northern Kenya–Somali desert, the aridity of which is principally caused by the stable nature of the maritime air masses that pass over it throughout the year. The stability of the air masses is induced by their passing over the cool body of water offshore. In addition to aridity, the desert climate is characterized by high mean monthly temperatures; the diurnal (daily) temperature range is, however, greater than the annual range of the mean monthly temperature.

Regions with a semi-arid climate type fringe the desert areas and include the greater part of the land south of the Zambezi River. They differ from true desert regions in being just within reach of the Intertropical Convergence Zone and therefore receiving more rainfall. Temperatures are about the same as in the desert regions.

The tropical wet-and-dry region is often called the savanna climatic region; this implies, incorrectly, that all areas with savanna vegetation experience this type of climate. This region covers a little less than 50 percent of the total surface area of the continent, extending toward the Equator from the semi-arid areas. The great distinguishing feature of this climatic region is the seasonal character of its rainfall. During the period of high sun, the maritime air masses produce up to six months of rainfall—the length of the rainy season depending on nearness to the Equator. The rest of the year is dry. In a few places, for example, on the coast of Mauritania and Senegal, there is also a little rainfall in the period of low sun. As in the desert and semi-arid climatic zones, mean



African Deserts.
(Left) The Sahara near the Ahaggar Mountains in southern Algeria. (Right) The Kalahari in southwestern Botswana.
(Left) Emil Schulthess—Black Star, (right) Gerald Cubitt

monthly temperatures show less variation than daily temperatures. In West Africa, lower than normal temperatures in December and January characterize what is known as the harmattan season (the harmattan is a warm, dry northeasterly or easterly wind) and are caused by the relative coolness of the tropical continental air mass.

Regions with the equatorial, or tropical wet, type of climate, or variants thereof, are the wettest in Africa. There are two peaks of rainfall from the equatorial maritime air mass during the year; these occur shortly before the sun reaches the tropics of Cancer and Capricorn and shortly after it recrosses the Equator. The first rainy season is often more important than the second, although there are years or places in which the second rainy season is the more important. Since areas with an equatorial climate are constantly covered by warm maritime air masses, variations in their monthly and daily temperatures are less pronounced than in the tropical wet-and-dry region.

Marked variations in the rhythm of equatorial climate sometimes occur; for example, the rainfall may be monsoonal (characterized by rain-bearing winds), and the second rainy season may be all but nonexistent. But the most notable anomaly is that to be observed on the West African coast from Cape Three Points (Ghana) eastward to Dahomey, where, although the two-peak regime prevails, the total annual rainfall is less than 40 inches. Among the explanations that have been suggested are that the presence of a cold body of water offshore chills the lower layers of the maritime air mass and makes it stable; that the body of cold air that forms offshore diverts the incoming air streams to the west and east of the anomalously dry area; that there is a strong tendency for the winds to blow parallel to the shore during the rainy seasons; and that the absence of high mountains deprives the region of orographic (mountain) rainfall.

The Mediterranean climatic zone

The northern and southern extremities of the continent experience a dry summer subtropical, or Mediterranean, type of climate. Rain falls only in winter (*i.e.*, December–January in North Africa, June–July in South Africa), although in some localities it may fall in autumn (September in North Africa, April in South Africa). Mean monthly temperatures are lower than in tropical climates, dropping to about 50° F (10° C) in winter, while summer (June–July in North Africa, and December–January in South Africa) temperatures may sometimes exceed those of tropical climates. Clear blue skies are characteristic.

The humid subtropical marine climatic region is restricted to the southeastern coast of South Africa. This region is characterized by rainfall throughout the year but with heavier rainfall occurring in summer. South of Natal, the winter rainfall is more pronounced, and the temperatures are a little lower than they are to the north. Thus, at Port Elizabeth, South Africa, there are six months when temperatures are below 62° F (17° C), while at Durban, South Africa, mean monthly temperatures do not fall below 64° F (18° C).

The warm temperate upland climatic region is found on the Highveld of southern Africa. Its rainfall regime is similar to that of the tropical wet-and-dry climate, but temperatures are greatly modified by the altitude; frost, for example, occasionally occurs in Lesotho. Toward the coast, the climate shows maritime characteristics, and there is a tendency toward winter rainfall.

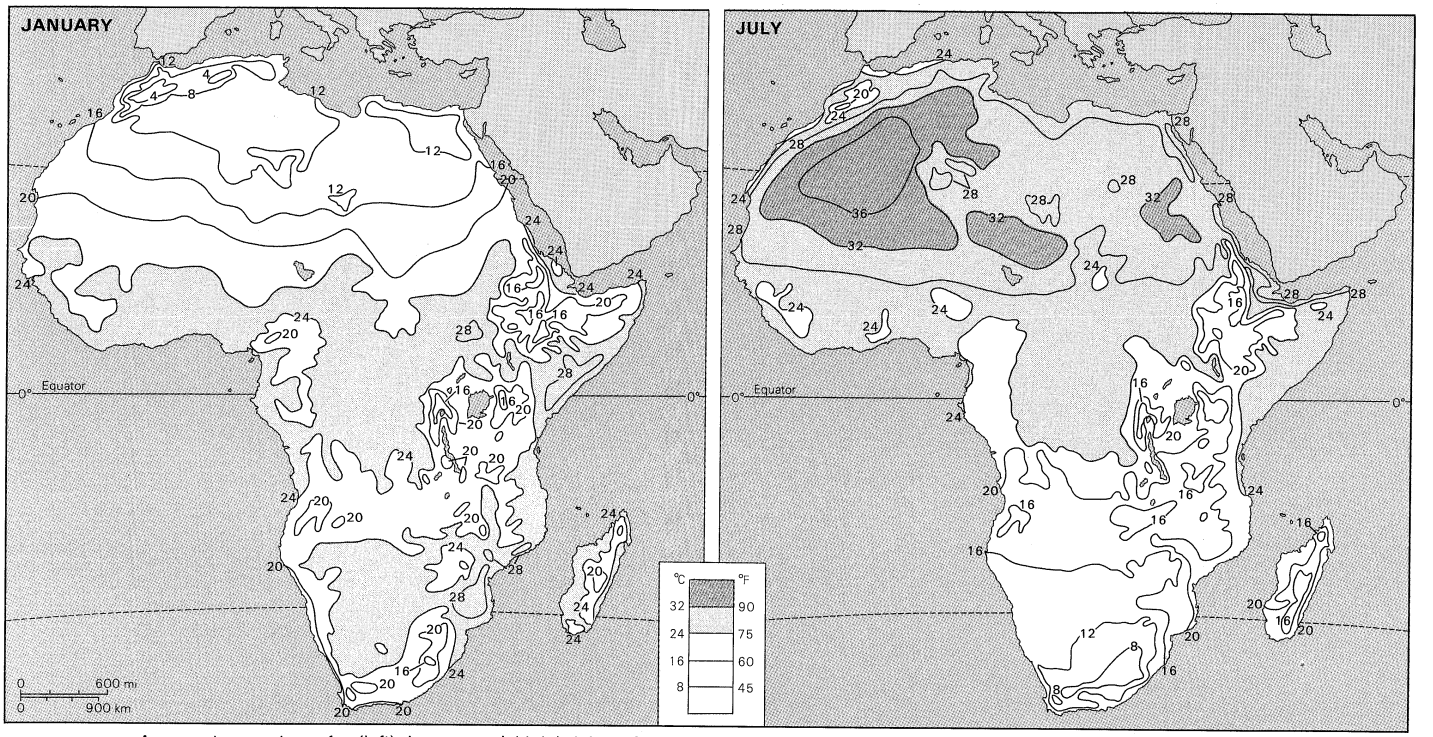
The mountain climatic region includes the high mountain areas of Ethiopia and the lake region of East Africa. In some respects the climate is similar to the warm temperate uplands climate, except that temperatures are even lower and snow occurs on the tops of the highest peaks, such as Kilimanjaro. The rainfall regime is similar to that of the adjacent lowland areas.

DRAINAGE

The uplifting and warping of the surface of the continent that occurred during the Pliocene and Pleistocene epochs produced a number of structural basins; these are now either individually occupied by, or are linked up with, drainage systems. With the exception of the Chad Basin, all the major drainage basins have outlets to the sea. In addition, minor drainage basins, similar to that of Lake Chad, are situated in the East African rift valleys. Some, again like Lake Chad, constitute the focus of centripetal drainage (*i.e.*, drainage directed toward the centre), while others are linked to river systems. Although the East African lakes are climatically and economically significant, relatively little is known of their hydrological characteristics.

Climate, geology, and the history of land movements have imparted certain common characteristics to African rivers. Spatial variations in the incidence and amount of rainfall are reflected in their hydrological regimes, so that, for example, in areas with one rainfall season experiencing pronounced drought throughout the rest of the year, the rivers flood in the rainy season and shrink in the dry.

Whatever their regimes, all the important African rivers are interrupted by rapids, cataracts, and waterfalls. This is explained by several factors. One of the most important of these were the land movements that caused ridges to be formed across the courses of the rivers. Waterfalls are often to be found where the rivers are still engaged in cutting downward as they flow across these ridges; Cabora-Bassa (falls) on the Zambezi and the Augrabies Falls on the Orange River are examples. Another factor contributing to the creation of rapids or falls is the incidence of bands of rock that have proved resistant to the erosive effect of the rivers' flow. (Tropical riv-



Average temperatures for (left) January and (right) July in ° C for Africa.

ers do not generally carry large quantities of stone or rock; instead, they tend to carry loads of fine silt, produced by chemical weathering.)

Although the Nile, the Zambezi, and the Niger rivers have large deltas, their size does not compare with that of, for example, the Ganges–Brahmaputra deltas. In Africa, their generally poor development is due mainly to the restricted extent of the coastal plain, together with the relatively narrow continental shelf, which provides neither sufficient room nor shallow enough water for the deposition of delta-forming material. The great speed with which most of the rivers flow into the sea is another factor inhibiting delta formation.

Major drainage basins

The major drainage basins of the African continent are those of the Nile, the Niger, the Congo, the Zambezi, and the Orange rivers, and of Lake Chad.

The Nile Basin. Two theories exist concerning the development of the Nile, which, it appears, originally consisted of two sections. The first theory is that the Lower Nile had its source at about latitude 20° N, whence it flowed directly into the sea, while the Upper Nile, issuing from Lake Victoria, flowed into an inland lake that covered the Sudd region (as-Sudd) in what is now The Sudan. The lake became filled with water, which then spilled over at its northern end and flowed into what is now the Upper Nile. According to the second theory, the upper section originally flowed into a vast lake between Jabal as-Silsilah (near Luxor, in Egypt) and what is now Aswān; this was tapped by the lower section of the Nile after the so-called Sebile erosion (which takes its name from the fact that the break-through by the lower Nile was identified at Sebile).

The Nile (*q.v.*), which is about 4,132 miles (6,650 kilometres) long, is the longest river in the world. From Lake Victoria it flows, as the Victoria Nile, into Lake Albert, from which it emerges as the Albert Nile. Further north it is known as the Bahr el-Jebel. Thereafter, having received several tributaries, it becomes the White Nile, and finally the Nile, emptying at last into the Mediterranean. Its major tributaries are the Baḥr al-Ghazāl (the largest left-bank tributary), and the Sobat, Blue Nile, and ʿAṭbarah (on the right bank). Because of the numerous rapids and waterfalls, the Nile descends fairly rapidly from source to mouth, as do its major right-bank tributaries. This is especially true of the Blue Nile, which, issuing from Lake Tana in the Ethiopian Highlands at a height of about 6,000 feet, flows for most of its length through a steep gorge. Swamps also interrupt the river's course. Of these the largest is the Sudd, a vast area of floating swamp reeds, mostly papyrus.

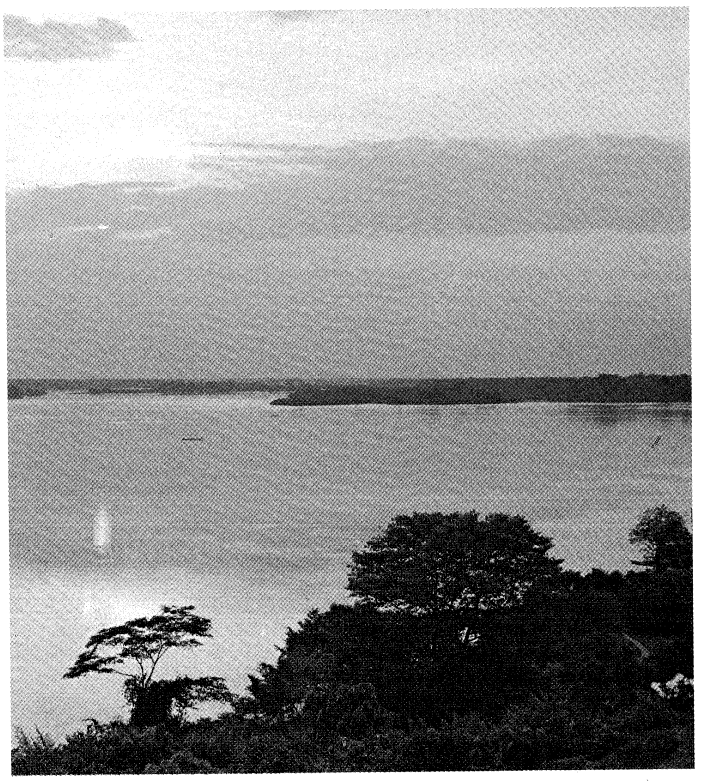
The river's regime is now controlled by a series of dams situated on the Nile itself or on one or other of its various tributaries; of these, the largest is the Aswān High Dam on the main Nile.

The Niger Basin. The Niger Basin is the largest West African river basin. The Niger River (*q.v.*), which rises in the mountains of the Republic of Guinea, enters the sea through its delta in Nigeria; it is about 2,600 miles (4,200 kilometres) in length. Rapids interrupt its course at several points, for example, below Bamako and at Bussa.

The Niger receives its largest tributary, the Benue, which flows in from its left bank, in Nigeria. The valleys of both the Niger downstream from Taoussa and of the Benue appear to be faulted troughs dating from the Lower Cretaceous Period (from about 136,000,000 to 100,000,000 years ago). Originally, the Middle Niger was separate from the Upper Niger, which flowed into an inland lake, the remnants of which now form the inland Niger Delta. The Middle Niger flowed southeastward to the sea; its valley eroded toward its headwaters, eventually tapping the inland lake and linking the Middle with the Upper Niger.

The Congo Basin. With a total area of about 1,335,000 square miles (3,457,000 square kilometres), the Congo Basin consists of a vast shallow depression that rises by a series of giant steps to an almost circular rim of highlands through which the river has cut a narrow exit into the Atlantic Ocean. The present exit is comparatively recent, the previous exit being to the north of the present one.

The Congo River (*q.v.*) is 2,900 miles (4,700 kilometres) in length. Its many waterfalls and rapids cause its valley, like that of the Nile, quickly to lose height. The river's course is often constricted by gorges. Of the falls, the best known are the Stanley Falls at Kisangani, where the river swings through an arc to flow westward; in fact, the Stanley Falls are no more than a series of unevenly spaced rapids at no great height, extending along a 60-mile stretch of the river. Downstream from Kisangani, the Congo is joined first by the Oubangui (Ubangi) from the right and then by the Kasai—which rivals the Ou-



Congo River, in Zaire.
Emil Schulthess—Black Star

bangui in the size of its drainage basin—from the left. Below its confluence with the Kasai, the main river cuts through the Crystal Mountains (Monts de Cristal) in a deep gorge, which at one point expands into Stanley Pool (Malebo Pool in Zaire), a shallow lake 22 miles long and 14 miles wide. The Congo enters the sea through a swampy estuary about six miles wide at its mouth.

The Zambezi Basin. The Zambezi River is about 2,200 miles (3,540 kilometres) in length; it occupies a basin with an approximate area of 514,000 square miles. Originally, there were two rivers, corresponding to the upper and lower courses of the present river; the valley of the lower section eroded toward the headwaters until it captured the waters of the upper section. Although there are stretches of the river where the gradient is very gentle—about three inches to the mile—the valley as a whole has a fairly steep gradient. There are numerous waterfalls, the most spectacular of which is the Victoria Falls (*q.v.*). After these falls, the river winds through a number of deep gorges cut out of basalt and, after flowing through a broad valley, enters Kariba Gorge, which is more than 16 miles in length and is cut through paragneiss (a gneiss, or coarse-grained rock, in which bands rich in granular minerals alternate with bands containing schistose minerals, formed out of sedimentary rock). The Kafue and the Luangwa, the two main tributaries, which both flow through gorges, join the Zambezi on its left bank downstream from Kariba. At the mouth of the main river is a delta about 37 miles wide.

The Orange Basin. The Orange River (*q.v.*) is the longest in South Africa. Flowing across almost the entire width of the country, it makes its way from the highlands in the east through the Kalahari depression in the west to empty into the South Atlantic Ocean. Its major tributary, the Vaal, is one of its northern headwaters; the two rivers together have a combined length of about 1,300 miles (2,090 kilometres). Together with other major rivers on the continent, the Orange–Vaal river system shares the characteristic of flowing over steep gradients for numerous stretches of its course. The largest drop (about 400 feet) occurs at the Augrabies Falls.

The Chad Basin. The Chad Basin constitutes the largest inland drainage area in Africa. Lake Chad (*q.v.*), a large sheet of freshwater with a mean depth of between 3½ and 4 feet, lies at the centre of the basin but not in its lowest part. Lake Chad is fed by three major streams, the Komadugu Yobe, Logone, and Chari, but these are in danger of having their waters claimed by the drainage systems of rivers that flow in opposite directions. Lake Chad itself, with an area of only some 5,000 square miles, was formerly far greater.

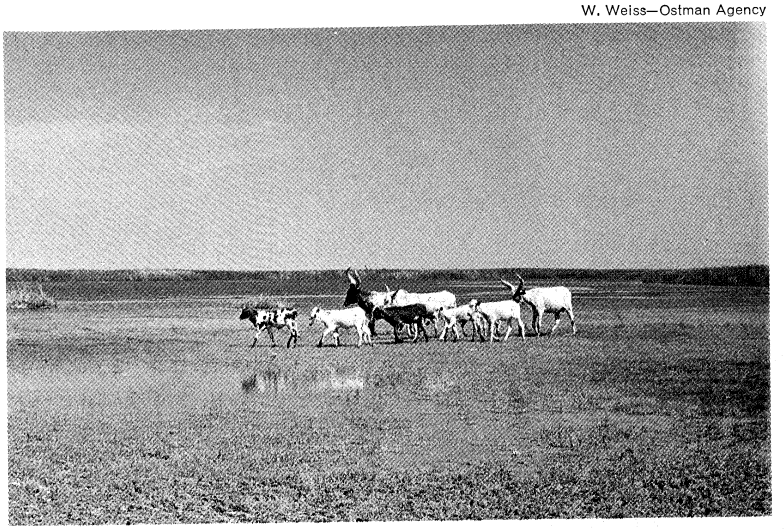
SOILS

Soil types. In general terms, soil types on the African continent may be divided into five or six broad categories—(1) desert soils; (2) chestnut-brown soils, which border the deserts; (3) chernozem-like soils (very black soils rich in humus and carbonates), which are found immediately south of the chestnut soils from the Democratic Republic of the Sudan westward to just beyond the Niger Bend (the bend in the middle course of the Niger River), and pockets of which are also found in East Africa, Zambia, Rhodesia, and South Africa; (4) black soils (often grouped with chernozems and found on the Accra Plains in Ghana); (5) red tropical soils and laterites (leached red iron-bearing soils), which occur in the tropical wet-and-dry and equatorial climatic zones; and (6) Mediterranean soils, found in the Atlas Mountains of North Africa and the Cape region of South Africa.

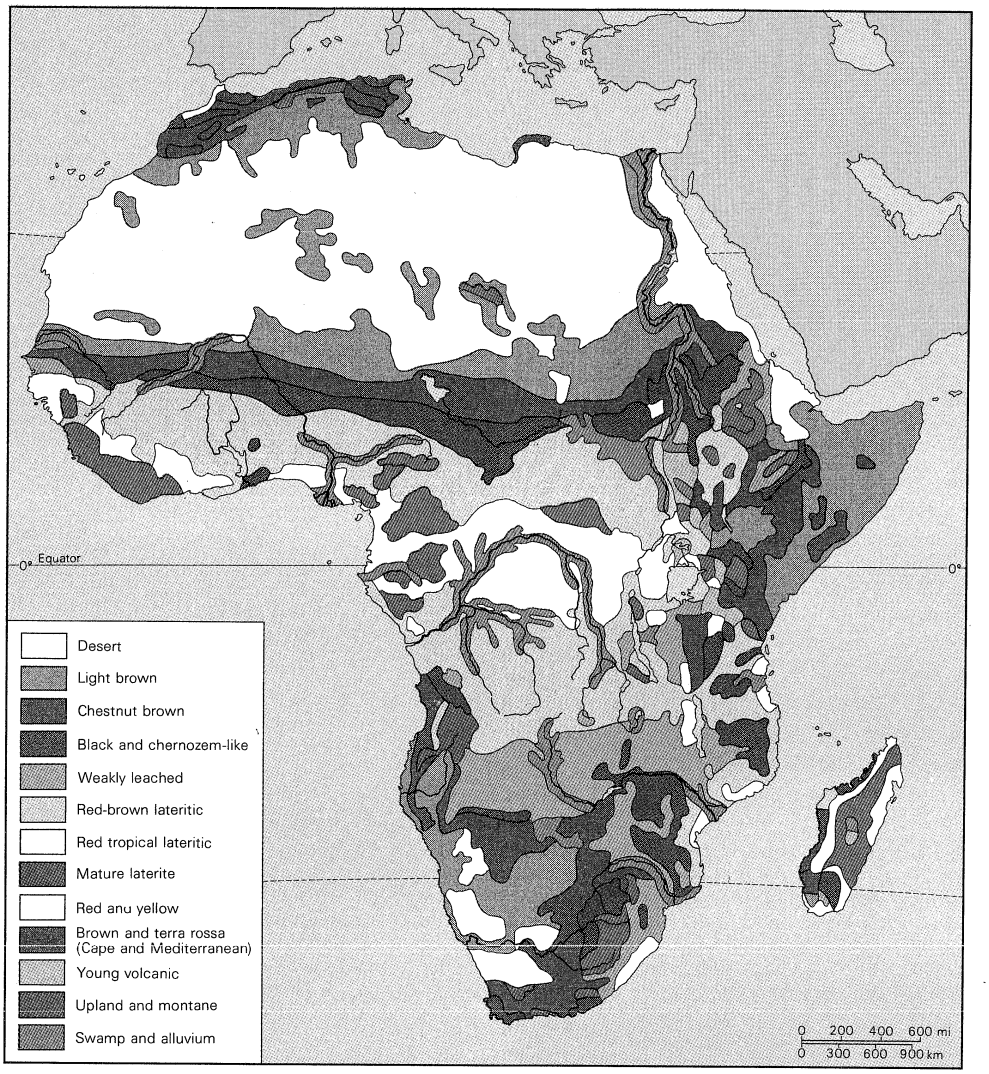
Factors affecting soil formation

The important factors affecting soil formation are climate, geology, relief, drainage, vegetation cover, and the passage of time. Where, as in Africa, the land has been generally stable and fairly flat for prolonged periods, climate becomes the major determinant of the major soil

W. Weiss—Ostman Agency



Zebu cattle grazing in the northern region of Lake Chad near Nguigmi, Niger.



Soils of Africa.
After S.J. Schokalskaja, reproduced by permission of UNESCO

groups. The different rocks are deeply weathered and are broken down into their common component elements to produce broadly similar soils under the same climatic conditions. Given sufficient time under a tropical climate, the differences in humus content of the great soil groups, which are introduced by vegetation types, are minimized. But within these groups there will naturally be differences in soil types as a function of local differences in physical factors.

Desert soils. These soils are characterized by the insignificance of their organic content; by the types of rock reflected in them, the chemical breaking down of which has been inhibited by the lack of water; and by the crusts or concretions of soluble salts on or just below their surface. While these crusts are in general thought to have been formed as a result of evaporation, it is nevertheless possible that they may have been formed under a wetter climate during the Pleistocene epoch.

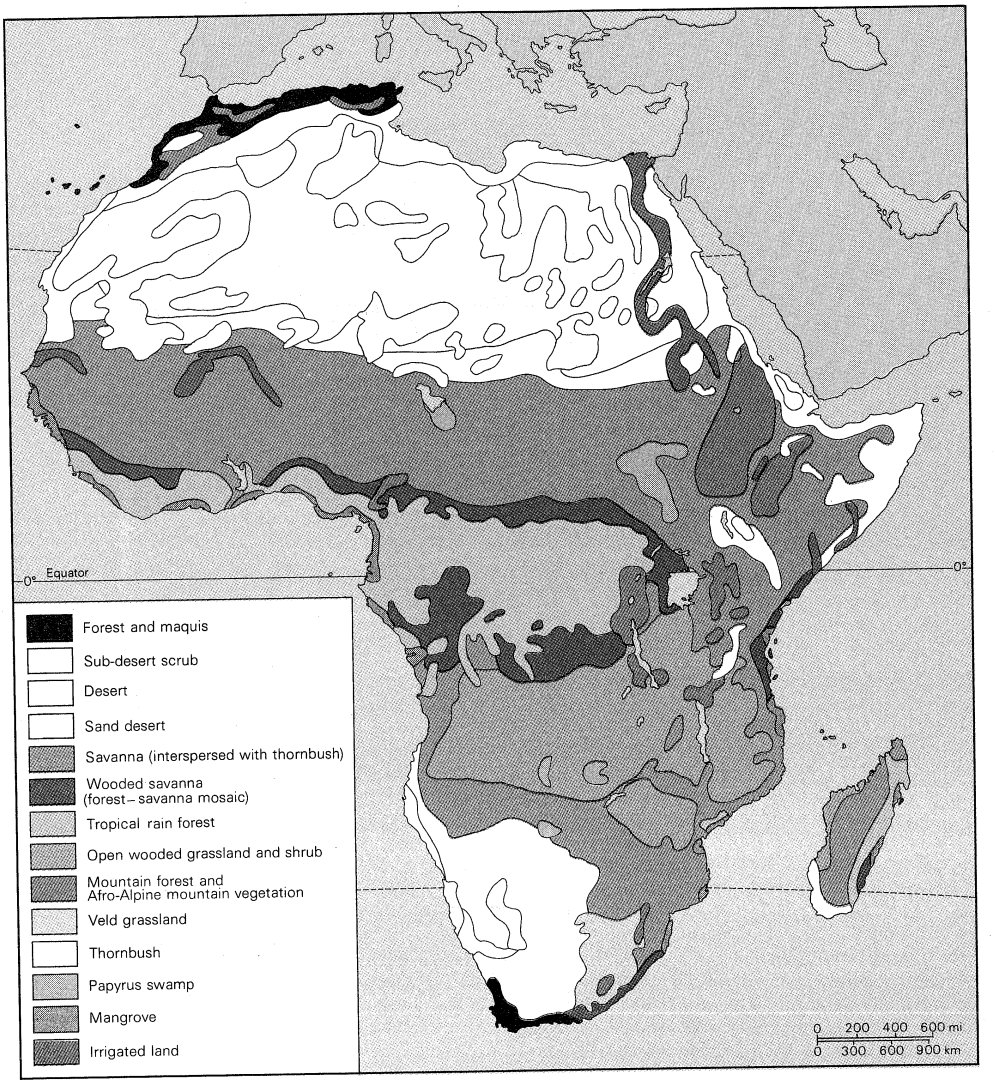
Chestnut-brown soils. In the semi-arid areas bordering the desert, increased rainfall makes grass vegetation more plentiful, results in rocks becoming more weathered than in the desert, and produces better developed soils with a higher humus content. It is the humus content that, according to the amount present, gives the chestnut soils their characteristic dark or light brown colour. Chestnut soils also differ from desert soils because they receive enough water to wash out some of the salt accumulations either on the surface or immediately below it.

Chernozem-like and black soils. An unfailing characteristic of the chernozem is the presence of a subsurface zone of calcium carbonate, sometimes accompanied by calcium sulfate, which is left behind after all the soluble salts have been washed out. Grouped with them are the black soils, which should, perhaps, be differently classified, for their black colour is not necessarily due to high humus content but rather to the presence of certain minerals, as in the black soils of the Accra Plains, in Ghana.

Red tropical soils and laterites. The majority of tropical soils have shades of colour varying from yellow and brown to red. The reddish colour reflects the presence of iron oxides that form as a result of chemical weathering. At one time, all tropical red earths or soils were indiscriminately referred to as laterites, although it is now clear that the term laterite should be confined to those tropical soils with large concentrations of sesquioxides (insoluble compounds) that have formed a hard pan at the surface or a little below it. At the most advanced state of laterization, bauxite, from which aluminum is extracted, is formed. Most tropical soils are, nevertheless, in varying stages of laterization, which is to say they are at various stages of accumulating insoluble compounds as the soluble elements are washed out. The compounds accumulate more readily in areas with a pronounced dry season and where the water table is not too far below the surface. If the top horizons (layers) of the soils should erode, the subsurface concentrations of sesquioxides are then exposed to the atmosphere, whereupon they crystallize irreversibly to form true laterite concretions.

Laterization of soils

Mediterranean soils. Mediterranean soils are generally deficient in humus not so much because of sparse vegetation cover as because of the slowness of the chemical processes that convert the vegetable matter to humus.



Vegetation zones of Africa.

Low rainfall, occurring when temperatures are lowest, retards chemical decay. The uneven surface relief of the regions where these soils occur also makes it difficult for mature soils to develop, since the land, except in the valley bottoms, is not sufficiently flat over wide enough areas to allow the soil-forming (parent) materials to remain *in situ* and thus to be thoroughly weathered.

Soil problems. Soil is the foundation of Africa's economic life, and as such its detailed study is most important. Failure to appreciate the physical and chemical properties of the soils has led to disastrous results for several projects for agricultural improvement (one example of this was the failure of the East African Groundnut Scheme in the late 1940s).

In studying the soils of Africa, it is essential not to lose sight of the importance of such social factors as the ability or inability of mostly uneducated farmers to judge the quality of the soil. Thus, schemes for transforming traditional systems of farming that are based on soil classification but that do not take into account local perception may have little chance of success.

For desert soils to be productive, they should be irrigated, as they are on the desert margins of North Africa; their excessive salinity or alkalinity should also be reduced. Compared to desert soils, the chestnut-brown soils are easier to work and are more productive under irrigation. Black soils tend to have a markedly crumbly structure and are sometimes difficult to plow. In the wet season, the black soils of the Accra Plains swell and become slippery, while in the dry season they shrink once more and crack to such an extent that they are said to plow themselves.

Red tropical soils need careful handling. Despite their luxuriant vegetation cover, high temperatures coupled with humidity promote the rapid decay of organic matter and keep the humus content low. Erosion is a constant threat if the soils are exposed to the elements for any length of time; the soils will only remain cultivable if the sesquioxides remain below the surface.

In the Atlas and Cape regions, there is a clearer relationship between soil characteristics and parent material than in the humid tropical areas. Over expanses of limestone, for example, the soils contain large amounts of calcium compounds, some of which must be washed away or neutralized before the soils can become fully productive. (K.B.D.)

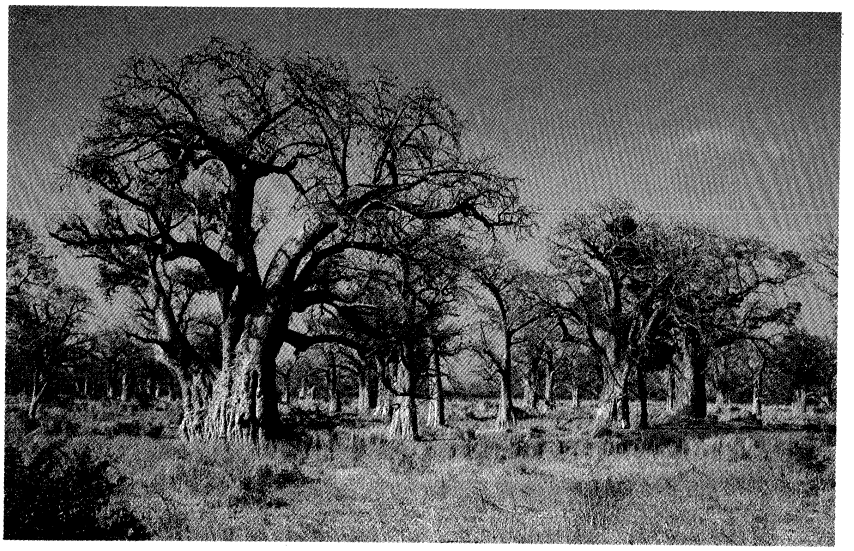
III. Vegetation and animal life

VEGETATION

African vegetation develops in direct response to the interacting effects of rainfall, temperature, topography, and type of soil; it is further modified by the incidence of fire, human agriculture, and depasturing by livestock.

Vegetational zones. With few exceptions, it may be divided into zones of decreasing humidity. The main zones are the tropical rain forest, mountain forest, woodland or wooded savanna, wooded grassland, thornbush, subdesert scrub, and desert zones.

Tropical rain forest. This vegetation develops in tropical climates wherever the rainfall exceeds 50 inches a year, with at least two inches of rain falling in any month. The wetter and warmer the climate, the more luxuriant the forest, although excessive rainfall, cloud cover, and relatively low temperatures may discourage growth. Rain



Baobab trees, native to Africa, growing in the savanna regions of Senegal in western Africa.
Scholz—Shostal

forest is the natural vegetation under such conditions, except where edaphic (soil) factors, notably swampy conditions, inhibit tree growth.

Characteristically, tropical rain forest is composed of a ground story, from six to 10 feet tall, of shrubs, ferns, and mosses; a middle story (or understory) of trees, palms, and lianas 20 to 60 feet in height; and a dominant top canopy consisting of trees of up to 150 feet high with straight unbranched trunks, buttressed roots, and spreading crowns of perennial leafage, the large branches of which support orchids, ferns, and mosses, or which are festooned with lianas.

Mountain forests. Occurring on the slopes of high mountains in Ethiopia, Cameroon, Angola, and East Africa, the mountain forests resemble those that grow at sea level in Cape Province. Like the tropical rain forests, they are characterized by a shrub ground layer, an understory, and emergent dominants. They have fewer species, however, and are usually less tall and luxuriant. Local variations may be discerned by the dominance of particular trees, which may include primitive conifers absent from true tropical rain forests. Mountain forests are usually isolated by large tracts of more xerophytic (drought-resistant) vegetation, but occasionally—as in Cameroon and in western Ethiopia—they may merge in unbroken altitudinal succession with lowland tropical rain forest.

Dominance of various trees

Woodland or wooded savanna. This type of vegetation covers huge lowland tracts from latitudes 10° to 15° north and south of the Equator, receiving 40 to 55 inches of annual rainfall, but with a dry season of at least four months. In the equatorial Intertropical Convergence Zone 4° north and south of the Equator, such woodlands also occur where 35 to 45 inches of annual rainfall is received. The Guinea savannas to the north of the Equator strikingly resemble the woodlands to the south, with wooded ridges of sandy loam alternating with swampy clay-based valley grasslands.

Such woodlands typically include a ground layer of grass and fire-resistant shrubs from four to six feet tall, the luxuriance of which varies according to the density of the canopy of deciduous fire-resistant trees, which rise from it to heights of from 30 to 50 feet. These trees shed their leaves in the dry season to spread them again just before the rains break. In the southern woodlands, *Brachystegia* (a leguminous hardwood tree, the bark of which is used for making cloth and roofing huts) and *Julbernardia* (a spreading leguminous tree which may grow to heights of from 40 to 50 feet) cover vast tracts with uniform vegetation, as do *Isoberlinia* (a tree resembling *Julbernardia* in many of its characteristics), *Daniellia* (a leguminous tree with white bark which reaches heights of from 50 to 70 feet), and *Lophira* (a spindly tree with fissured bark and simple strap-shaped leaves) in the northern savannas. In areas of lower rainfall, *Combretum* and *Terminalia* (hardwood trees with simple leaves and winged seeds) tend to dominate.

The boundary between forest and savanna is usually abrupt. In forested areas receiving marginal rainfall, the resultant interlocking fingers of savanna and forest are known as forest–savanna mosaic and are often characterized by forests along the watercourses and by savannas on the ridges.

Wooded grassland. A gradual transition to wooded grassland, or acacia savanna, occurs as the rainfall decreases from 30 to 20 inches a year. Small (20 to 30 feet) deciduous or fine-leaved trees (mainly acacia) occur at intervals, rising from a continuous sheet of perennial grass two to four feet tall that also contains a few small shrubs.

Thornbush vegetation. Occurring in areas with an annual rainfall of 20 to 12 inches, thornbush, or bushveld, vegetation is characterized by the reduction of perennial grasses and the predominance of xerophytic, thorny, sometimes succulent or semisucculent trees, such as acacia, *Commiphora* (the myrrh tree), or *Boscia* (an evergreen hard-leaved tree). The baobab in these areas reaches heights of up to 100 feet. A relatively luxuriant shrub layer, often forming dense thickets, is admixed with succulents, such as aloes, *Sansevieria* (a fibrous succulent species), and *Adenium*, or Desert Rose (a succulent shrub with smooth grey bark, a huge water-storing base, and beautiful red or pink flowers), and smaller euphorbias. This type of vegetation covers vast tracts in East Africa, Somalia, Botswana, The Sudan, and West Africa.

Subdesert scrub. In areas where annual rainfall decreases from 12 to 5 inches, tree growth and perennial grass—surviving in narrow strips along watercourses—separate much larger areas of sparse annual grasses and scattered low (four to six feet) shrubs, often mainly acacias. Shrubs may often be salt-tolerant. While shrubs may die as a result of inadequate, erratic rainfall, they are little affected by the rare fires that occur. This type of vegetation covers huge areas on the southern edge of the Sahara, in the Horn of Africa, and in the Kalahari of South Africa and Botswana.

Salt-tolerant and fire-resistant shrubs

Desert vegetation. More than a quarter of Africa has less than five inches of rainfall and is thus typified by the absence or extreme sparsity of vegetation, except where water that has run off the surface and has collected in depressions encourages the growth of small trees such as acacia. Desert vegetation consists of two main types: (1) perennials with huge root systems and sparse aerial parts, often protected by waxy cuticles, thorns, or hairs; and (2) ephemerals, with slight root systems and little foliage but with the ability to flower profusely immediately after oc-

casional storms and then to seed quickly and abundantly —the seed remaining dormant until the next storm, for years if necessary.

Other vegetational types. Special types of vegetation, not necessarily controlled by the rainfall pattern, include sudd, edaphic valley grasslands, macchia, and Afro-Alpine mountain vegetation.

Sudd consists of an association of such sedges as papyrus with reeds and other water plants, such as the floating Nile cabbage (*Pistia stratiotes*); this mass of vegetation usually floats and is almost totally unproductive.

Edaphic valley grasslands consist of associations of tall grasses, which occur in seasonally inundated areas of larger river valleys. They are partly maintained by intense dry-season fires, which burn off the dead grass and promote renewal; they constitute true edaphic climax vegetation (that is to say, they represent a predominant type resulting more from soil than from climate) and, unlike the sudd, are intensely productive of food for both domestic and wild animals.

Macchia, a dense growth of xerophytic or sclerophyllous shrubs (*i.e.*, with foliage hardened so as to retain water) and small trees, often fire-resistant, occurs in northwestern (Mediterranean) Africa and at the Cape. Mediterranean macchia includes *Quercus* (a genus of hardwood trees and shrubs, often evergreen, including typical oaks), merging gradually on mountain slopes with pines and cedars. South African macchia is florally much richer and is the remnant of a once much larger tract of similar vegetation. Ericaceous plants (a family that comprises the heaths) abound, and many indigenous families occur. Giant heath (*Erica arborea*) is common to both areas. To the north, beyond the mountains, southern Africa macchia merges into the subdesert scrub of the Karroo region, which is characterized by small shrubs and by *Salsola* (the Russian "thistle").

Heath and scrub

Afro-Alpine mountain vegetation occurs on Ethiopian and East African mountains at elevations above 10,000 feet, where nightly frosts occur. This vegetational type is made up of associations of mainly ericaceous plants together with tussock grasses and sedges; superficially, these associations are similar to macchia. A unique feature is the profusion of giant *Senecio* (groundsel or ragwort) and *Lobelia* species (plants with a basal rosette of smooth, soft leaves, from which emerges a tall spike of flower), often specific to particular mountains. *Helichrysum* (a large genus of African and Australian plants with long-lasting flowers) are equally characteristic but are individually more widespread. Vegetation dies out at about 15,000 feet on East African mountains and at 13,500 feet on those of Ethiopia.

Influences on vegetation: past and present. Though they may have been modified by climatic changes, movements of the Earth's crust, and the varying position of the caloric Equator, the vegetational zones of Africa have probably existed in approximately the same climatically controlled series and with the same characteristically developed plant and animal life for a long period of time. The main features of existing vegetational distribution are adequately explained by climatic changes that took place during the last 1,000,000 years when the zones expanded and contracted at least four times. Successive glacial advances in the Northern Hemisphere were probably associated with wet (pluvial) periods in the African tropics, while glacial retreat conversely corresponded to increased aridity and warmth.

Past climatic influences. If, as is assumed, the rainfall over much of Africa in the last pluvial, which ended about 10,000 to 15,000 years ago, was 140 to 150 percent that of the present day, the mean annual temperature would have been depressed by about 9° F (5° C), and glaciers would have descended to about 10,000 feet on the East African mountains. The upper vegetational limit would then have been depressed to about 9,000 feet, Afro-Alpine vegetation would have covered huge tracts of present-day savannas on plateaus at heights between 5,000 to 6,000 feet, and continuous forest would have extended from Senegal to Tanzania and from Ethiopia to South West Africa.

Conversely, during warmer, drier interpluvials, mountain vegetation would have retreated to higher levels, to die out altogether on isolated mountains of less than 6,000 to 7,000 feet high. With rainfall perhaps half of the present amount, acacia steppe (treeless plains) would have advanced into woodland, and woodland into forest, which would then have been reduced to small areas where comparatively few species could survive. Kalahari sand appears to have been blowing in eastern Angola 11,000 to 13,000 years ago, with arid conditions extending to the Zambezi. Much of the present *Brachystegia* woodland was acacia steppe, and the woodland itself occupied much of the present-day equatorial forest areas. In interpluvials, arid tracts coalesced in Central and East Africa, permitting the north-to-south spread of adapted species.

The historical expanse of Kalahari sand

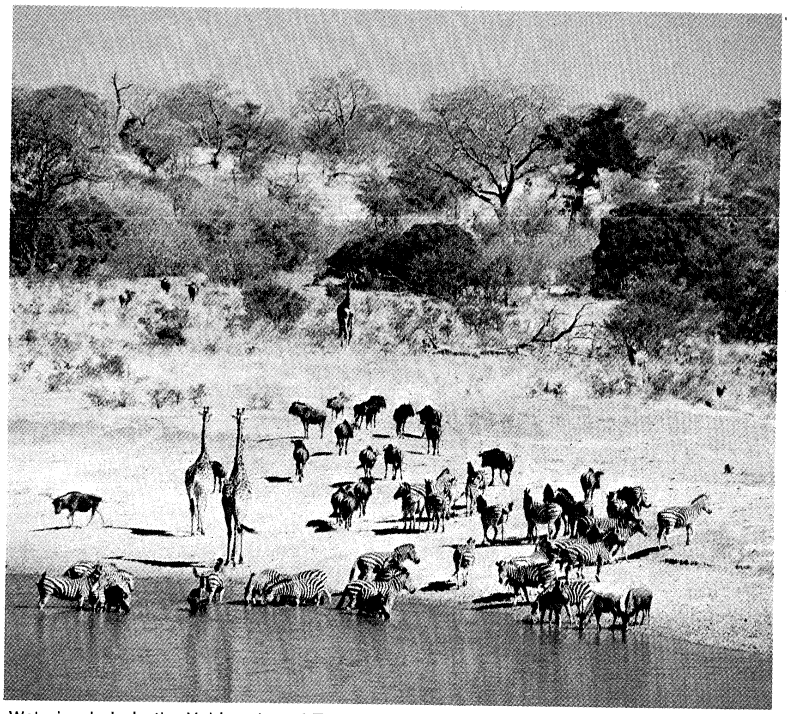
During the pluvials, forest separated the northern savannas from the southern, isolating related forms and encouraging increased speciation. Savanna vegetation, with flowing rivers, advanced northward across the Sahara to meet Palearctic vegetation (*i.e.*, belonging to the biogeographic region that includes Europe, Asia north of the Himalayas, northern Arabia, and Africa north of the Sahara) forced southward by glaciation. The European lime—no longer extant in North Africa—was found at Ahaggar in the Sahara. The whole western Sahara has probably become desert within the last 5,000 years.

Influences of man. In the last 2,000 years, and within the last century especially, man has impoverished African vegetation by the use of fire, by cultivation, and by the introduction and depasturing of livestock. From perhaps 350,000 years ago, when primitive man first acquired the use of fire, its previously sporadic effects intensified, for then as now a few hunters could burn disproportionately large tracts of growth. In most of Africa, human cultivation has become catastrophic in its effect only within the last 50 years, the process being greatly accelerated by the trebling of the population since 1900 and the provision of better tools. Great tracts of savanna have in consequence been impoverished, and in forested zones a type of man-made savanna—derived savanna—now covers large areas. Many mountain forests have been almost totally destroyed.

Destruction caused by population increases

In the main, the depredations of livestock have adversely affected the vegetation of semi-arid areas close to permanent water and free from the deadly tsetse fly. These effects have been accelerated, especially in the last 30 years, by the effective control of epizootic diseases (*i.e.*, those affecting many animals simultaneously) and by continuing bad management (grazing of too many livestock, excessive use of fire, and failure to use selection to improve stock when breeding). Typically, by eating down the grass, cattle and sheep reduce inflammable material and so diminish the frequency and intensity of fires to such an extent that uneatable shrubs and bush gain ascendancy. The animals continue to graze until the grass is completely suppressed and the ground is covered with tall unpalatable thickets. Browsing livestock (goats and camels) then tend to replace the grazers, or the tsetse fly may invade the area and exclude them all. The overgrazing effect is worsened by the fact that the bare soil under trees no longer absorbs so much moisture, and the amount of runoff is increased.

The ecological balance. Where relatively unaltered vegetation still exists, it shows close adaptation to topography, soil type, and drainage. Where temperature and rainfall are relatively constant in wide belts of savanna, a succession of gently rolling ridges and valleys is covered with broad-leaved woods on sandy ridge soils, while swampy grassland is covered with more xerophytic acacias on valley clays. In a small area of higher rainfall of only a few square miles, there may be a visible succession from actual swamp with papyrus, through swampy grassland and broad-leaved savanna to a patch of forest on rich hillside soil, and finally to succulents and xerophytes on a nearly naked rock summit. These effects, not explained by climatic variation, demonstrate that, even in the absence of man and livestock, the African countryside maintained a mosaic of natural habitats.



Watering hole in the Veld region of Transvaal, South Africa.
John Rushmer—Ostman Agency

Other influences. In response to climatic and other influences, the fluctuation and varying extent of African vegetational types have led both to the diversification of and to the emergence of distinct species of native plants and animals; at the same time, such variations have largely controlled or affected the activities of man. Where the tsetse fly, which is associated with forest or well-watered savanna, effectively excludes livestock, man must fall back on fire and agriculture. In semi-arid areas where crop cultivation becomes impossible, man affects the vegetation through fire and livestock. Some 50,000,000 pastoralists thus inhabit three-fifths of the continent.

ANIMAL LIFE

Zoogeographical regions

Africa includes parts of two main zoogeographical regions: (1) the Ethiopian region, which comprises the continent south of the Sahara and part of Arabia, as well as Madagascar; and (2) North and northwest Africa south to the Tropic of Cancer.

Genera and distribution. Of all the continents, Africa is the richest in large ungulates (hoofed mammals) and freshwater fish. There are some 90 species of ungulates and 60 of Carnivora (the order of flesh-eating mammals). Among ungulates of the order Artiodactyla (composed of mammals with an even number of toes), several families, including giraffes and hippopotamuses, are now endemic (peculiar to a locality). Some families of Carnivora, such as civets (catlike carnivorous mammals of the Viverridae family) and their relations the genets, which are small carnivores, as well as hyenas are chiefly African. The rodent family of jumping hares (Pedetidae) is endemic, and one order, the aardvark (Tubulidentata), a large nocturnal burrowing mammal, with one species, is exclusively African.

The primates include some 45 species of Old World monkeys, as well as the world's largest ape, the gorilla. Presimian primates—such as pottos (African lemurs) and galagos (bush babies, or small arboreal lemurs), as well as Lorisidae (a family of arboreal lemurs, moving with a slow, delicate crawl), are mainly small and nocturnal, but in Madagascar, where there are no true monkeys, a varied animal life of large and small diurnal and nocturnal presimian lemurs survives. Madagascar also has a remarkable insect-eating family, the tenrecs (animals with long pointed snouts, some of whom are spiny and tailless). Salt-water aquatic mammals include one Mediterranean and one South African seal, and two Sirenia (an order of aquatic, herbivorous mammals)—the dugong and the manatee.

South of the Sahara, the birdlife includes some 1,480 resident species, to which must be added another 275 species that are either resident in Northwest Africa or else are Palearctic winter migrants; these migrants total perhaps 2,000,000,000. Birds are mainly of Old World families, but of those that are endemic, the most noteworthy are perhaps the ostrich, shoebill, hammerkop (a brown heron-like bird), secretary bird (a large long-legged predatory bird), and the touracos (brightly coloured birds, some with helmet-like crests). Other families, such as bustards, sand grouse, honey guides (small dull-coloured birds, some species of which are noted for leading men to honeybees' nests, in order to feed on them after the nests have been broken), and larks are predominantly African.

Reptiles, of which there are few endemic families, have mainly Old World affinities. Those most likely to be seen include lizards of the agamid family, skinks (a family of lizards characterized by smooth overlapping scales), crocodiles, and tortoises. Endemic reptiles include girdle-tailed and plated lizards. Lizards of the iguana family and boa constrictors occur only in Madagascar. Large vipers are abundant and varied. A wealth of both colubrine snakes (with fangs at the posterior end of the upper jaw) and elapine snakes (with fixed poison fangs at the front of the upper jaw) include such highly venomous elapine species as mambas.

Amphibians mainly belong to Old World groups. Salamanders and hylid tree frogs (having teeth in the upper jaw) are confined to the Palearctic Northwest. Abundant commoner frogs and toads include such oddities as the so-called hairy frog of Cameroon, whose hairs are auxiliary respiratory organs. The frog family Phrynomeridae is exclusively African.

Archaic fish species

Freshwater fish, of which there are altogether some 2,000 species, include both remarkable archaic forms and examples of sudden recent evolution. The coelacanth, an archaic marine form, believed extinct for 70,000,000 years, was rediscovered in 1938, since which time others have been found. Other ancient freshwater forms, such as lungfish (*Protopterus*), bichirs or lobefins (*Polypterus*), and reedfish (*Calamoichthys*) can all breathe air—a property that certain catfish (Clariidae),

which are able to travel overland for some distance in wet weather, also possess.

A rich and varied invertebrate animal life on the east and west coasts includes marine organisms typical of the Indo-Pacific and Atlantic oceans. Coral reefs and associated organisms are mainly eastern, while the southwest coast, washed by the cold Benguela Current, abounds in fish. Among arthropods (a grouping that includes insects and other segmented invertebrates) are to be found large butterflies of the *Charaxes* (brush-footed) and *Papilio* (swallow-tailed) genera, stick insects, and mantises, grasshoppers, driver ants (tropical ants that travel in vast swarms), termites, and dung beetles. Spiders and scorpions also abound. At one time most African fauna was thought to derive from elsewhere. There is no doubt, however, that as little as 15,000 years ago an amelioration of the present Saharan climate enabled such typical Ethiopian forms as clariid catfish to reach the river systems of North Africa. Likewise, Palearctic animal life and vegetation appear to have extended far south into the Sahara, and the white rhinoceros apparently lived beside elklike, typically Palearctic deer.

Within the Ethiopian region, repeated climatically controlled expansion and contraction of vegetational zones resulted firstly in organisms establishing themselves in numerous ecological communities (niches) of plants and animals, and secondly in the proliferation of those species that successfully adapted themselves to the prevailing conditions. The spreading of forests during the pluvials, separating northern and southern savannas, led to the evolution of such closely related northern and southern species of antelope as the kob and puku, the Nile and common lechwe, as well as of the northern and southern races of white rhinoceros.

Some subfamilies of Bovidae (the ox family), like the spiral-horned antelopes (Tragelaphinae), have adapted to almost every ecological environment—forest, savanna, and Afro-Alpine zones, and even to papyrus swamp. Others, like the hartebeests (Alcelaphinae), which inhabit savannas and grasslands, are less adaptable.

Freshwater fishes demonstrate the existence of the relation to one another of past river systems and lakes. Large rivers containing Ethiopian fish evidently existed quite recently in the northern Sahara. The fish life of the now isolated Lake Rudolf, in East Africa, demonstrates that the lake was once connected to the Nile, though Lake Victoria, the present source of the White Nile, was not. Lake Kivu, too, was formerly connected with the Nile, but as a result of recent volcanic activity is now part of the Congo drainage system.

In the recent past, the animal life was even more remarkable than today. Fossil deposits reveal sheep as big as present-day buffalo, huge hippopotamuses, giant baboons, and other types similar to existing species. These huge types probably lived in pluvial periods, dying out as aridity increased. Smaller types survived.

The effects of man. Until the introduction of firearms, man had relatively little effect on the animal life. Within the last century, however, and particularly since 1940, direct or indirect human wastage of Africa's animal life has been intense. The antelope known as the Zambian black lechwe, for example, believed to number 1,000,000 in 1900, has been reduced to less than 8,000.

Though European hunters and colonists were rightly blamed for much of the early decline, hunting, destruction, or the disturbance of habitats by Africans has recently become more important. Rinderpest, an acute, and usually fatal, infectious disease of livestock, entering Africa with domestic stock in the 1890s, decimated many indigenous antelopes; the accelerated spread of agriculture and stock raising involving the destruction of forests, and excessive grazing and burning of vegetation eliminated large animals from wide tracts. In the 1960s, for example, in the southern Sudan, political strife and warfare have entirely eliminated wildlife from some areas. The demand for fancy leather and fur also endangers the Nile crocodile and the leopard.

Man has, however, been of benefit to many smaller species. Dams and irrigation schemes benefit waterfowl, frogs, and fish, and the spread of grain crops encourages certain pests. Even the patchy cultivation of forests results in the development of a mosaic of habitats.

Ecology. Sufficiently large tracts of relatively unspoiled country still exist in which animal life may be studied in its environment. The recently revealed complementary roles of wild ungulates, for example, show that in any area inhabited by a wide variety of species, the grass is grazed in regular succession and at different stages of growth—for example, by zebra, wildebeest, hartebeest, and gazelle—while specific adaptations enable a still greater variety to survive. A much smaller variety of domestic stock cannot duplicate such effects. Overpopulation by domestic or wild species may upset the delicate natural balance, as may be seen by the example of elephant overpopulation in Kabarega Falls National Park in Uganda and in Tsavo National Park (Kenya); whether the elephants survive or not they have ineradicably altered the environment to the detriment of many other typical species.

Complementary grazing systems

Animal life of particular interest. Animal life of particular interest to man includes four main groups that are not mutually exclusive. They are: (1) species potentially or actually useful to man as food (large ungulates); (2) dangerous or pest species that may have to be controlled or eliminated (locusts, tsetse flies, *Quelea* finches or black-faced diochs, which do immense damage to grain crops, and some ungulates or carnivores); (3) species that provide a spectacle and bring economic benefit (elephants, the larger plains ungulates, primates, or carnivores); and (4) endangered, rare, or unique species.

Intensive studies, associated with human behaviour and psychology, have, in the 1950s and 1960s, been made of such primates as baboons, the Ethiopian gelada, to which baboons are related, and especially of the chimpanzee and the gorilla. Studies of such primitive primates as the lemurs of Madagascar show them to be threatened by the destruction of their habitat, and reveal that the forest lemurs, including the largest (*Indri*), are especially in danger. The unique aye-aye (*Daubentonia*), a nocturnal lemur feeding on insects and fruit, has been shown to be nearly extinct, not only because of habitat destruction but also because it is killed for superstitious reasons.

Other studies show that, in areas made inaccessible to domestic stock by tsetse fly, controlled populations of wild ungulates may help to compensate for the chronic protein shortage among African peoples. Studies of predators, such as the lion, have shown that they do not generally control the numbers of ungulates as do disease or starvation. It has also been established that the hyena is as much a potent predator as a scavenger. Since about 1960 the territorial behaviour of many antelopes has also been revealed. Such pest species as the desert locust, tsetse, mosquito, and *Quelea* finch have been intensively studied as a first step in their control.

Conservation. Many countries have now set aside large tracts as national parks, game reserves, or forest reserves. Of these parks, only some are large enough to be self-contained ecosystems, and most have been set aside for large mammals. In East Africa there are also sanctuaries for birds and marine organisms. The conservation of vegetation is undertaken mainly in forest reserves but also in national parks.

The oldest and best known national park is the Kruger National Park in South Africa, where representative populations of most savanna species are maintained. The Kalahari Gemsbok National Park, which South Africa shares with Botswana, conserves a tract of arid country with such associated types of antelope as springbok and gemsbok; smaller reserves and parks conserve particular species. Only one large mammal species, the blaubok (or blaauwbok), has become extinct, though several subspecies have nearly disappeared; one such subspecies, the quagga, a race of zebra, has vanished.

East and Central African countries have large national parks, which during the 1960s have been expanded in size or have increased in number as a result of the economic benefits of tourism. Kenya's parks include Tsavo, one of the largest, with an area of 8,000 square miles,

Nakuru Lake Forest Reserve for flamingos, montane national parks, and a marine park. Uganda has three large national parks, Tanzania has the famous Serengeti National Park, with its unrivalled populations of plains ungulates, and the parks of Ruaha, Lake Manyara, and Arusha. Zambia has the Kafue National Park and the Luangwa Valley Parks; Rhodesia has Wankie; Mozambique has the Gorongosa; and Botswana has the Chobe Game Reserve and the Central Kalahari Game Reserve. All these parks conserve representative woodland, savanna, or thornbush species.

Depletion of wildlife

Elsewhere the situation is less satisfactory. In Zaire, national parks were all seriously depleted after independence. But in 1972 Zaire announced a program that would, by 1975, place 15% of the country within 7 existing, and 10 to 15 new, national parks. Though The Sudan is one of the countries with the largest numbers of remaining ungulates, little of it is known, and it is not conserved by an adequate park network. Ethiopia has parks largely on paper, and Somalia only a rudimentary system consisting mostly of game reserves and one area (Luc Badana) nominally a national park, but failing to meet UN criteria for such parks. The "W" park is shared between Upper Volta, Niger, and Dahomey, but most West African countries have only small national parks if any at all.

Animals not conserved in parks are not necessarily threatened. Indeed, many large species are still plentiful in forest or game reserves or in controlled hunting areas. Well-managed forest reserves in particular provide secure habitats for many smaller unprotected forms.

Animals affecting land usage. Historically, the abundance of large ungulates may have stimulated some men —notably Europeans in southern Africa—to occupy certain areas. Trade routes were established and early hunter-explorers were influenced by the availability of elephants, whose ivory tusks slaves could carry.

Currently, large or small animals may affect humans by competing with them or with their stock. They may prey on man or on stock or carry diseases affecting either. The bilharzia snail and the *Simulium* fly (host to an organism causing blindness), the tsetse fly, and the mosquito collectively affect human beings and their stock far more than do such individually large or formidable species as lions or elephants.

The larger ungulates may compete with man's domestic stock for forage but are more keenly resented and hunted by agriculturists, whose crops they may eat. Pastoralists are much more concerned about lions that occasionally kill cattle than about locusts or rats, which, in depleting forage, indirectly cause the death of much stock.

Danger to humans from large animals, carnivorous or otherwise, has been greatly exaggerated, but disease carried by living organisms is still a serious problem. Indigenous African ungulates are resistant to the diseases (such as trypanosomiasis, a parasitic organism that causes sleeping sickness) carried by tsetse, but domestic stock, none of which has been in Africa much more than 4,000 years, is not. Despite tremendous efforts to eradicate tsetse, the insect still occupies more than 2,000,000 square miles of forest and savanna, from most of which it cannot economically be cleared. (L.H.B.)

IV. Natural resources

MINERAL RESOURCES

Africa's mineral wealth

Africa's known mineral wealth places it among the world's richest continents, with a very large share of the world's mineral resources: coal (7.5 percent), petroleum (8 percent), natural gas (12 percent), uranium (30 percent of the reserves in the non-Communist world), radium (the chief source), thorium (a grayish-white, lustrous, radioactive metal; 20 percent), iron ores (3 percent), chromium (almost the entire world reserve), cobalt (90 percent), copper (20 percent), lead (8 percent), zinc (5 percent), tin (over 15 percent), bauxite (27 percent), titanium (the deposit at Sherbro, in Sierra Leone, is greater than any yet found in the world), antimony (about 7 percent), gold (50 percent), platinum (40 percent), tantalum (a gray, hard, rare metallic element; 80 percent), germanium (a scarce, metallic, grayish-white element; the bulk of world reserves), lithium (a soft, silver-white metal; more than any other continent), phosphates (over half of world reserves), and diamonds (the bulk of world reserves).

Mineral fuels. Information about Africa's more important mineral fuels is often fragmentary in character. Moreover, any assessment of mineral reserves is necessarily valid only for a short span of time, as old mines are depleted and new ones discovered.

Coal. Major deposits are confined to three groups of coal basins. The largest is in southern Africa, with a combined total of 80,000,000,000 tons; South Africa possesses much of this total, amounting to possibly 6.5 percent of world reserves. Other deposits exist in Rhodesia (more than 1,000,000,000 tons), Madagascar (about 500,000,000 tons), Malawi (several hundred million tons), Mozambique (100,000,000 tons), Tanzania (400,000,000 tons), and Zambia. Two groups of basins in Zaire are to a large extent unexplored.

The second group of coal deposits is in North Africa, and contains less than 1,500,000,000 tons. They consist of a number of large basins extending from Algeria into Morocco. The third potential source of coal is in West Africa, particularly in Nigeria, where reserves total 350,000,000 tons.

Petroleum. Proven petroleum reserves (1971) in North Africa occur in Libya (28,000,000,000 barrels), Algeria (9,840,000,000 barrels), Egypt (1,000,000,000 barrels), Tunisia (410,000,000 barrels), and Morocco (7,750,000 barrels). So far, exploration has been concentrated north of the Aïr-Ahaggar massifs; there may also be major Saharan reserves to the south. The other major oil reserve is in the West African coastal basin. Nigerian basins contain 10,000,000,000 barrels, with the Gabon Basin containing 602,000,000 barrels, People's Republic of the Congo 8,000,000 barrels, and Angola 1,000,000,000 barrels.

Natural gas. Africa's natural-gas reserves are concentrated in basins of North and West Africa. Algeria has the lion's share (156,000,000,000,000 cubic feet in 1971), thought to be among the world's largest. The other countries with the most commercially promising reserves are Libya (27,100,000,000,000 cubic feet), Egypt (2,000,000,000,000 cubic feet), Morocco (about 17,000,000,000 cubic feet), Nigeria (10,000,000,000,000 cubic feet), and Gabon (858,000,000,000 cubic feet).

An unusual source of natural gas exists in Central Africa, where the waters of Lake Kivu, shared by Rwanda and Zaire, contain a solution of about 2,100,000,000,000 cubic feet of methane gas. Mention should also be made of the low-grade source of oil, estimated at 15,000,000,000 tons, in the bituminous tar sands and oil shales that are widespread in southern Africa, particularly in Angola, Zaire, and Mozambique.

Uranium and fissionable elements. Southern Africa is said to be one of the world's seven major uranium provinces. In South Africa the unusual degree of knowledge of reserves derives from the fact of the joint occurrence of uranium with gold, a fact that also decreases the cost of production. Total immediate reserves amount to 225,000 tons, with a further 17,000 tons possible. In Zaire, uraninite veins are associated primarily with cobalt and nickel occurrences at Shinkolobwe–Kasolo and at Kalongwe, in reserves totalling 7,000 tons of low cost uranium. Elsewhere in Africa the mineral occurs in unusually high concentrations with phosphates. Deposits occur in Morocco; in Algeria; at Thiès, in Senegal; and at Bakouma, in the Central African Republic, where a vast reserve of about 10,400 tons of uranium metal is known.

Uranium deposits

In the early 1970s, deposits of major international significance were known to exist in Central and West Africa. At Mounana, in Gabon, a reserve of 13,500 tons of uranium is known to exist, and, in Niger, ores in sandstone are thought to contain 26,000 tons. In Somalia, exploration of the Bur region has revealed the Alio Ghelle deposit, known to include 290 tons of metal in 300,000 tons of ore.

Africa is thought to possess about 100,000 tons of thorium. So far, however, exploration for this element as such has been limited. The chief ore mineral is monazite, in which thorium is contained in varying proportions. The primary ore is found in South Africa, Malawi, Burundi, Kenya, Niger, and Nigeria. Elsewhere, secondary concentrations of monazite are found in the Ivory Coast, in the beach sands of Madagascar, and in the black sands of the Nile Delta in Egypt.

In Zaire, Africa possesses the world's chief source of radium—located in the uranium ores of the Shinkolobwe–Kasolo area.

Metallic deposits. *Iron and iron ores.* Although iron-ore reserves are known to be widespread, knowledge of the deposits is somewhat limited.

In North Africa, reserves of iron ore are concentrated in the Atlas and western Sahara. Egypt, however, has 433,000,000 tons of low-grade reserves, of which the most important are in the al-Kāḥāt al-Babrīyah. The ore deposits in Morocco and Tunisia, which were of considerable importance in the past, have been severely depleted. Algeria has proven reserves of about 586,000,000 tons. The largest deposit of North Africa is at Gara Djebilet, Algeria, where there is a minimum of 432,000,000 tons of ore. Algeria is also estimated to have 1,100,000,000 tons of potential ores of which all 1,100,000,000 tons are located at Gara Djebilet. Africa's most significant iron reserves, however, are to be found in West and southern Africa. Potential lateritic ores are widespread in Ghana (more than 200,000,000 tons), in Nigeria (60,000,000 tons), and in Guinea, where the laterities of the Presqu'-île de Camayenne have iron-ore content. Malagasy lateritic crusts also have high iron content.

West African iron-ore deposits

It is the sedimentary Precambrian rocks themselves, however, particularly in West Africa, that are proving the basis of Africa's role as a major world producer of iron ore. The most significant deposits are in Liberia, where 713,000,000 tons of known reserves exist in the Bomi Hills, Bong and Nimba ranges, and the Mano Valley deposits; in the extension into Guinea of the Nimba–Simandou ranges, where 1,000,000,000 tons of hematites have been located; in Mauritania (465,000,000 tons); in Zaire, where vast reserves are present; and in Gabon (1,000,000,000 tons). Weathering of such formations has produced the Cassinga ores of Angola (more than 100,000,000 tons), while the massive parent rock is assumed to contain further supplies.

In Southern Africa most iron-ore reserves lie in South Africa itself. The chief deposits are at Postmasburg, which has 200,000,000 tons of known and 600,000,000 tons of indicated reserves; in the Bushveld Complex, which has at least 200,000,000 tons; in the eastern and northern Transvaal; at Thabazimbi; and finally in the vast, low-grade deposits of Pretoria (1,000,000,000 tons). There are also substantial reserves in Rhodesia, which has 200,000,000 tons of ore near Que Que, as well as over 200,000,000 tons elsewhere in the country.

While this outline completes present known reserves of international importance, it excludes those that may be only of national significance on account of inaccessibility or low grade. Mention should be made, however, of the considerable Zambian reserves; of more than 90,000,000 tons of potential ore in Uganda; of 78,000,000 tons in Tanzania; and of outcrops of banded ironstones in Kenya, Cameroon, and the Central African Republic; as well as of the as yet unevaluated reserves of the People's Republic of the Congo.

Ferroalloy metals. Africa's reserves of minerals used as ferroalloys in the steel industry are even more striking than its vast share of world iron-ore reserves. This is particularly true of chromium. Almost the entire world reserve of chromium is found in southern and, to a much lesser extent, in western and northeastern Africa. The highest concentrations are found in Rhodesia, where the Great Dyke contains reserves estimated at 600,000,000 tons. South Africa contains the other major world reserve of chromite, estimated at 2,000,000,000 tons. As compared with these two sources, reserves elsewhere in Africa are relatively small.

Manganese reserves are considerable. In South Africa, tens of millions of tons of the ore are thought to be contained in the Kalahari Manganese Field and elsewhere. The Mouanda deposit in Gabon is thought to be one of the largest in the non-Communist countries and is assessed at 96,000,000 tons. Ghana is another important source of manganese, having both low-grade and high-grade reserves.

Manganese reserves

Elsewhere in West Africa, manganese deposits are situated in Upper Volta (more than 10,000,000 tons) and the Ivory Coast (with reserves of similar size), as well as in Zaire and Cameroon. In North Africa, manganese is found in Morocco and in Algeria (the latter of which has reserves of 1,500,000 tons of ore).

Africa's contribution to world resources of other ferroalloys is, by comparison, insignificant. Nickel is of some importance, occurring in other metalliferous ores in Southern Africa.

There are only two significant reserves of tungsten on the continent, although wolframite ore (an important ore of tungsten) occurs in Rhodesia, South Africa, and fairly extensively in East Africa. Economic reserves are concentrated at Mokama, Zaire, and at Gifurwe, Rwanda.

Deposits of vanadium (a rare element used to toughen steel and to increase its shock resistance) are widespread in southern Africa, but the concentrations are rarely sufficient to make them commercial, except in South Africa.

Africa is seriously deficient in reserves of molybdenum, a metal most needed in the production of steel; some is found in Morocco and in South Africa.

Nonferrous base metals. Most of Africa's copper is contained in the Central African Copperbelt, stretching across Zambia and into the Shaba area of Zaire. In 1970 Zambia was estimated to have 30,000,000 tons and Zaire 20,000,000 tons of contained copper, out of a total for African reserves of almost 48,000,000 tons. Accompanying minerals vary with the geological layer, but cobalt dominates.

Outside the copper belt a number of countries possess lesser but still significant reserves of copper. Elsewhere in southern Africa, Rhodesia has 475,000 tons of copper, South Africa has known reserves of almost 1,000,000 tons of copper metal, and in South West Africa (Namibia) more than 500,000 tons of copper are associated with the lead–zinc ores of Tsumeb. Botswana also possesses more than 30,000,000 tons of proved copper reserves, while, to the north, Mauritania has the largest reserves of West Africa, 460,000 tons of copper metal. In East Africa, Uganda has a reserve of more than 200,000 tons of metal. In North Africa the ore is found in Algeria, Morocco, and the Niger Sahara. Only Nigeria and Zaire contain tin reserves of any significance.

Lead and zinc

Although it is difficult to consider Africa's reserves of lead and zinc separately, of the two, lead ores are considerably more widespread. Total lead reserves are about 4,000,000 tons. North Africa is the largest traditional producing region, but most Moroccan ores are approaching exhaustion, and the major remaining reserve in the Oued Moulouya River Basin contains only 300,000 tons of metal; Algeria has more than 400,000 tons, and there are more than 5,000,000 tons of ore in Tunisia. South of the Sahara, the largest reserve is in South West Africa at Tsumeb.

African reserves of zinc metal are estimated at 6,000,000 tons; they are located along the Moroccan–Algerian frontier, in the copper belt of Zaire, in Nigeria, in Zambia, and in South West Africa.

Africa possesses about 27 percent of the world's reserves of bauxite, the chief aluminum ore. Virtually all of this occurs in a major belt of tropical laterite stretching 1,200 miles from Guinea to Togo. The largest reserves are in Guinea (estimated to contain over 3,500,000,000 tons of bauxite); Cameroon, with about 1,000,000,000 tons, has the second reserve in Africa. The other major West African reserves, small by comparison, are those of Ghana (500,000,000 tons) and Sierra Leone (about 10,000,000 tons). Reserves also exist in Mali and in Malawi.

Ninety percent of the world reserves of cobalt can be

found in a single African country, Zaire, where an estimated 750,000 tons of cobalt metal is located. A continuation of the geological formation into Zambia gives the country a reserve estimated at 375,000 tons of cobalt content. The only other deposit of any significance is found in Morocco, but it is now severely depleted.

The titanium ores ilmenite and rutile are widely distributed in Africa but are rarely considered as minable reserves. The chief of these is the Sherbro deposit in Sierra Leone, where 3,000,000 tons of proved and 30,000,000 tons of probable ore containing titanium are present—the proved deposits being second only to the 4,000,000 tons found in Australia. Elsewhere along the West African coast, ilmenite is found in the black sands at the mouth of the River Saloum, Senegal, and near Grand Lahou, Ivory Coast (400,000 tons). In South Africa, the Morgan Bay deposit, containing 25,000,000 tons of titanium oxides and heavy minerals, is the only commercial reserve. Other reserves include those of the Nile Delta and on the Somalia and Malagasy coasts.

Almost all Africa's antimony resources lie in the Murchison Range, South Africa, where reserves estimated at 280,000 tons of metal are the second largest (after Bolivia) of any non-Communist country. The major concentrations of beryllium are in Madagascar, Mozambique, Zaire, Rhodesia, and Rwanda. The principal sources of cadmium are in South West Africa and Zaire. Deposits of mercury are restricted to North Africa, notably to Tunisia and, more particularly, Algeria.

Gold reserves

Precious metals. Gold and allied metals are widely disseminated, reaching their greatest concentrations in South Africa, where reserves of 500,000,000 ounces of gold probably constitute about 50 percent of the world total. Gold is also obtained in Rhodesia, in the Zaire belt, and in Ghana. There are numerous alluvial sources of gold in Burundi, Ivory Coast, and Gabon.

South Africa possesses the most important deposit of platinum in the free world, amounting to an estimated 200,000,000 troy ounces of metal. Silver reserves of the continent are not important, amounting to about 2 percent of the world's reserves, and are significant only through their association with other commercially interesting minerals. The major sources are in Zaire, South West Africa (Namibia), Morocco, and South Africa.

Africa contains a major share of world resources of tantalum (80 percent) and of niobium (a steel-gray metallic element resembling tantalum in its chemical properties and used in alloys; 15 percent). African reserves of niobium total almost 1,400,000 tons and tantalum 13,000 tons; 60 percent of the niobium and 80 percent of the tantalum are found in Zaire. Further reserves are located in Nigeria and Kenya.

One of Africa's many sources of zirconium (a metallic element resembling titanium chemically) is the Jos Complex, Nigeria. Greater reserves, however, are contained on the Senegal coast, on the east coast of South Africa, at Sherbro (Sierra Leone), and in the Nile Delta.

Another rare metal of which Africa contains a majority of world reserves is germanium, concentrated in Zaire and South West Africa. Africa possesses the second largest deposits of lithium in the world. Africa's largest deposits are found in Rhodesia.

Nonmetallic deposits. *Clay minerals.* Clays are widespread and are found in North Africa, where brick and pottery clays occur in Algeria and in Morocco; in West and Central Africa, where clays are located in Togo (ceramic); in the Central African Republic; in the Ivory Coast (ceramic); and in East and Southern Africa.

Kaolin (china clay) occurs in Algeria. Outside North Africa, it is widespread. In West Africa it occurs most notably on Nigeria's Jos Plateau, as well as in Mali, in Ghana, and in Guinea. Similar deposits occur in Central and East Africa, as well as in southern Africa.

Bentonite (a clay formed from decomposed volcanic ash, which is able to absorb large quantities of water and to expand to several times its usual size) is found in the Moroccan Atlas, in Tanzania and Kenya, and in South Africa. The continent's principal reserve of fuller's earth (an absorbent clay) is in Morocco.

Economic mica deposits occur, among other countries, in southern Africa (South Africa, Rhodesia, and Tanzania) and in Madagascar.

Sulfur and salt. Africa has none of the world's major reserves of sulfur. It reaches economic concentrations only in South Africa's Witwatersrand, in Zambia's Copperbelt, and in Morocco. Large quantities of sodium deposits remain to be evaluated. Sodium chloride is the principal salt, the largest deposit being in the Danakil depression of Ethiopia. The principal sources of salt in Africa, however, are inland or coastal basins, from which it is extracted by evaporation. Major coastal reserves of this type lie along the North African Mediterranean coast and along the Red Sea and Indian Ocean coasts of East Africa and Madagascar. Inland, the chief reserves are in the Oran *sebkha* ("salt pan"), in Algeria; in Botswana around Lake Makarikari; and in Uganda.

Natron deposits

The second sodium mineral is natron, or sodium carbonate. This is more limited in occurrence, but Africa contains several significant deposits. It is found in Lake Magadi, Kenya, and in Lake Natron, Tanzania, as well as in West Africa, where beds have been deposited from the waters of Lake Chad.

Phosphates and potash. North Africa has been a traditional exporter of phosphates, and West Africa has large reserves. Morocco alone possesses the vast reserve of 21,000,000,000 tons of phosphate, and Algeria and Tunisia between them possess a further 3,000,000,000 tons. To the east, phosphate-bearing sediments outcrop on the Red Sea coast. Egyptian reserves total about 200,000,000 tons. In the western Sahara, total possible additional reserves in Rio de Oro in the Spanish Sahara are estimated at 2,000,000,000 tons. Togo has about 120,000,000 tons and Senegal 40,000,000 tons. The Thiès deposit in Senegal is of particular interest in constituting the world's only source of aluminum (as opposed to calcium) phosphate, of which there is more than 50,000,000 tons. Other phosphate deposits occur in Tanzania (10,000,000 tons), Uganda (180,000,000 tons), and Malawi (18,000,000 tons).

The potash deposits in the People's Republic of the Congo are believed to be among the largest in the world. The other large reserve (150,000,000 tons of ore) is in Ethiopia.

Refractory substances. In Madagascar, Africa possesses the world's largest known accumulation of flake graphite deposits. Continuations of these deposits in Mozambique and southeast Kenya contain further reserves of graphite; 15,000 tons are also known to exist in Malawi.

While deposits of low-grade sand suitable for construction and engineering work are widely distributed, reserves of sands with a sufficiently high silica content for glass manufacture are more localized. Deposits exist in West Africa (the Ivory Coast, Liberia, Nigeria, and Ghana), East Africa (Uganda and Tanzania), and South Africa. Glass sands are also found in Egypt.

Kyanite (cyanite), a mineral aluminum silicate, used as a refractory, occurs most typically in southern Africa. Apart from South African reserves, there are deposits in Kenya, Malawi, Ghana, Cameroon, and Liberia.

Abrasive substances. Of the abrasive substances, industrial diamonds are most closely associated with Africa. The continent contains 99 percent of the total Western-world reserves. The stones are found in a number of major belts south of the Sahara. The principal known reserves of diamonds in their primary form are in the South African Vaal belt. Elsewhere in Africa, primary deposits are found in Tanzania, Botswana, and Lesotho.

Industrial diamonds

Another major belt of diamondiferous rocks encircles the Congo Basin and includes the world's largest deposit, located in Zaire, which is estimated to contain 85 percent of Western reserves of industrial diamonds. The same belt has secondary deposits that occur elsewhere in that country, as well as in the Central African Republic and Angola. In West Africa, known reserves are located primarily in alluvial gravel fields. They are found in Sierra Leone, Guinea, Ivory Coast, Liberia, and Ghana.

A considerable proportion of the world reserves of co-

rundum (a common mineral, aluminum oxide, notable for its hardness) is located in southern Africa. The principal deposits are in South Africa, Mozambique, and Rhodesia.

Pumice is found in areas of volcanic activity such as the Atlantic islands, the coastal Atlas of northeastern Morocco, and in the East African Rift, notably in Kenya, Tanzania, and Malawi. Joint reserves constitute, however, only a small percentage of the world total.

Building materials. Reserves of building materials are characterized by their wide distribution, to such an extent that the commercial status of such deposits depends more on their location relative to areas of development than on their extent and quality. While almost all African countries have reserves of building materials, knowledge of such reserves is so strictly related to the country's present level of development that only some are discussed here, and no meaningful estimate of the size of reserves can be made.

Granite is located in Morocco and Nigeria, and there are vast reserves in Upper Volta. Quartzite (a granular rock, consisting essentially of quartz) is important as a building stone in Uganda and in Zaire. Dolerite (a coarse-grained variety of basalt) is produced in South Africa, and basalt, which is crushed for use in road construction, in Senegal. Marble is found in Mali, Togo, Nigeria, and South Africa.

Limestone is important because of its use in the cement industry. Deposits are fairly widespread. North Africa is a particularly important source. In West Africa a belt of limestone runs from the Central African Republic to the Atlantic coasts, with major outcrops in northern Nigeria, Niger, Upper Volta, and Mali. Elsewhere there are deposits in Togo and Ghana. East African deposits include those in Kenya, Tanzania, Uganda, and Zambia; there are also deposits in South Africa.

Gypsum deposits

North Africa possesses major reserves of gypsum on the Mediterranean coast, as well as in outcrops along the Gulf of Suez and the Red Sea. Somalia has a reserve of 30,000,000 tons—one of the largest known deposits. East Africa and Madagascar have further reserves, and in southern and West Africa superficial deposits are particularly important—for instance, north of Nouakchott, Mauritania.

Gem minerals. Many of the major deposits of the most important commercial gem mineral, diamond, have already been described in the discussion of industrial diamonds. One major deposit, that of South West Africa, however, consists almost entirely of gem diamonds.

There is no other gem mineral in Africa comparable in importance with these diamond reserves, but deposits of a number of such stones are found, especially in southern and eastern Africa, where diamond fields contain beryl, garnets, amethyst, rose quartz, topaz, opal, jasper, emeralds, and other stones. Madagascar contains the largest deposit of garnet. Tourmaline is found in Madagascar and South West Africa. Agate is particularly associated with the volcanic areas of eastern and southern Africa, and malachites with the Shaba (formerly Katanga) Copperbelt, while sapphires are found with diamonds in Ghana.

Miscellaneous deposits. Africa contains no major world deposits of talc, but the mineral is found in Morocco, The Sudan, Rhodesia, and South Africa. Reserves of asbestos are much more important, and southern Africa has a number of deposits of world significance.

Major deposits of fluorite (a common mineral, calcium fluoride, used as a flux in metallurgy), constituting 7.5 percent of world reserves, are particularly associated with deposits of lead and zinc. In South Africa the chief deposit is in the Transvaal. North African reserves lie primarily in Tunisia.

Africa produces only about 2 percent of the world supply of diatomite (a fine siliceous earth, used as an abrasive); the most important deposit is in Kenya.

WATER RESOURCES

Freshwater derives firstly from rainfall that has run off the land to collect on the surface in rivers and lakes and secondly from groundwater that results from rainfall infiltrating through and then collecting in the soil. Generally speaking, the seasonal distribution of river flow reflects the seasonal rainfall pattern; the amount of groundwater entering the river channels during the dry season is comparatively small. Important modifications in the flow of some rivers are caused by the presence of large lakes and swamps, which act as natural storage reservoirs.

Natural storage reservoirs

Surface water. Although the surface area of Africa is about 21 percent of the Earth's land surface, the combined annual flow of African rivers is only about 7 percent of the world's river flow reaching the oceans.

Northern Africa's few perennial rivers originate in the mountains of the Maghreb; their water is used extensively for irrigation. The large number of *wādīs* (watercourses) to be found throughout the Sahara and the eastern Mediterranean coastal lands become filled with water as a result of the rare and erratic storms that occur over mountainous areas; otherwise, however, they remain dry.

From the relatively well-watered areas of western and Equatorial Africa, the Sénégal, the Niger, the Logone–Chari, and the Nile rivers flow through the drier inland zones. Of these, the Niger River, originating in the Fouta Djallon, in Guinea, is retarded in the lake and swamp area south of Timbuktu in Mali, and the Logone–Chari feeds Lake Chad.

The Nile, the world's longest river, receives more than 60 percent of its water from the highlands of Ethiopia, although its source is much farther south in the mountains of Burundi. Since the completion of the Aswān High Dam (opened 1970), only a small proportion of the total river flow reaching Egypt enters the Mediterranean Sea.

A number of rivers flowing in a more or less southerly direction into the Atlantic Ocean drain the southern part of West Africa. Many flow rapidly over bedrock before entering the coastal plains, draining into the system of lagoons and creeks along the coast. During the dry season the upper reaches of these rivers are without water; but in Guinea, Sierra Leone, and Liberia, where the dry season is comparatively short, the rivers flow throughout the year.

In the well-watered western part of Equatorial Africa the total average annual flow of the Congo River is 42,300,000,000,000 cubic feet. River flow at the lower end of the basin has two maxima: one that corresponds with the rainy season north of the Equator, the other with the rainy season that occurs when it is summer in the Southern Hemisphere. The waters in the lower reaches of the river are slightly acid after traversing the large swamps situated in the centre of the basin.

Rift Valley lakes

East Africa's many lakes stretch along the Rift Valley from the Red Sea to the mouth of the Zambezi. Evaporation from most of them exceeds their surface rainfall, and in consequence their outflow is less than the quantities brought in annually by their tributaries. They often govern river flow by acting as natural storage reservoirs —decreasing the flood flow and increasing the dry-season flow. A number of the Rift Valley lakes are situated in closed basins and contain high percentages of dissolved salts. The largest of these are Lakes Rudolf, Natron, and Eyasi.

Rainfall over much of southern Africa is small, and the majority of the rivers originating there have an intermittent flow. Some large perennial rivers (*e.g.*, the Okavango, the Zambezi, and the Orange) flow from areas of abundant rainfall into the drier zones.

Groundwater. The conditions under which groundwater is found and the quantity and quality of groundwater reserves are closely related to geological structure. Large inland depressions in Africa's basement rock, having been filled with sedimentary layers of continental origin, sometimes form important groundwater reservoirs, notably those in the Taoudeni–Niger region, in the central Sahara between the Atlas and Ahaggar mountains, in the Libyan Desert, and in Chad, the Congo Basin, the Karoo area of South Africa, and the Kalahari.

The East African plateaus usually contain little or no



Great Fish River Canyon In the Namaqualand region of South West Africa.
John Rushmer—Ostman Agency

quantities of groundwater, and aquifers (geological formations containing water)—generally of local importance—are found only in humid areas where the crystalline rock is weathered or fractured.

The chalky shales (rocks of laminated structure formed by the consolidation of clays) and dolomitic limestones (*i.e.*, those containing calcium magnesium carbonate), which here and there cover the basement rock, may contain important aquifers; those in Zambia and the Republic of South Africa are major sources of water.

In the Sahara a rock stratum called the Continental Intercalary, which dates to the Cretaceous Period and which includes Nubian Sandstones, is the most important water-bearing layer. It extends over very large areas and reaches a thickness of more than 3,000 feet; in Egypt and Algeria it provides artesian water. In The Sudan it sometimes lies directly on the Precambrian rock and contains underground water layers of local importance. Overlaying the Continental Intercalary but generally separated from it by a thick marine deposit is a layer called the Continental Terminal, which is the second-largest aquifer in this area. Both these layers contain "fossil" water —that is to say, water that entered the layers when rainfall in and around the Sahara was much more abundant than today. Near the surface, aquifers are found in such recent deposits as alluvial deposits and sand dunes.

"Fossil" water layers

In the coastal areas of Senegal, Ivory Coast, Ghana, Togo, Dahomey, Nigeria, Cameroon, Gabon, the People's Republic of the Congo, Angola, Mozambique, East Africa, and Madagascar, aquifers are found in sandstone, limestone, and sand and gravel sediments. Intensive exploitation, however, may result in saltwater intrusions.

The Jurassic limestones of the mountainous area of the Maghreb are much more abundant in water sources than are dolomitic limestones. Around the cape in South Africa, sandstones and limestones contain very little water.

Yields from aquifers with good porosity, such as coastal sedimentary rocks or alluvial deposits, vary from a few cubic feet per hour in the fine-grained sands found in many parts of the continent to 35,000 cubic feet per hour in the coarse gravels of the Nile Delta. Capacity of wells in the Continental Terminal is generally somewhat lower, although 280 cubic feet per hour have been reported. The Continental Intercalary and the Karroo formations can also deliver moderate to good yields.

In northern Africa, limestones containing many cracks and fissures may yield hundreds of cubic metres of water per hour, while, in Zaire, Zambia, and South Africa, large yields are drawn from dolomitic limestones.

The harder sandstones, sandy shale, and quartzites of Precambrian and Paleozoic age are not generally very porous, and water is obtained only from fractured or weathered deposits. In West Africa, Angola, and Tanzania, wells in these formations produce only a few cubic feet per hour. Crystalline and metamorphic rocks are almost impermeable except where fractured or weathered. Volcanic rock, especially the basalts, may yield up to 1,060 cubic feet per hour.

Densities of mineral solutions

Most of the exploited groundwater is generally fit for consumption, for the dissolved minerals in water from shallow wells, particularly in the sandy aquifers of West Africa, is quite low. Groundwater from deeper marine layers, however, such as occurs in parts of North Africa, Mozambique, Ethiopia, and South Africa, may have a high content of dissolved salts. In moist tropical countries, water from Precambrian rocks generally contains only small amounts of dissolved solids. In the volcanic areas of East Africa, groundwater may have so high a content of fluorine as to make it unfit for human consumption. Here and elsewhere in Africa, hot (possibly medicinal) springs with high mineral properties occur.

Availability for human use. The pronounced seasonal character of rainfall and the fact that many rivers stop flowing during the dry season necessitate the development of groundwater for human use, whether for husbandry or industrial development.

Large-scale irrigation has, until now, been practiced mainly in northern Africa, Egypt, The Sudan, the Republic of South Africa, Mali, Rhodesia, and Mozambique. Medium-scale irrigation projects exist in Senegal, Somalia, eastern Africa, and Ethiopia, and in many countries (such as Ivory Coast, Upper Volta, Kenya, Nigeria, Ghana, and Zambia) medium- to small-scale projects are being undertaken.

More than 50 river basins are shared by two or more countries, and the development of their resources requires the cooperation of the basin states and several intergovernmental agencies—such as the Organization for the Development of the Senegal River, the Niger River Commission, and the Chad Basin Commission.

Several large reservoirs have been built in recent years, such as the Aswān High Dam, Roseires, and Khashm al-Qirbah reservoirs in the Nile Basin, Kainji on the Niger, Akosombo on the Volta, and Kariba on the Zambezi, while others (such as Cabora Bassa on the Zambezi in Mozambique; Kossou on the Bandama in Ivory Coast; and Kafue on the Kafue in Zambia) are under construction or being planned. At a number of the new man-made lakes, research centres have been set up to study resettlement problems, the full utilization of the new ecological conditions, and the control of health hazards that sometimes occur.

BIOLOGICAL RESOURCES

The biological resources of Africa consist of an immensely varied vegetational cover, a vast insect life, and a diverse animal life that populates a wide variety of ecological environments, including the biological resources like bacteria, earthworms, and termites that influence soil and soil fertility; hydrobiological resources include the fisheries of lakes and rivers, as well as the oceans surrounding the continents.

Botanical resources. Out of the total land area of the continent, forests cover about 20 percent, savannas about 40 percent, and deserts about 40 percent.

The forested area cannot be ascertained exactly, as original forests are progressively being converted to farming areas, and few governments have undertaken comprehensive land utilization surveys. The tropical forests consist of both rain forests of virtually evergreen trees (found in the equatorial belt in areas with an annual rainfall of more than 100 inches, as well as at high altitudes) and tropical moist semideciduous forests (in which many species of trees shed their leaves at some time of the year) in areas with less than 70–100 inches of annual rainfall. Forests of this description, in the rainfall range between 50 and 150 inches, are mainly found in the equatorial belt. The tropical high forests contain most tree species commercially in demand.

A large proportion of the land surface of Africa bears vegetation in which grass is an essential feature. Grasses occur only rarely in forest environments, where they are usually tall and fibrous and seldom very palatable to livestock. In certain savanna areas, however, fire, flooding, or soil conditions may produce open grasslands free from woody vegetation. The broad physiognomic vegetation types and their grass cover may be divided into the six following groups: (1) forest, dense, woody growth with virtually no grass cover; (2) woodland, woody growth of varying density with some grass cover; (3) savanna, woody growth of varying density and conspicuous grass cover; (4) grassland, good, dense grass cover, the absence of woody growth or its confinement to watercourses; (5) steppe, very open, scattered woody growth with sparse grass cover and including desert conditions; and (6) undifferentiated types, variable grass cover, much of it ungrazeable for climatic or topographical reasons.

Vegetation types

Natural vegetation, especially grasses—which are basic to the well-being of domesticated livestock as well as to ruminant game species—are of particular importance in those parts of the continent where livestock are economically important. Better management of pastures is required in much of Africa where indiscriminate burning and widespread overgrazing and undergrazing are rife.

The savannas, of which there are some 30 types in Africa, cover more than 5,000,000 square miles and stretch latitudinally across the continent in two broad zones north and south of the Equator, where they are divided by the tropical forests of West Africa and the Congo but join each other in the east. Savanna in North Africa is represented in the Atlas Mountains and foothills.

Arable land and land under permanent crops occupies about 7 percent of Africa's total area. Arable agriculture is largely confined to subsistence farming and is considerably dependent on the wasteful system of shifting cultivation, in which land is temporarily cultivated and then abandoned for a time. Over most of Africa, where arable land is generally allocated through tribal affiliation and not through individually acquired title, the intensity of arable production tends to be low, and only a few agriculturally enlightened countries make strenuous efforts to raise productive levels by selecting newer varieties of seeds and planting materials or by increasing the use of mineral fertilizer.

Individual sites for arable cultivation are usually chosen because they have well-drained soils, because their soils are easy to work with hand tools and are free from large stones, and because they are reasonably near to the homestead.

One tree of economic interest is worthy of special mention. This is the Albida acacia tree of the orchard savannas of West Africa. Unlike almost all other dry savanna woodland trees, whose leaf shedding normally occurs at the onset of the dry season, it appears to have a period of partial dormancy during the rainy season and springs to life only at the beginning of the dry season. At such periods its foliage is much in evidence and—being a palatable leguminous species—is much prized as browse for sheep, goats, and cattle. The smaller leafy branches are frequently fed to stock. The tree flowers and produces fruits, which are harvested, dried, and fed as a protein concentrate to stock at the height of the dry season.

The importance of the Albida acacia tree

Another plant of particular interest is genetically improved maize (corn). Maize is not in itself a balanced food, being deficient in certain amino acids. A recent breakthrough, resulting in an increase of the content of an amino acid called lysine in certain new varieties of maize, may significantly change the value of diets in Africa, which until now have consisted mainly of maize preparations.

Animal resources. *Draft animals.* Although water buffalo, oxen, horses, mules, donkeys, and camels are used primarily as draft, pack, or riding animals in Africa, they also provide milk, meat, hides, or skins.

The water buffalo is an offshoot of the Asiatic buffalo (*Bubalus bubalis*); it arrived in Africa in relatively recent times and is at present found almost exclusively in Egypt, where it may number about 1,800,000. The domesticated African water buffalo is used to cultivate irrigated land (principally in the Nile Delta) and to provide milk and meat.

Because of its intractability and wild nature, the African elephant, unlike the Asian elephant, is not used for draft or haulage purposes.

Horses and ponies are principally represented in Morocco, Algeria, Tunisia, Chad, Ethiopia, Mali, Niger, Nigeria, Senegal, Upper Volta, Egypt, South Africa, and Lesotho; in the early 1970s they were estimated to number approximately 3,800,000 head. Horses are largely used as riding or pack animals and in certain areas for the production of mules. Few are kept in areas where the tsetse fly, deadly to livestock, is present.

In Africa five main types of horses have been defined: the Darfur pony, the Dongola horse, the Ethiopian-Galla horse, the Somalia pony, and the South African horse (a type that includes the Basuto pony). In Northern Africa, types also have evolved as a result of selection and crossing with exotic Arab, Barb (Barbary), and Thoroughbred horses. In the South, Arab and Thoroughbred influence may also be noted.

The distribution of the ass roughly corresponds to that of the horse, except that it also extends into the livestock areas of East and Central Africa.

Mules may be observed in Algeria, Ethiopia, Morocco, Somalia, South Africa, Tunisia, and Egypt. In Africa as a whole, mules number approximately 2,100,000. They provide farm draft power and are used as pack animals and for riding. The ability of mules to perform work in hot, dry climates is superior to that of most other farm animals.

The camel, or one-humped dromedary, is widely dispersed in the drier regions of North and northeastern Af-

Camels

rica. The camel is also found frequently in Ethiopia, Egypt, Algeria, Mauritania, Mali, Niger, northern Kenya, Morocco, Chad, northern Nigeria, Somalia, and Tunisia. In the early 1970s the total camel population of Africa was estimated to be about 9,000,000. Though used principally as a pack animal, it also is used for land cultivation, water pumping, and riding. The dromedary is essentially a bush browser and, if reasonably well fed and watered, may produce five to six kilograms of milk daily, in addition to that fed to the calf. The milk is prized by the camel herders and their families. Camel meat and camel hides find a ready market among Muslim communities.

Oxen are widely used in Africa for agricultural purposes, especially for plowing and cultivation; they are also trained to thresh grain, pump water, and act as pack animals. Bullock (male castrated oxen) plowing is well-developed in the countries of North Africa, in Ethiopia and The Sudan, and further west in Chad, in northern Nigeria, and in the savanna climatic zone of West Africa. Plowing and cultivation by oxen is also well developed in tsetse-free areas of eastern and southern Africa. Females used for work may also be milked. Work oxen may also be used for meat and to provide hides.

Game animals and wildfowl. The main group of herbivores—which includes the principal game animals—are the African antelopes, which belong to the four following subfamilies of the ox family (Bovidae): (1) the ox-like Bovinae, which are further subdivided into (a) the African buffalo and (b) the twisthorned antelope, including the eland (the largest of all antelopes), kudu, nyala, and bushbuck; (2) duiker, a small, primitive bovid that lives in thickets, bush, and forest; (3) the "horse antelope," further divided into (a) sabre-horned sable, roan, and oryx antelope, (b) the "deer antelope," kongonis, hartebeest, topi, gnu, and blesbok, all mostly inhabitants of the open plains, and (c) the "marsh antelope," waterbuck, lechwe, kob, puku, and reedbuck; and (4) the antelope proper, divided into two distinct tribes, the first of which includes royal, dik-dik, klipspringer, oribi, steenbok, and grysbok and the second, gazelle, impala, springbok, and gerenuk.

Predators and scavengers, important in maintaining a natural equilibrium, are the lion, leopard, cheetah, wild dog, hyena, serval (a long-limbed African cat), wildcat, jackal, fox, civet, mongoose, eagle, vulture, and marabou (a large stork).

Other vertebrates include birds, reptiles, amphibia, and small mammals. There are also many invertebrates.

African wildfowl are too numerous to list in detail, but they include pheasants, of which there are two subfamilies in Africa, and various genera that include peafowl, quail, francolin, partridge, guinea fowl, sandgrouse, duck, goose, snipe, and many others.

Sources of animal products. Cattle provide hides, and sheep, goats, and pigs provide skins. Skins of the Maradi, Sokoto, and Kano red goats from Niger and Nigeria are greatly prized by the Morocco leather trade.

North and south of the tropics, African sheep are covered with wool, but in the tropics they are hairy. In elevated areas, such as Ethiopia, where temperatures are modified by altitude, some sheep may be partially woolled, at least on the back and buttocks. The woolled sheep of North Africa are largely of the woolly Barbary type, which was originally introduced to Africa from the Middle East.

Many countries are attempting to conserve wildlife by refusing export licenses for certain kinds of skins, especially those of the leopard, cheetah, and zebra.

Fish and other aquatic animals. The most useful categories of fish found in the African waters include many species of freshwater fish; important among the marine fish are flounder, halibut, sole, redfish, bass, conger, jack, mullet, herring, sardine, and anchovy. Crustaceans are important for local consumption and for export.

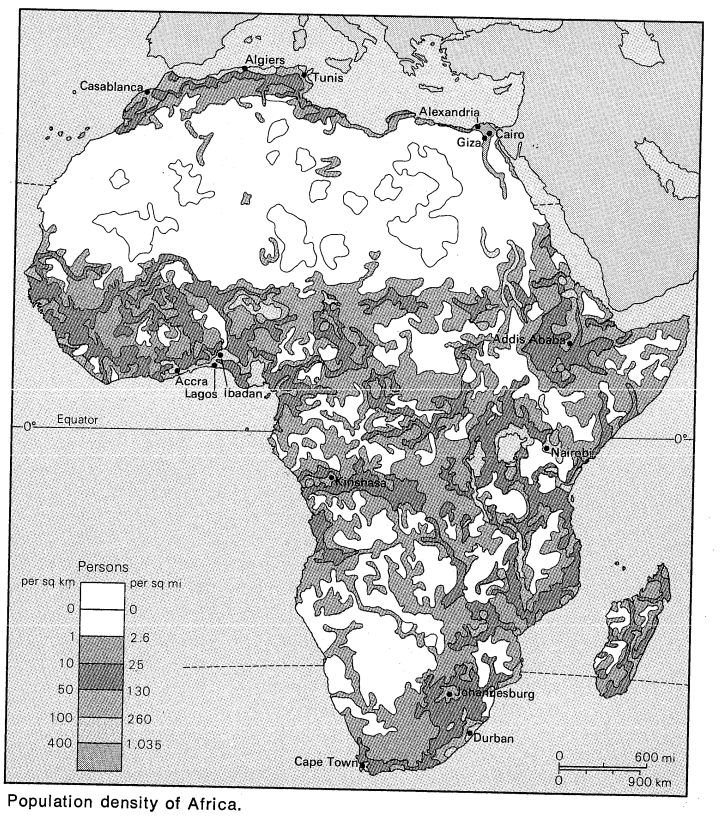
Other economically important aquatic animals and plants include the Cape fur seal, turtles, porpoises, and dolphins, various whales, oysters (for pearls), trochus shells, corals, and sponges. (R.K.A.G.)

V. Human resources

THE HUMAN ECOLOGY

The pattern of settlement. The basic pattern of settlement in Africa may be described as follows.

Africa north of the Sahara is inhabited mainly by peoples speaking Semitic and Hamitic languages. The Hamitic-speaking people are represented, for example, by the Berber of Morocco, Algeria, and Tunisia. They are light-skinned people with dark hair, thin noses, and non-everted lips. In many regions, as a result of culture contact and of intermarriage, the Berber have become largely assimilated to the Semitic-speaking Arabs who migrated into North Africa from Arabia in a number of waves, the first of which occurred in the 7th century AD. The distinctive nature of Maghrebian, or western, Arab culture resulted from this admixture. In the Sahara, such Arab peoples as the Shuwa exist side by side with such Berber peoples as the Tuareg. In the region of West Africa lying south of the desert, known as the Western and Central Sudan, the pastoral Fulani show in their physical characteristics clear evidence of the infusion of lighter-skinned elements into a basic Negro stock.



Population density of Africa.

On the East African coast, contact and intermarriage between Arab settlers and local Bantu Negroes gave rise to the distinctive Swahili culture. Ethiopia and Somaliland are populated by members of the Cushitic branch of the Hamitic-speaking peoples. The movement of Semitic-speaking peoples into the region has, however, occurred in many waves, one of the earliest of which was that of the Sabaeans from Arabia as early as the 7th or 8th century BC. Their intrusion created the conditions for the emergence of the early Aksumite (Axumite) civilization.

The remainder of the continent is dominated by peoples of Negro stock: brown to dark brown in colour, with characteristically woolly hair, broad noses, and everted lips. Four principal though not clearly distinguished subgroups may be identified: the Sudanic peoples of the region extending across the continent south of the Sahara desert, the Guinea Coast people of the forests of West Africa, the Nilotic of East Africa, and the Bantu of central and southern Africa. Many of the Sudanic peoples, such as the Malinke (Mandingo), Hausa, Songhai, and Bornu, are Islāmized, and that religion has also achieved

substantial gains among such Guinea Coast people as the Yoruba of Nigeria and the Temne of Sierra Leone. Much conversion to Christianity has also occurred, most notably in the coastal regions of sub-Saharan Africa.

The Bantu-speaking peoples

The Bantu-speaking people extend over most of the continent south of the Equator. They include such well-known societies as the Lozi in the middle of the Zambezi River region, the Herero of South West Africa, and the Zulu of South Africa. Although originally probably stemming from the same stock as the Guinea Coast Negroes of the Cameroons and east central Nigeria region, the early Bantu homeland was in the woodland region to the south of the Congo rain forests. From there, over the course of several millennia, they spread in all directions, possibly being the first carriers of an iron-age culture into many areas. Throughout the vast region occupied by Bantu-speaking peoples, remnants of earlier pre-Bantu populations survive. The best known of these are the Pygmies of the rain forests and the Bushmen of the Kalahari (desert)—both hunter-gatherers—and the pastoral Hottentots of South West and South Africa. Everywhere the Bantu cultivators expanded at the expense of these earlier societies, either absorbing and assimilating them into Bantu culture, or else relegating them to the most dry or most humid areas.

Finally, in the southernmost parts of the continent, are found white settler populations, largely of Dutch and English origin; smaller communities originating from the Indian sub-continent; and mixed communities, of Hottentot, European, and Asian extraction, known as the Cape Coloured. The Dutch migrations into South Africa occurred over the past three centuries. Originally settling on the coast, the Dutch—or Boers—later moved inland to the high veld region, where a series of military conflicts occurred between them and the Bantu in the 19th century. Other European settlement took place mainly in the 19th century: the British particularly in Natal, but also inland in what is now Zambia and Rhodesia and in the East African highlands; the Portuguese in Angola and Mozambique; and the Germans in South West Africa. The presence of large settler populations has delayed the achievement of self-government by the African people of South and South West Africa, Rhodesia, Angola, and Mozambique, and has resulted today in much bitterness between the races. In North Africa, by contrast, where the extensive settlement of Europeans from France, Italy, and Spain occurred, the growth of Arab nationalism and the emergence of independent states such as Morocco, Algeria, and Tunisia led to the return of many of the colonists to their homelands and political dominance of the indigenous peoples.

Migrations and population movements. There have been many migrations or movements of population within the African continent, from outside into the continent, and from the continent outwards. The major movement within the continent in historic times has been that of the Bantu-speaking people, who, as a result of a population explosion that is not yet fully understood, came to expand over most of the continent south of the Equator. The major movements into the continent—of the Arab settlers in northern Africa, and of European and Asian settlers in southern Africa—have been described. The great outward movement of people was that of the Negroes, particularly from West Africa, to the Americas and the Carribean, which took place during the period of the slave trade from the 16th to 19th centuries. Older estimates that between 15,000,000 and 20,000,000 Africans were transported across the Atlantic have now been revised; a figure of 10,000,000 appears more realistic. While their contribution to the development of the New World was of crucial importance, the effect of the loss of manpower to the African continent was considerable and has yet to be satisfactorily analyzed.

Few major movements of population have occurred in the 20th century, although an extensive settlement of Hausa from northern Nigeria took place in what is now the Republic of the Sudan. Some significant displacements of population have also resulted from situations of conflict, usually involving minority groups different in religion, ethnicity, or culture from the dominant majority. In 1966 the Ibo people of northern Nigeria, for example, returned en masse to their homeland in the Eastern Region, the number of refugees being estimated as in excess of 500,000. Another displacement of people occurred when the government of Ghana expelled perhaps 500,000 aliens from the country in 1969. The 17-year civil war in the southern Sudan created an estimated 800,000 refugees, of whom about 300,000 had fled to surrounding countries. Most were expected to return to their homes after the settlement of March, 1972. Other population movements have been those associated with such phenomena as pastoralism and urbanization.

Population distribution. Africa has about 10 percent of the total population of the world, distributed over a land area representing about one-fifth of the world's total land surface. Such areas as the Sahara and Kalahari, however, reduce the amount of habitable land.

The main regions where there is a high population density are along the West African coast around the lower reaches of the Niger River and the coast from The Gambia to the Congo River; along the east coast from Kenya down to South Africa; on the north coast along the shores of Tunisia, parts of western Morocco, and the Lower Nile area; and in the regions of the East African Plateau around the great lakes and the Ethiopian Plateau. The territories with the highest populations in Africa are Nigeria (55,000,000), Egypt (34,000,000), and Ethiopia (25,000,000). Others with populations of more than 10,000,000 are Zaire, Kenya, The Sudan, Morocco, Uganda, Algeria, and the Republic of South Africa. Madagascar has a population of about 7,000,000.

The largest cities

The largest cities (*i.e.*, those with over 1,000,000 population) include Cairo and Alexandria, in Egypt; Kinshasa, in Zaire; Lagos, in Nigeria; Casablanca, Morocco; and Johannesburg, South Africa. Many other larger cities are seaports along the coasts or central marketing towns, linked by rail or river with a coast. Accra, in Ghana, Lagos, in Nigeria, and Capetown, in South Africa, are examples of seaports, while Ibadan and Onitsha, in Nigeria, Nairobi, in Kenya, and Addis Ababa, in Ethiopia, are examples of inland cities.

FORMS OF GOVERNMENT

The main task facing many African countries is to devise forms of government that will promote internal unity, since almost every African country comprises different communities and ethnic groups that are sometimes hostile to one another. The fact that many of these countries still have a high degree of illiteracy often makes it difficult for inhabitants of rural areas and sometimes of towns to give loyalty and allegiance to the central government instead of to village or tribe, as has been the custom for centuries. Nevertheless, in some areas a sense of supra-tribal identity was already growing up in pre-colonial times, and kingdoms such as Ashanti and Dahomey were developing centuries ago as multi-national states.

Kingdoms, such as Morocco and Ethiopia, have hereditary heads of state. Although these countries have elected houses of representatives, most of the power is nevertheless vested in the ruler. In some countries, however, such as Egypt, Libya, Burundi, and Rwanda, the hereditary rulers have been deposed and their power seized by either military or civilian governments.

When independence was granted to the former colonial territories, the constitutions adopted usually provided for democratic elections and a multiparty system of government of western European types. The democratic multiparty system did not survive long and at present exists only in Sierra Leone and The Gambia.

The prevalence of the one-party state

In many African countries the form of government is that of the one-party state, where the ruling party is absolute and power is concentrated in the hands of the party and especially of its leader, who is usually president and head of government. The justification for the one-party system, it is argued, is that for the economic development of the country only one party is necessary and that an opposition party works against the interests

of the people. An opposition party is also said to encourage tribalism and secessionist tendencies, since sooner or later political parties tend to become identified with certain regional interests. Disagreements or differences of opinion, it is felt, should be within the party itself; in effect, in some countries elections are held in which rival candidates both represent the single ruling party.

The danger inherent in a one-party system is that, when there is unemployment or dissatisfaction, the citizens of the country see no way of changing their government except by revolution or a coup d'etat. This is in fact one of the main reasons for the many military coups d'etat that have taken place in Africa in the past few years and that have left the world with an impression that political instability prevails in the continent. In the decade of the 1960s, for example, there have been military interventions in at least 15 African countries. While the existence of a one-party system is not the only reason that coups d'etat take place, in many instances it provides fertile ground for military intervention. Military governments frequently rule with the aid of civil servants. The civil service thus constitutes a stabilizing force in African countries that are subjected to frequent changes in their forms of government.

In some countries, traditional rulers, especially when cooperative with the central government, are the mainstay of local government in their territories, continuing the role given them during colonial government as Native Administration authorities.

In southern Africa, in particular, several countries are still ruled by white minorities that have refused to give way to democratic rule. Among these countries are the Republic of South Africa, South West Africa, and Rhodesia. Angola and Mozambique are still governed by the Portuguese as provinces of metropolitan Portugal, despite the opposition of militant African liberation movements that have seized control of significant areas.

In sum, the best utilization of the potential of the human resources of Africa will constitute one of the major problems of the final decades of the 20th century.

(D.S.H.W.N.)

VI. Political geography

EXPLORATION AND SETTLEMENT

Although Africa was the first continent whose entire coastline was known, much of its interior remained unknown to the rest of the world until the 19th century. Until a century after Vasco da Gama, in 1497 to 1498, had sailed round the Cape of Good Hope to India, the extreme northeast coast of Europe was not known to western Europe, and the coasts of Asia, North America, and Australia were not fully known until the 19th century. It is also noteworthy that Africa is the only continent shown in full on the reconstruction of Herodotus' world map, drawn in about 450 BC.

The explorers of antiquity

Phoenicia and Carthage. The Phoenicians were probably the first maritime explorers of Africa. From their coastal citadels of Tyre and Sidon they sailed to many parts of the Mediterranean and founded colonies—the most famous of which was Carthage, which was established on the North African coast about 820 BC and which later surpassed its mother cities in size, power, and influence. The Phoenicians (Carthaginians), indeed, did more than any other people of the ancient world to make known the shape and size of the continent of Africa.

About 480 BC a Carthaginian naval expedition set out under the command of Hanno to explore the Atlantic coast of Africa as far as Rio de Oro, where a trading post was established at Cerne. From Cerne the expedition sailed on perhaps as far as the Sénégal and Gambia river basins and beyond to Sierra Leone.

Greeks and Romans. The Greeks had founded Cyrene on the North African coast in about 630 BC. Diogenes, a Greek sailor who had been blown onto the East African coast, wrote an account of a journey inland to two great mountains where, he thought, the Nile had its source. In the 2nd century AD, Ptolemy, the Greek geographer and mathematician, showed the "Mountains of the Moon" on his maps. Greek and Roman knowledge of Africa was essentially confined to the Mediterranean coastal strip. The Greeks, mainly seafarers, were not very interested in terrestrial exploration, the Romans rarely ventured farther than was required by the exigencies of military conquest, though at least two minor expeditions across the Sahara took place—those of Septimus Flaccus and of Julius Maternus in the 1st century AD.

The Arabs. The invasions of North Africa from Arabia, which began in 640 AD, extirpated almost all vestiges of Greek and Roman influence and eventually unified the diverse peoples of the region. These invasions also carried the Arabs right across the Sahara to the south, and close links were developed between North Africa and Africa south of the Sahara. These links, however, did not go beyond the northern fringes of the tropical rain forests, where impenetrable jungle and unknown dangers halted the Arab advance. From the 10th century on, Arab travellers brought to Europe the first definite knowledge of trans-Saharan Africa as well as of the east coast, where the Arabs established themselves from Zanzibar to Sofala.

The Portuguese. The Portuguese sent out by Prince Henry the Navigator (1394–1460) were the first to undertake systematic voyages of discovery southward along the African coast. These and later Portuguese voyages were made in the hope of establishing contact with the rich gold-producing centres, which were known to exist in West Africa, and also of discovering a sea route to India around the tip of Africa, which would enable the Portuguese to participate in the rich trade of the Indies.

By 1434 Cape Bojador had been rounded, and Senegal and Cape Verde were reached in 1446. In 1481 the Portuguese built the fort São Jorge da Mina (Elmina) on the Gold Coast. The exploration of the African coast was carried past the mouth of the Ogové River by Diogo Cão, who, in 1482, discovered the mouth of the Congo. In the final years of the 15th century, another Portuguese emissary, Pêdro da Covilhã, visited most of the Arab settlements on the East African coast as far south as Sofala and travelled to Ethiopia.

By the 16th century, however, the Portuguese were losing ground to the English; by about 1553, English ships were venturing into the Gambia region and had sailed as far as the Gold Coast (contemporary Ghana). By 1595 the Dutch had also appeared on the West African coast and were establishing themselves there.

The exploration of the Niger

Ventures into the interior. The exploration of the interior of Africa is largely the story of the search for the sources of some of the world's greatest rivers. Systematic exploration may be said to have begun with the establishment of the African Association, a group of English scientists and scholars who resolved in 1788 to explore the course of the Niger River to the sea. Under the association's auspices, and later under commission by the Colonial Office, Mungo Park, a young Scottish ship's doctor, made two attempts, in 1795 and 1805, to sail down the Niger to the sea. Although he was unsuccessful in reaching the coast to the southeast of his starting point near Timbuktu, he managed on the second attempt to sail as far as Bussa, 1,200 miles downstream, where he lost his life. Several more expeditions finally culminated in the successful attempt by two Englishmen; Richard Lander and his brother, John, who in 1830 were to paddle down the river from Bussa to the sea.

The source of the Nile. For more than 20 centuries the source of the Nile had excited the imagination of man; yet its discovery had proved extremely baffling. In 1856 the British Royal Geographical Society sent two explorers, Richard Burton and John Speke, to find the great lakes that were rumoured to exist in the African hinterland. They found Lake Tanganyika, and in 1858 Speke discovered Lake Ukerewe, which he named Victoria Nyanza. He also heard of a great river that flowed from the lake at Jinja and assumed that this could only be the Nile. In 1860 he returned to Africa with another English explorer, J.A. Grant, and confirmed that Lake Victoria was indeed the principal source of the Nile.

The Zambezi and the Congo. Another 19th-century explorer, David Livingstone (*q.v.*), went to South Africa as a medical missionary; but his contact with the slave

trade changed his interest from missionary work to exploration. He was convinced that the only way to destroy the slave trade was to open up the interior of Africa to normal trade, and he vowed to break Africa's isolation from the rest of the world or perish in the attempt. In 1853 he set out to trace the course of the Zambezi to its source; after great hardship he succeeded in reaching the Portuguese territory of Angola on the Atlantic coast in 1854. On his return downstream to the east coast, he discovered Victoria Falls. He returned to England after 16 years in Africa, having established that much of the interior of the continent was not a sandy desert but a well-watered country, with large tracts of fertile soil covered with forests and valleys. In 1858 Livingstone returned to Africa and discovered Lake Shirwa, the southern end of Lake Tanganyika and Lake Mweru.

Subsequently, Henry Morton Stanley, a Welsh explorer, continued the explorations that Livingstone had begun. On a major expedition into the interior, which lasted from 1874 to 1877, he sailed around Lake Victoria and Lake Tanganyika and then moved on to the Lualaba River, following it downstream until he had traced the course of the Congo River to its mouth at Boma, on the west coast of Africa.

Colonization and settlement. Although European countries had established trading stations at many points on the African coast, there had been no thought initially of settlement or colonization. But when the industrial revolution, with its demand for raw materials and markets, gained momentum in western Europe, interest began to develop in African colonization and settlement.

By 1884 there had begun a veritable scramble for African territory, which ended in the parcelling up of the continent into a bewildering mosaic of European colonial possessions. The scramble resulted in frequent disputes between the European powers, as their governments manoeuvred to obtain special rights and territories in various parts of Africa. This situation led to the convening of the Conference of Berlin (1884–85) to agree upon the orderly partitioning of Africa in general and to decide the future of the territory of the Congo in particular. The "balkanization" of Africa that resulted had long-lasting effects harmful to its unity and an even more inhibiting effect on African economic development. The Conference of Berlin, however, and subsequent colonial conferences, confirmed the claims of the various powers to their respective spheres of influence.

The Berlin conference

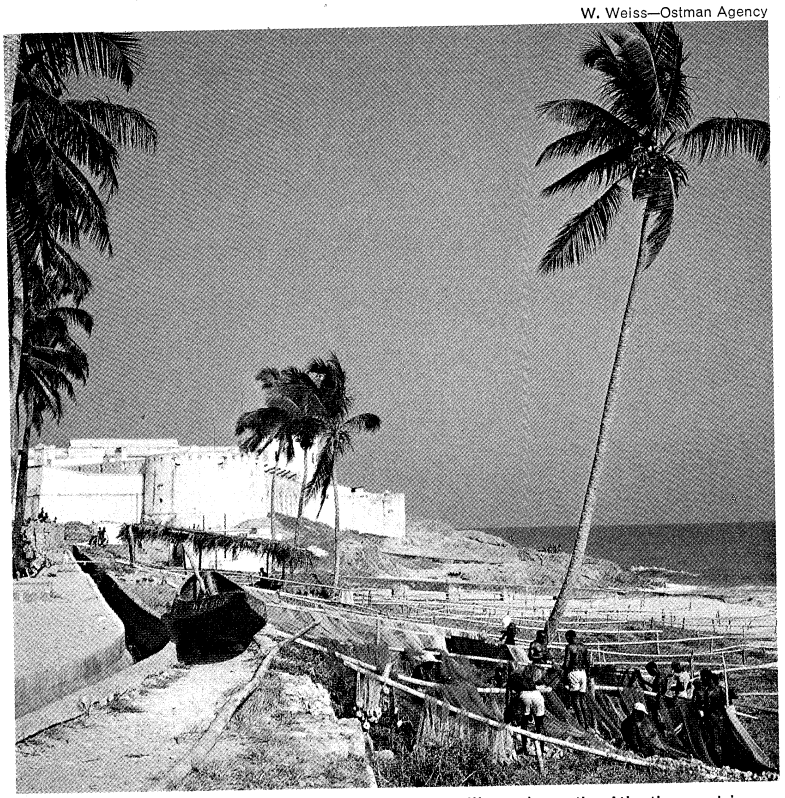
In the 17th century the Dutch had begun to challenge Portuguese supremacy in the East Indies trade and in 1652 sent Jan van Riebeeck to establish a base at the Cape of Good Hope as a supply depot for the provisioning of ships on the long voyage from Europe to India. Although the Dutch had no intention of establishing a permanent colony, the Dutch East India Company began farming to provide fruits and vegetables for their ships. In time, the settlers pushed inland and formed permanent settlements; and in the 1790s these were established as independent South African republics. After the British had established themselves at the Cape, tension developed between the British and the Dutch (Boers) in South Africa; after a series of conflicts, these elements were combined in the Union of South Africa in 1910.

Although the French had a strip of coast in Gabon and settlements in Senegambia, their main possessions were in North Africa, where in the early 19th century they had acquired considerable territory in Algeria, primarily for strategic and geopolitical reasons.

French strategy during the period of colonization was to drive east from their coastal base in Senegal, blocking British expansion north from Freetown, Accra, and Lagos. Simultaneously, the French sent expeditions to the north from the Congo River, and the two pincers met on the shores of Lake Chad, effectively containing the British. Further attempts to extend French territory east across Africa were, however, prevented by the British in 1898, at Fashoda, in the southern Sudan.

The British expanded inland from coastal settlements. In the mid-1880s Cecil Rhodes, a British financier and empire builder, pushed British power north from Cape Colony and succeeded in interposing British territory between Portuguese Angola and Mozambique, thus thwarting the Portuguese ambition of linking these two territories. Britain also established a protectorate over the vast territory of what is now Kenya and Uganda and moved into the Sudan to link up with Egyptian soldiers under British command. In West Africa, British posses-

W. Weiss—Ostman Agency



Castle dating from the days of European exploration still stands on the Atlantic coast in Ghana.

sions expanded considerably as a result of the activities of the famous colonial adventurer, Lord Lugard, who established British control over what is now Nigeria.

Although the Germans did not establish any colonies until the latter part of the 19th century, they quickly acquired the territories of South West Africa, German East Africa (Tanganyika, Rwanda, and Burundi), Togoland, and Cameroun between 1883 and 1885. They were to lose all these territories as a result of World War I.

At the beginning of the 19th century most colonial possessions had been coastal enclaves, but by the beginning of the 20th century the whole of Africa, except Ethiopia, Liberia, and nominally Egypt, was under colonial rule; it was to remain so until after World War II, when the era of African independence broke upon the international scene.

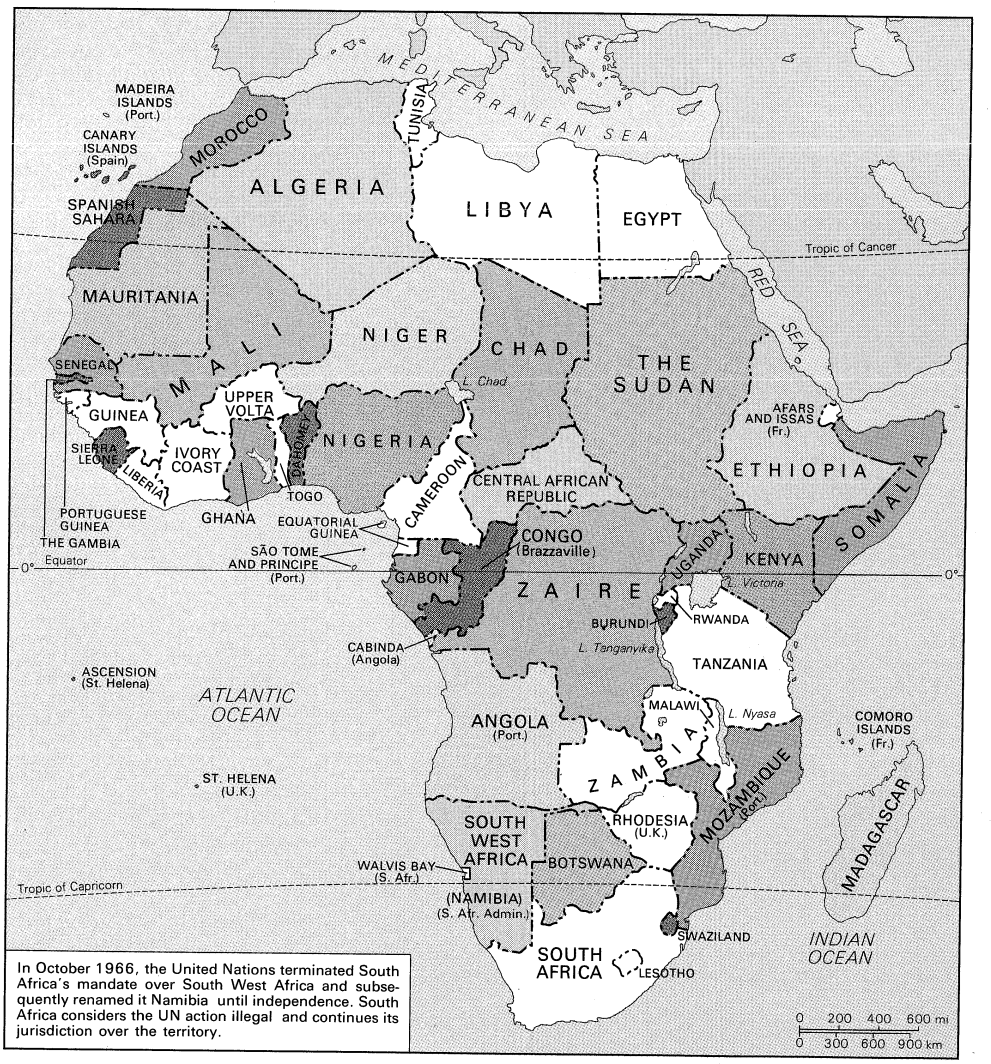
CURRENT POLITICAL GEOGRAPHY

In 1920 every square mile of the African continent—excepting the independent states of Ethiopia, Liberia, and South Africa—was under colonial rule or protection or was claimed by one or the other of the European colonial powers. By the early 1950s, however, African nationalism and the rise of independence movements were in flood, a development that marked the beginning of the rapid decline of imperial power in Africa.

The recession of colonialism

Three major factors were important in the recession of colonialism in Africa. First was the whole political and intellectual climate of the 20th century, which found expression in the concept of self-determination and the right of all peoples to govern themselves. It was also exemplified by the increasing concern for the dignity of the human personality and the entitlement of men to equality of opportunity. This more enlightened international atmosphere made possible the full operation of a second factor: the increasing awareness by the African of his own potential and the translation of this awareness into effective emancipatory action. The third factor was the dramatic shift in the world balance of power, resulting from the decline of the strength and influence of the European colonial powers—following two disastrous world wars—and their supercession by new international power structures. Western Europe no longer had the military or economic capability, unilaterally, to perpetuate its imperialism. Reinforcing this situation was the rise of the Soviet Union and the United States—both professedly anticolonial nations—as the pre-eminent powers of the latter 20th century.

The history of colonial emancipation has naturally differed from country to country. Generally speaking, however, where machinery existed—even in a vestigial form—to enable political and constitutional change to occur and where the colonial power was sufficiently sensitive and flexible to the requirements of circumstance, the transition from colonial rule to full sovereignty was relatively peaceful. On the other hand, where the colonial power had frozen its policies within the context of the 19th-century approach, it was obliged to resort to repressive colonial wars to maintain the status quo; the inevitable transition thus became bloody and rough. An index of the extent of the change that took place may be seen in the fact that by 1970, 42 African states had acquired their sovereignty and had acceded to membership in the United Nations organization.



In October 1966, the United Nations terminated South Africa's mandate over South West Africa and subsequently renamed it Namibia until independence. South Africa considers the UN action illegal and continues its jurisdiction over the territory.

African Independent States
(1972)

	year of independence	population (1971 estimate)
Algeria	1962	14,800,000
Botswana	1966	700,000
Burundi	1962	3,600,000
Cameroon	1960	5,800,000*
Central African Republic	1960	1,600,000
Chad	1960	3,800,000
Congo (Brazzaville)	1960	1,000,000
Dahomey	1960	2,800,000
Egypt	1922	34,000,000
Equatorial Guinea	1968	300,000
Ethiopia		25,200,000
Gabon	1960	950,000†
Gambia, The	1965	400,000
Ghana	1957	8,900,000
Guinea	1958	4,000,000
Ivory Coast	1960	4,400,000
Kenya	1963	11,700,000
Lesotho	1966	900,000
Liberia	1847	1,600,000
Libya	1951	2,000,000
Malagasy Republic	1960	7,000,000‡
Malawi	1964	4,500,000
Mali	1960	5,100,000
Mauritania	1960	1,200,000
Mauritius	1968	900,000
Morocco	1956	15,200,000
Niger	1960	4,100,000
Nigeria	1960	56,500,000
Rwanda	1962	3,800,000
Senegal	1960	4,000,000
Sierra Leone	1961	2,600,000
Somalia	1960	2,900,000
South Africa	1910	22,100,000
Sudan, The	1955	16,100,000
Swaziland	1968	400,000
Tanzania (Zanzibar)	1961(63)	13,600,000
Togo	1960	2,000,000
Tunisia	1956	5,100,000*
Uganda	1962	10,100,000
Upper Volta	1960	5,500,000
Zaire	1960	22,500,000
Zambia	1964	4,300,000

*1970 estimate. †1970 census. ‡Questionable reliability.
Source: UN population and Vital Statistics Report, July 1, 1972; official country figures.

By 1972 the only major colonial territories remaining in Africa were the Portuguese-administered territories of Angola, Mozambique, and Portuguese Guinea, where wars of colonial emancipation were still raging; and Rhodesia, where a minority white-settler regime had seized government from Britain, the administering power. South West Africa (Namibia) remained under South African jurisdiction despite a 1971 ruling by the International Court of Justice that the South African presence was illegal. Minor colonial territories were the French Territory of the Afars and Issas, the Spanish Sahara, and other Spanish possessions.

The heritage of colonialism

The colonial heritage has posed many problems for the African states and has made efforts at cooperation among them more difficult. There is the problem of the mosaic of arbitrary and illogical boundaries, most of which mark the extent of colonial conquests or imperial advance, and which usually have no relation to geography or ethnicity. Colonialism has resulted also in a degree of political and economic identification with, and dependence upon, the metropolitan country (particularly strong in the former French colonies) that persists even after independence. Consequently, many African states still find it difficult to pursue international policies without inhibition from extra-African considerations.

The relatively brief period of European dominance in Africa also resulted in a pattern of institutions and habits inherited from the various colonial powers. This superimposition of mainly Anglo-Saxon and French cultures on the indigenous African cultures created a difference of outlook between English-speaking and French-speaking African countries, which has tended to complicate relations between these two groups of states.

Many African leaders, conscious of this dichotomy, have at various times worked for the development of a Pan-African approach to the continent's problems. A product of these efforts was the establishment of the Organization of African Unity (OAU) in 1963, founded to create a sense of unity and to identify areas in which practical cooperation among African states might usefully be pursued.

MAJOR LANGUAGE PATTERNS

Indigenous African languages fall into such broad categories as the Hamito-Semitic grouping in northern Africa and what some scholars consider the Niger-Kordofanian, Nilo-Saharan, and Khoisan families south of the Sahara. The sub-Saharan language classification is subject to considerable scholarly dispute. (For indigenous African language groupings and characteristics, see AFRICAN LANGUAGES and HAMITO-SEMITIC LANGUAGES.)

Meanwhile, the cultural impact made by the European powers during the colonial era is reflected in the continuing official use of languages of the former administering powers in large areas of Africa.

In general, Arabic is the current language, both officially and unofficially, in states north of the Sahara, as well as in The Sudan. In many countries, Arabic is also the language of religion. Between the Sahara and the Zambezi River, either English or French is widely understood. French is habitually the official language in the states that formerly comprised French West Africa and French Equatorial Africa, as well as in Madagascar and in Zaire. Similarly, English is habitually the official language in the states of West, Central, and East Africa that were formerly under British administration and is also the official language in Liberia. Portuguese is current, officially and otherwise, in the territories under Portuguese administration. In South Africa there are two official languages—Afrikaans (which developed out of a colloquial version of 17th-century Dutch) and English. Amharic is the official language in Ethiopia.

Among Negroes on the northern fringe of West Africa south of the Sahara, Hausa, a language with a Hamitic base, is spoken by 15,000,000 people (1971 estimate). Indeed, Hausa has become, like Swahili in parts of East and Central Africa spoken by an estimated 13,000,000 people (1971), a lingua franca that is in use over a wide area, traversing ethnic and political boundaries. Further west, Dyula, a dialect of Malinke, has also acquired the status of a lingua franca.

Hindu, Gujarati, Pakistani, and other languages of the Indian subcontinent are spoken among the Asian communities of East and southern Africa. (F.S.A.)

VII. Resource development

The countries of North Africa, unlike those of the rest of the continent, have wide-ranging and ancient traditions of manufacture. At the end of the 19th century, however, Africa as a whole was regarded solely as a potential source of raw materials or as a natural market for Europe. In the course of time, limited industrialization tended to converge around the relatively large expatriate settlements, with technical considerations operating in favour of the industrialization of some areas and transport costs constituting the dominant development factor in others. Though World War II led to a general acceleration in the process of industrial development, by 1950 the total factory output of manufacturing industries (excluding South Africa) had an estimated value of only about $2,800,000,000, or 7 percent of the gross domestic product (GDP).

Industrial development

After 1950 output rapidly increased—to about 12 percent by 1970. The substantial increase and its range were due to such factors as increased demand, the substitution of home-produced for imported goods, and the encouragement of manufacturing by individual African administrations.

Major weaknesses nevertheless were evident, among them high capital costs; the political division of Africa into more than 40 countries, which inhibited mass production and mass marketing; and a scarcity of indigenous skilled personnel.

Despite its expansion in recent years, the relative sig-

nificance of manufacturing remains considerably smaller than in the advanced countries and smaller also than in Asia (excluding Japan) and Latin America. Furthermore, the share of manufacturing in the GDP varied widely in different African countries in the early 1970s. At the lower end of the spectrum were countries such as Mali, Togo, and Upper Volta, where the contribution of manufacturing was a minuscule 1 percent of the GDP; at the upper end of the spectrum were countries such as Egypt and Rhodesia, where the share of manufacturing was about 17 to 20 percent. To put it somewhat differently, the total output of manufacturing in the Republic of South Africa alone was some 90 percent of the output of the rest of Africa.

Of Africa's output (excluding South Africa), Egypt produced a quarter, and nine other countries, in the order given—Morocco, Nigeria, Algeria, Rhodesia, Tunisia, Zaire, Mozambique, Kenya, and Ethiopia—accounted for one-half. The remaining quarter of the output was distributed among more than 30 countries.

INDUSTRIES

Manufacturing in Africa tends to concentrate on comparatively simple items and on those where some special advantage is available to the African producer, although the range of products was widening rapidly in the 1960s. Toward the end of the 1960s, industrial production included electric motors, automobiles, and tractors, while airplanes were also being assembled, and petrochemical complexes had been established. Most industrial plants, however, were of the relatively simple kind, being engaged in food processing or in manufacturing textiles, cement, or other building materials.

Mining. The mining industry is an increasingly significant source of national income, foreign exchange, and raw materials for the development of local processing industries. In 1960 it accounted for 5 percent of the total GDP of developing Africa; by 1968 it had risen to about 10 percent. In South Africa it declined from 14 to 12 percent because of the increased share in GDP of other sectors such as manufacturing and commerce. The industry is very unevenly distributed: in 1966, for example, 49 percent of mineral earnings came from southern Africa alone and 28 percent from central Africa, while western and northern Africa each accounted for about 10 percent and eastern Africa for less than 2 percent.

Use of iron and steel

Heavy industry and engineering. Except in South Africa, iron and steel are used in every African country for construction rather than for engineering. Integrated iron and steel plants exist in Algeria, Tunisia, Egypt, Rhodesia, and South Africa, while smaller production facilities —often based on the transformation of scrap—exist in several other countries. At the end of 1969 existing capacity in developing Africa was estimated to be less than 2,000,000 tons, while in South Africa it was about 5,000,000 tons. The main producing countries supply between 40 and 60 percent of their own iron and steel requirements; the rest depend mainly on imports.

Petroleum-refining and petrochemical industries. Petroleum-refining capacity (excluding South Africa) exceeded 680,000 barrels per day in the early 1970s. Based on domestic crude-oil output in a few cases and on imported crude oil in others, 21 countries had at least one petroleum refinery. The development of the petrochemical industry followed the establishment of refineries in some countries. In 1965 there were only three major petrochemical complexes in Africa—in Rhodesia, Egypt, and South Africa; but by 1971 seven additional countries had refinery capacities exceeding 1,000,000 metric tons—Algeria, Ghana, Kenya, Morocco, Nigeria, The Sudan, and Tunisia.

Manufacturing. During 1970 about 2,875,000,000 square yards of textiles were produced. Most of the output is processed in bleaching, dyeing, and printing establishments that form an integral part of composite spinning and weaving units. With the exception of Egypt, African producers have concentrated on the home market and on the manufacture of cotton textiles. Rayon–synthetic and woollen materials are, for the most part, imported. Ready-made clothing, both domestic and imported, has emerged as a major market factor.

By 1970 most African countries were producing their own beer. Zaire was the leading producer, followed by South Africa and Nigeria. Production facilities for soft drinks are widespread, and there are more than a dozen plants for manufacturing various types of distilled liquor.

Cement production

By 1970 at least 22 countries were known to have cement plants. South Africa was the leading producer (6,000,000 tons a year), followed by Egypt (about 4,000,000 tons). The transport costs of cement make its price variable. On the North African coast the price may be equivalent to U.S. $20 or less per ton, whereas in some of the inland countries it might be higher than U.S. $100 a ton. On the west coast the price ranges from U.S. $35 to $40 and on the eastern coast from U.S. $25 to $45.

Timber, fisheries, animal husbandry. In 1967 there was an estimated 500,000,000 acres of closed high forest, of which 5,000,000 consists of plantations: 86 percent of the output of wood is used for fuel, 6 percent is converted into sawed wood and veneer logs, and 8 percent is directed into other categories. In 1970 between 800 and 900 sawmills, spread throughout the continent, produced about 145,000,000 cubic feet of sawed wood, of which almost a third was exported. Plants for the manufacture of plywood, particle board, and fireboard had considerable excess capacity.

The pulp and paper industry is concentrated in northern Africa and in southern Africa, although seven small paper mills are established in West, Central, and East Africa. The African output of wood pulp amounted to about 1,200,000 tons in 1970, 35 percent of which was exported—mainly to Europe. The main products of the paper industry proper comprised newsprint (about 200,000 tons), printing and writing papers, and paperboard (about 210,000 tons) and industrial paper (550,000 tons). The bulk of the output was directed to national markets.

In 1970 the world fish catch, including both saltwater and freshwater fish, was estimated to amount to 76,000,000 tons, of which Africa's contribution amounted to 4,600,000 tons. The countries in Africa that contributed the most to this total were Morocco, Ghana, Ivory Coast, Senegal, Sierra Leone, Angola, Cameroon, Chad, Zaire, Madagascar, Malawi, Tanzania, Uganda, Zambia, South West Africa, and South Africa.

Herring, sardines, and anchovies contribute most to the total catch, followed by redfish, bass, conger, jack, and mullet in tropical waters and tuna, bonito, skipjack, cod, hake, and haddock in northern and southern waters. A series of commissions has been set up around the African continent to control fishery development; these agencies collect statistics, conduct research, and attempt to regulate the operation of the fishing industry. The production (largely by traditional methods) of freshwater fish in 1969 was estimated to be a little more than 1,000,000 tons, of which 432,000 tons were produced in East Africa, 312,000 tons in central Africa, 269,000 tons in west Africa, and 103,000 tons in North Africa. There is very little freshwater fish production in southern Africa (*i.e.* south of the Limpopo River).

Africa's wool production in 1971 amounted to almost 175,000 tons (before being cleaned of grease), or to about 83,000 tons of clean (washed) wool. The chief producing countries are Algeria, Morocco, Libya, Egypt, and South Africa. South Africa alone produces approximately 70 percent of the entire continental clip from merino or crossbred merino sheep.

Statistics for the output of cattle hides and of sheepskins, goatskins, and pigskins are notoriously unreliable; it is, however, sometimes assumed that the annual output of hides is in the region of 10 percent of the total population of cattle and that of sheepskins and goatskins in the region of 25 percent. The number of game hides and skins processed and sold annually is not accurately known.

POWER

A spectacular development in the use of electric energy took place in the 1960s, partly because of the growth of

the petroleum industry and partly because of the establishment of large hydroelectric plants. From 1937 to 1970 electric-energy production increased nearly 13 times—from 6,500,000,000 kilowatt-hours (kW-hr) to 89,000,000,000. The average annual growth rate from 1957 to 1967—a period of great effort and activity—was more than 8 percent. The proportion of thermal (as distinct from hydroelectric) production, however, most of which was in South Africa, decreased. The general development trend indicates that the public sector will probably absorb the output formerly used by the private sector, which, especially in its industrial aspects, has shown a marked decline. The increased quantity and quality of electric energy have given rise to problems of transmission and distribution. Unlike thermoelectric plants, which may be sited where the consumer demand is greatest, new hydroelectric installations are not flexible; the type of tension lines in use has therefore changed. Though in the 1950s, for example, it was common practice to use tensions below 220 kilovolt lines (kv); the 1960s marked a change to the use of high-tension lines. In Nigeria 330-kilovolt lines were laid, and similar lines were used in the Rhodesian system, which feeds the Copperbelt, in Zambia, and Salisbury and Bulawayo, in Rhodesia. This same system is interconnected in the north with the large Shaba (formerly Katanga) hydroelectric power stations. Two 500-kv lines supply the consumer centres of Egypt from the hydroelectric power stations of Aswān and Saad El Aali. The possibility of supplying landlocked states with energy from the large hydroelectric plants in the coastal states is more likely to be considered in the future.

Electricity distribution systems

Over the past 50 years, electric-energy consumption in large urban centres, especially when they are near coastal towns and mining areas where industrial activity has taken shape, has increased considerably. In 1929 per capita consumption of electricity was only 20 kilowatt-hours. In 1949 this rose to about 60 kilowatt-hours and by 1970 had risen to more than 312 kilowatt-hours, although outside the urban centres and industrial zones consumption was on average less than 15 kilowatt-hours per annum. Although some countries have extended networks to the rural areas or have increased the numbers of isolated, low-powered stations and independent networks, progress in rural electrification has not been especially noteworthy.

Hydroelectric power. The technically exploitable hydroelectric resources of Africa represent 1,630,000,000,000 kilowatt-hours per annum, or approximately one-third of the world's hydroelectric potential. Hydroelectric resources are unevenly distributed: North Africa has only 5 percent, West Africa 10 percent, East Africa 22 percent, southern Africa 17 percent, and Central Africa 46 percent. This type of energy, which represented about a quarter of overall electric production in 1969, tends to increase at an average annual rate of about 13 percent. Whereas only 18 countries produced hydroelectric energy in 1957, it was produced by 31 in 1969, of which nine thereby met more than 80 percent of their own electrical needs.

Countries producing more than 50 percent of their needs in 1967 were: the Central African Republic (almost 100 percent); Rwanda (98 percent); Zaire (almost 98 percent); Uganda (97 percent); Cameroon (96 percent); Ghana (96 percent); Rhodesia (90 percent); Angola (88 percent); Malawi (88 percent); and Tanzania, Ethiopia, the People's Republic of the Congo, Morocco, Ivory Coast, Kenya, and Madagascar (over 50 percent). Many hydroelectric installations will be completed by 1980.

Other sources of power. A number of steam power stations are located in ports and cities near the coasts. The largest installations of this kind operate in Tunis (Tunisia), Casablanca and Oujda (Morocco), Dakar (Senegal), Abidjan (Ivory Coast), and in Lagos (Nigeria). Steam power stations using coal are much the largest, especially in South Africa, which had an installed capacity of 8,924,000 kilowatts (kW) in 1969.

Although geothermal energy is potentially a major source of electric production, especially in East Africa where many active volcanoes and hot springs are associated with the Rift Valley, there is as yet only one small (220 kW) geothermal power station in existence, in Shaba, Zaire. The most favourable zones for such installations are Kenya, Ethiopia, and Cameroon.

The geothermal power potential

Africa has no electric energy-generating nuclear power stations and only a few small research installations.

AGRICULTURE, IRRIGATION, AND LAND USE

Roughly 35 percent of the total land area of Africa is devoted to agriculture: arable land and land under permanent crops represents about 7 percent, permanent meadows and pastures about 28 percent.

Cereals and grains. The distribution of crops is influenced by a series of factors in which rainfall and its distribution play a significant part.

In 1971 Africa had about 41,300,000 acres planted in maize (corn), yielding approximately 22,000,000 tons of grain. Highest yields per acre were recorded in Egypt, Rhodesia, and in Zambia. Irrigation is practiced in both Egypt and Rhodesia. The total area devoted to the production of millet and sorghum in 1971 was about 63,000,000 acres. The countries that have developed production of these cereals to the greatest extent are Burundi, Cameroon, Chad, Egypt, Ethiopia, Ghana, Guinea, Mali, Niger, Nigeria, Rhodesia, South Africa, Senegal, Somalia, The Sudan, Tanzania, Uganda, and Upper Volta.

Rice production and consumption were becoming increasingly important in the early 1970s. Production of unhulled "paddy" rice in 1971 was about 5,600,000 tons. The most important rice-producing countries are Zaire, Guinea, Ivory Coast, Liberia, Mali, Nigeria, Senegal, Sierra Leone, Upper Volta, Togo, Tanzania, Egypt, and Madagascar.

The principal wheat-producing countries are Algeria, Ethiopia, Kenya, Libya, Morocco, South Africa, Tunisia, and Egypt.

Fodder crops are not widely produced. Cereals for fodder production in subtropical and tropical zones include maize (corn), sorghum, and millet; protein-rich legumes, sown together with these crops, include velvet beans, cowpeas, soya beans, and lablab (a type of bean). In temperate Mediterranean zones, oats admixed with broad beans and vetches are also sown. In climatically similar parts of south central, southern, and eastern Africa, pure strands of alfalfa (lucerne) are fairly widely grown. In the northeast, particularly in Egypt but also in The Sudan, berseem (a clover used for forage) is produced in considerable quantity with the help of irrigation.

Fruits and vegetables. Among the important fruits are bananas, pineapples, dates, figs, olives, oranges and tangerines, and grapefruits; vegetables include potatoes (sweet potatoes), yams, plantains, cassava, tomatoes, onions, pulses, and cowpeas.

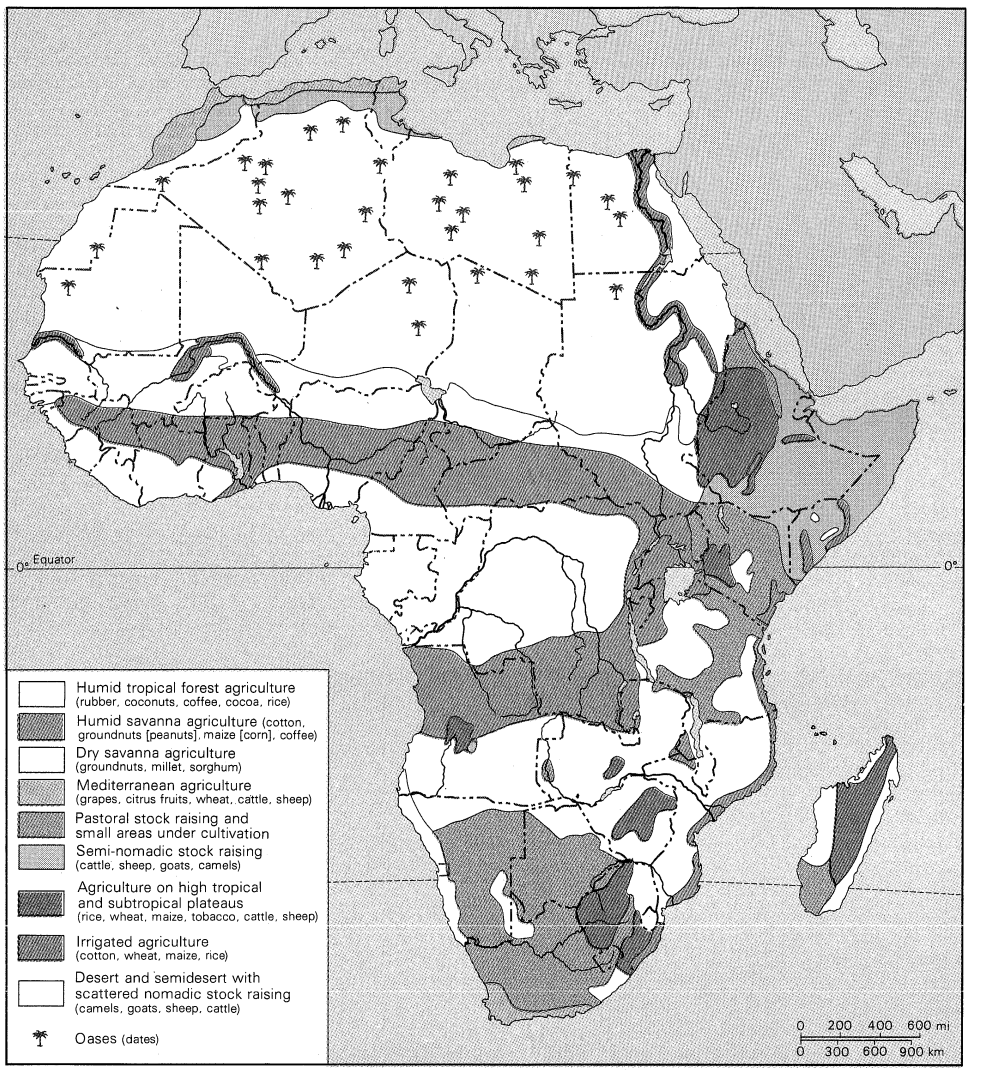
Well distributed throughout Africa except in the north and south, the banana has been cultivated to some extent in Cameroon, Zaire, Ivory Coast, and as an intensive enterprise assisted by irrigation in Somalia, Madagascar, and Egypt. Ivory Coast, Kenya, Tanzania, and South Africa are the largest pineapple-producing areas.

Date palm and fig cultivation

A typical tree of desert oases, the date palm is most frequently cultivated in Egypt, Algeria, and Morocco. The fig is cultivated largely in Algeria, Morocco, and Tunisia. The olive tree is developed principally in Tunisia, Algeria, Libya, and Morocco. About two-thirds of the olive production is expressed into olive oil.

The principal orange-growing countries of Africa are Algeria, Morocco, South Africa, Egypt, Tunisia, Ethiopia, Guinea, Sierra Leone, Ghana, and Tanzania. The largest yields are produced in countries where basin irrigation is widely practiced. South Africa is the largest producer of grapefruit, and Morocco is the second largest. In 1971 Africa's total grapefruit production was nearly 135,000 tons; production of oranges was nearly 3,000,000 tons; production of lemons, limes, and other citrus fruits was about 115,000 tons.

Potatoes are cultivated in elevated regions in such countries as Burundi, Rwanda, Ethiopia, Kenya, and Mada-



Agricultural regions of Africa.

gascar. Countries with Mediterranean climates, such as Morocco, Algeria, Egypt, and South Africa, are also important producers. The sweet potato is also widely grown, as are two species of yam. The plantain, or cooking banana, is grown extensively in equatorial or tropical forest zones. Cassava (manioc) is widely grown, except in the north, the south, and in very dry regions. Except in Central and West Africa, where the crop is regarded as a staple, cassava is generally treated as famine reserve.

Tomatoes are widely grown, with the largest producing areas bordering the Mediterranean; more than half are grown in Egypt. Onion production is widely dispersed, with the countries bordering the Mediterranean producing for export. Countries producing green beans on the largest scale are Egypt, South Africa, Morocco, and Algeria. Production of beans and peas for drying is much more widespread; major producing countries are Angola, Burundi, Ethiopia, Madagascar, Rwanda, South Africa, Uganda, Zaire, and Morocco. Ethiopia, Egypt, and Morocco are the most substantial producers of dry broad beans, while Ethiopia and Morocco are the largest producers of chick-peas. Lentil production is significant in Ethiopia and Egypt. The cowpea is becoming increasingly important as a legume for human consumption. Best suited to warmer tropical environments, it is produced in Botswana. Madagascar, Niger, Nigeria, Senegal, Uganda, and Upper Volta. Heavy vegetables, such as cabbages and cauliflowers, are principally grown in countries along the Mediterranean coast, from where it is possible to export small quantities to southern Europe. Other vegetables of significance in tropical Africa include peppers, okra, eggplant, cucumber, and watermelon.

Vineyards. Grapes are produced in northern and southern Africa. Although attempts are being made to curtail production in Algeria, Morocco, and Tunisia (because of the loss of markets in the European Economic Community as a result of the Community's common agricultural policy), a distillery for producing brandy from surplus wine has been installed in Algeria by the Soviet Union, and there is some resumption of the traditional wine trade between Algeria and France.

Dairy products. Milk and its products are consumed according to national and development circumstances, either as whole fresh milk, skim milk, buttermilk, fermented milk, flavoured milk, yogurt, butter, ghee (clarified butter fat), soft cheese, processed cheese, manufactured creamery cheese, or as ice cream. No reliable figures are available for production and consumption of exclusively African milk and dairy products. Except in South Africa, Rhodesia, and Kenya, however, production is known to be insufficient to meet domestic needs.

Miscellaneous crops. The cashew tree is grown to a limited extent in East Africa and to a lesser extent in the coastal countries of West Africa. Yields vary widely. The macadamia nut tree, introduced from Australia to East Africa, yields about 60 pounds of nuts a year when it is 12 to 14 years old.

Coconuts and copra

Coconuts are important in the Comoro Islands, Madagascar, Mauritius, Mozambique, Nigeria, Tanzania, Zanzibar, and Togo. About 1,300,000 tons of copra (dried coconut kernels) were produced in 1971. Kola nuts (estimated production of which was 200,000 tons in 1965) are grown principally in the forested regions of Nigeria, Ghana, Ivory Coast, Sierra Leone, and Liberia. Yield

varies, but it averages from 300 to 500 pods per mature tree; each pod contains from six to 12 nuts. The oil palm is grown in areas of heavy rainfall. In 1971 production of palm kernels was about 860,000 tons and of palm oil (which includes pericarp oil from the ovary of the palm, as well as kernel oil) 1,200,000 tons. The largest producers in Africa are Nigeria and Zaire.

Groundnuts (peanuts) are widely grown, with production being of particular importance in Cameroon, The Gambia, Senegal, Malawi, Mali, Niger, Nigeria, South Africa, The Sudan, Uganda, and Upper Volta. In 1971 the estimated total yield exceeded 5,400,000 tons of groundnuts in shell.

The most important cotton-producing countries in Africa are Egypt, Uganda, Tanzania, The Sudan, Nigeria, Mozambique, and Chad. Production of cotton lint in 1971 was nearly 650,000 tons and of cottonseed 1,200,000 tons. Sesame is mainly produced in the Central African Republic, Ethiopia, Nigeria, The Sudan, Uganda, and Tanzania. About 180,000 tons of sunflower seed were produced in 1971, the main producing countries being South Africa, Ethiopia, and Kenya. Castor beans are grown in Ethiopia, South Africa, The Sudan, and Tanzania.

Production of cocoa in 1970–71 was 1,050,000 tons, the bulk of which was produced by Cameroon, Ghana, Ivory Coast, and Nigeria. Average yield per acre under average conditions of management is about 500 to 600 pounds of dry cocoa beans. The most important coffee-growing countries in Africa are Angola, Cameroon, Ethiopia, Ivory Coast, Kenya, Liberia, Madagascar, Tanzania, Uganda, and Zaire. Arabica, robusta, and liberica are the three main types. The principal tea producers are Kenya, Malawi, Mozambique, Tanzania, and Uganda, while the principal producers of rubber are Cameroon, Liberia, Nigeria, and Zaire.

The production of tobacco is widely distributed, the most substantial producers being Malawi, Nigeria, Rhodesia, South Africa, and Tanzania.

The largest cane sugar producers are the People's Republic of the Congo, Madagascar, Mauritius, Mozambique, Réunion, Rhodesia, South Africa, Swaziland, Uganda, and Egypt. Total production in 1971 was about 40,000,000 tons; about 1,400,000 tons of sugar beet were also produced.

VIII. Commerce

TRANSPORTATION

The fragmentation of African transport systems

Although highly developed transport networks existed in many parts of Africa in pre-colonial times, in the colonial era they were restructured to penetrate into the interior from the seaports and, in the main, to serve the commercial and administrative needs of the colonial powers. Their fragmentation, which led to interregional links being but thinly developed, resulted from the juxtaposition of varied and difficult terrains, the economic artificiality of certain national frontiers, the lack of a developed intra-African trade, and the strong orientation of commodity trade with the administering countries. All this was further complicated by the existence of vast unpopulated areas lying between the main centres.

The emergence in the 1960s of independent African governments who recognized the need to lift economies from their generally very low levels and, above all, to develop agriculture and embark on industrialization heralded improvements in economic planning, the development of transport networks, and the introduction of cheaper freight rates. But there remained a serious shortage of qualified African manpower to plan and manage transport systems at the national or multinational level and, simultaneously, to keep up with the rapid development of transport technology outside Africa.

Animal transport. There is some evidence that before the arrival of the camel, which was introduced into Africa via Egypt at the time of the Arab conquest, bullocks were used either as pack animals or to draw carts from the northern countries across the Sahara to the gold-producing areas of the ancient Sudan. From the 16th century onward, the Portuguese developed transport inland from the coast at Mozambique, and from the 17th century, first the Dutch and then British settlers from the Cape trekked northward and northeastward with their wagons. Except in such highland areas as Ethiopia where pack animals were and still are used, the tsetse fly often prevented the use of animal transport. With the steady progress in the development of transport infrastructure in many African countries, the use of bullocks in southern Africa, donkeys in northern and western Africa, horses in northern Nigeria, and camels in West and North Africa and the Horn of Africa has been reduced, but the extent of this reduction cannot be accurately gauged.

Motor transport. The arrival and rapid development of the internal-combustion engine in the 1920s transformed the collection and distribution of goods and personal travel. Roads were built, particularly in northern and southern Africa, but also in parts of the west and east. World Bank loans in the 1950s and 1960s, supplementing contributions to road and highway development from national budgets, financed the building and improvement of road networks in many African countries. An important new transport link under construction is a 900-mile highway to link Addis Ababa, the capital of Ethiopia, with Nairobi, the capital of Kenya. An economic feasibility study of a 1,800-mile highway across the Sahara to link Algeria and Tunisia in the north with Mali and Niger to the south is being carried out. A prefeasibility study has been completed for the construction of a trans-African highway about 4,000 miles in length which will link Mombasa in Kenya with Lagos in Nigeria, passing through Uganda, Zaire, the Central African Republic, and Cameroon.

Rail transport. The early railways were constructed partly to facilitate the administration of interior regions and to bring supplies from ports to central consumption or distribution points and partly—especially in the south—to enable valuable minerals or commodities to reach the coast for export. In Africa, as in Europe, the period of railway development extended from the end of the 19th century to the end of World War I, during which time, unfortunately, by constructing railways with different gauges of track and by operating them with rolling stock of different braking and coupling systems, the colonizing powers left a difficult and costly legacy for independent African countries who wished to link themselves together.

Investment in transport

As with roads, rail networks saw considerable improvements in the 1960s and, as a result, a lowering of transport costs. During that time investment in transport accounted for some 25 percent of all development expenditure. Between 1958 and 1966, for example, the World Bank and its affiliates financed the construction of 400 miles of new railway in Nigeria, more than 400 miles of railway extension in The Sudan, and advanced loans for railway extensions in Kenya. In addition, the United States granted loans to Tanzania for railway construction and financed the rehabilitation of the Senegal–Niger Railway running from Senegal to Mali. New railways for carrying iron ore were built in Liberia, Swaziland, and Mauritania. The People's Republic of China also extended financial assistance and labour to Tanzania and Zambia for the construction of a new railway from the Zambian Copperbelt to the Tanzanian port of Dar es Salaam; the railway, nearly 1,000 miles long, is scheduled for completion in 1973.

Air transport. Air transport is the primary means of international and sometimes of national travel in Africa. The late 1940s and the 1950s, as well as experiencing great advances in the extension and improvement of rail and road services, saw the emergence of a new transport factor in the introduction of internal and international scheduled air services. The rapid development of air transport increased the movement of goods and people and began to open up the hitherto largely closed interior of the continent. Transport became much quicker and usually cheaper. Since then, internal air services have steadily increased, and intercontinental air transport, especially of passengers, has greatly developed. The intro-

duction of capacious, supersonic, and vertical-take-off planes is expected further to reduce airfield costs. The largest international airports in operation are at Casablanca (Morocco), Cairo, Abidjan (Ivory Coast), Addis Ababa (Ethiopia), Nairobi (Kenya), Dowala (Cameroon), and Johannesburg (South Africa).

Navigation. Historically, over the vast interior between the Sahara and the Zambezi River, people or goods were transported by canoe or boat on the great river systems of the Nile, Sénégal, Niger, Congo, and Zambezi rivers and on the few but very large lakes. Where navigable conditions allowed, power-engined craft later supplemented or displaced canoes, but further development has been slight. Mention should be made, however, of the construction of lake ports and the installation of new rail ferries across Lake Victoria.

Meanwhile, on the coasts, artificial harbours have been developed. New berths have been added to existing ports, and new ports have been constructed: at Lomé (Togo), Cotonou (Dahomey), Safi (Morocco), Nacala (Mozambique), Safājah (Egypt), Suakin (The Sudan), Kismayu and Berbera (Somalia), Cap des Biches (Senegal), and elsewhere. In planning new ports, the choice of site, probable costs, and the possibilities of using containers or other unitized loads are taken into consideration.

TRADE

Internal trade. Intra-African trade records frequently understate the amount of trade—partly because of the lack of adequate statistics and partly because of the high rate of smuggling, which allows a substantial amount of traditional border trade to continue unrecorded. Apart from this, trade between African states has been handicapped by a tendency for trade to remain concentrated within the common currency areas and trade zones that first developed among African countries during the colonial era, by the often inadequate means of transport and communication, by the lack of complementary agricultural or other products, and by the absence of manufacturing industries.

On the average, the proportion of intra-African export trade to total export trade was about 6 percent in 1970, as compared to 11 percent between the Latin American countries and to 34 percent between developing Asian countries. Much of the intra-African trade consists of food, drinks, tobacco, sugar, cattle, and meat. There is also a good deal of re-export trade between the coastal and inland states, especially in machinery, transport equipment, and spare parts.

Common currency and trade zones that have evolved through the granting of preferences or the operation of common currencies inherited from former metropolitan powers include: the East African Economic Community, whose members are Kenya, Uganda, and Tanzania; the Central African Economic and Customs Union (UDEAC), comprising Cameroon, Gabon, the Central African Republic, and the People's Republic of the Congo; The Economic Community of West Africa, comprising Dahomey, Ivory Coast, Mali, Mauritania, Niger, Senegal, Upper Volta, and Togo; and the Maghreb Permanent Consultative Committee (CPCM), grouping Algeria, Morocco, and Tunisia.

External trade. Since the outbreak of World War II, there has been a considerable expansion in Africa's overall external trade. Between 1938 and 1970, total exports increased in value nearly fifteen times (from about US $1,000,000,000 to well over US $14,700,000,000), while imports rose tenfold from about US $1,500,000,000 to US $14,300,000,000. This growth compares favourably with that of the other developing regions, such as Asia and Latin America; exports now slightly outweigh imports. The large expansion in African exports is generally attributed to the increase in the demand for primary commodities during World War II and in the immediate postwar reconstruction period. Subsequently, the attainment of independence by a large number of African countries, especially in the early 1960s, followed by a new bid for economic development, strengthened the export expansion drive. Another reason for the rapid growth in African exports was the temporary increase in the price of primary commodities, although subsequently the general trend has been toward depressed commodity prices.

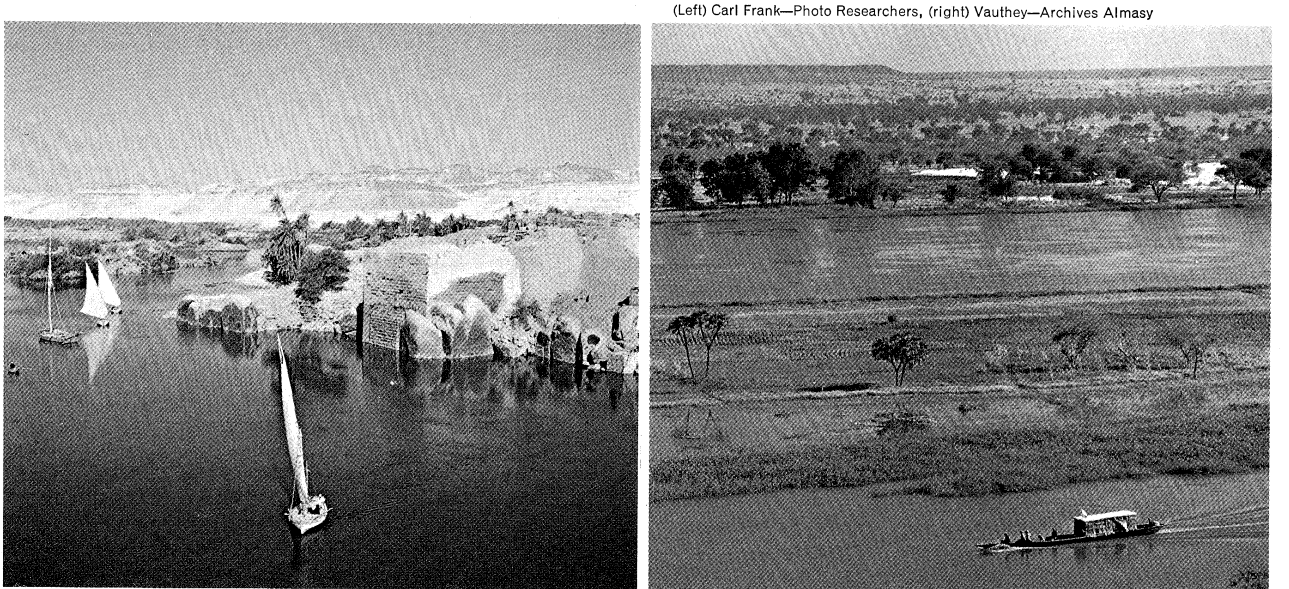
Exports. Not least among the factors that in recent years influenced the growth of African exports was the discovery of petroleum in Libya, Algeria, Nigeria, and elsewhere; the increased exploitation of much-demanded minerals, such as diamonds, in Sierra Leone, the People's Republic of the Congo, the Central African Republic, Zaire, and elsewhere; and the exploitation of other minerals, such as uranium ore.

Since independence, attempts have been made to diversify external trade relations, but the record of achievement has not been impressive, since Africa's trade patterns largely reflect the influence of traditional links with the countries of western Europe.

Trade with the EEC and EFTA

The biggest market for African exports is the European Economic Community (EEC), in which France, as the largest buyer, received more than 32 percent of all African exports to the EEC in 1970, West Germany 28 percent, Italy 17 percent, The Netherlands about 12 percent, and Belgium-Luxembourg the remainder. The European Free Trade Association (EFTA) ranked second in importance as far as exports were concerned, with the

(Left) Carl Frank—Photo Researchers, (right) Vauthey—Archives Almasy



African rivers.
(Left) The Nile River near Aswān, Egypt. (Right) The Niger River near Niamey, Niger.

largest buyer, the United Kingdom, accounting for 76 percent of the total in 1970. Other importers from Africa were Japan and the United States.

In most African states one or two primary commodities dominate the export trade: petroleum accounted for over 99 percent of Libyan exports in 1970 and about 66 percent of Algerian and 43 percent of Nigerian exports; iron ore represented more than 87 percent of Mauritanian exports and about 72 percent of those from Liberia; copper accounted for 95 percent of Zambia's exports and almost 72 percent of those from Zaire. Statistics for primary agricultural commodities were: cotton from Chad (almost 72 percent) and from Egypt (more than 61 percent). Coffee from Burundi, Uganda, Ethiopia, Ivory Coast, Rwanda, Madagascar, and Kenya also represented a significant proportion of the exports of these countries. Sugar accounted for over 95 percent of the exports from Mauritius.

Imports. The tremendous increase in recent years of Africa's import trade has resulted in the import bill of many African states consuming a substantial proportion of export earnings; in consequence, many governments subsidize many of the imports required. The bulk of imports comes from western Europe, especially the EEC and EFTA countries, which together accounted for more than 53 percent of Africa's total imports in 1970. France, the United Kingdom, Italy, and Belgium are the principal suppliers—mainly to their former colonies. There has, however, been a substantial increase in imports from the United States, Japan, West Germany, and South Africa. In 1970 the United States supplied 11 percent of Africa's total imports, Japan about 9 percent, eastern Europe about 6 percent, and Asia (exclusive of Japan) 14 percent. Imports are primarily required to develop manufacturing industries and are therefore confined for the most part to industrial goods, machinery, transport equipment, and durable consumer goods. Africa continues to allocate considerable foreign exchange to food imports, which in 1965, for example, accounted for more than 16 percent of the overall import total. The situation is unlikely to improve, since per capita food production in Africa had, by 1971, actually declined below levels of 1960–65.

Problems and prospects. The most difficult problem facing Africa in its export-promotion drive to the developed countries is the diminishing demand for agricultural primary commodities. The demand for these commodities in the developed countries has reached a saturation point, and this implies that further African exports can only be achieved through substantial population increases in those countries. Given, however, the generally very low rate of population growth in the developed countries, the demand for primary commodities is not likely to increase substantially. Second, African exports continue to meet stiffer tariffs and other trade restrictions in the developed countries, and this makes it more difficult for primary commodities to penetrate those markets. Third, there is growing competition from synthetics and substitutes for most of the African agricultural primary commodities, a fact that further reduces the demand. Again, trade in primary commodities among the developed countries themselves is on the increase. In other words, the developed countries are now increasingly becoming self-sufficient in such commodities.

Conditions affecting commodity trade

African countries also face competition from other developing countries in Asia and Latin America, particularly in the export of such commodities as rubber, natural fibres, coffee, tin, and copper. Further fluctuations in world market prices create severe problems of foreign-exchange earnings for most developing countries, particularly those that depend on one or two primary products for much of their export earnings.

There are, therefore, growing fears among the African countries that during the 1970s their export earnings will, in consequence, be adversely affected. This is recognized as a common problem facing developing countries. It is also a point of major concern to international organizations, especially the United Nations Conference on Trade and Development (UNCTAD), which has become an important forum for the discussion of the general problems affecting exports from developing to developed countries. These problems concern the access of primary commodities and of semi-manufactured and manufactured goods to the markets of the developed countries; quotas for certain categories of primary commodities, such as coffee, cocoa, sugar, and copper; and the general attitude of developing countries on these issues in their negotiations with the developed countries.

IX. Demographic patterns

CURRENT PATTERNS

With the advent of the colonial era the African continent was divided into small geographically and politically based units that took little or no account of ethnic distribution. These political boundaries persisted—despite the fact that in 1972 only seven mainland African countries remained under colonial rule—and the continent continued to be characterized by a large number of countries with predominantly small populations. Africa's total population, estimated at 354,000,000 by mid-1971, was divided among 58 countries (including the offshore islands). Of these, more than half had less than 5,000,000 inhabitants, 11 had more than 10,000,000, and only four had populations of more than 20,000,000.

Density. Covering about a fifth of the world's land surface, Africa had in 1971 a population density of about 30 persons per square mile—this is of the same order as the Soviet Union and the Americas but considerably lower than Europe and Asia—both about 120 per square mile in 1971—and less than half the world average of 65. Wide variations in density occur from country to country in Africa and within countries.

Areas of dense population

In general, the most densely populated areas are found bordering the lakes, in the river basins, along the coastal belts of North and West Africa, and in certain highland areas, while settlement is the most sparse in the desert and savanna areas. Thus, Burundi and Rwanda, situated in the East African highlands, are the most densely populated countries in Africa; while Mauritania and Libya, in the Sahara, and Botswana and South West Africa, in the Kalahari, are those where settlement is most sparse. Since much of Africa's total area is desert, semidesert, or swamp and is unable at present to support human—and in particular agricultural—settlement, a more realistic measurement of population pressure is the density per square mile of arable land. On average, one square mile of arable land in Africa has to support or at least to provide food for more than 250 people (as compared with 750 people in Europe, 1,400 in Asia, and 800 in South America). In some countries, the number is appreciably higher—over 2,500 in Egypt, 2,100 in Mauritius, 1,500 in Angola, and more than 500 in six other countries.

Internal migration. Mostly rural for centuries, Africa is rapidly becoming more urbanized. The population in towns of 20,000 or more has increased from about 21,500,000 (10 percent of the total population) in 1950 to about 36,500,000 (13 percent) in 1960 and to more than 60,000,000 (17 percent) in 1970. Indigenous urban centres had long existed in parts of West Africa. In the 19th century the population of Kumasi, capital of Ashanti, was about 30,000 while many Yoruba towns were considerably larger. Before World War II, growth was very slow in these and other tropical African urban centres. Since then, an increasing expansion of towns and cities has taken place in sub-Saharan Africa and in those countries along the Mediterranean coast where centres of trade and population have long existed in various stages of growth and decline.

These northern countries are the most urbanized in Africa, with between 20 and 40 percent of their population in towns of 20,000 or more. The biggest city in Africa is Cairo, with a population of more than 5,100,000 in 1971. Other large cities are Alexandria (more than 2,000,000 in 1971; Casablanca, Kinshasa, Lagos, and Johannesburg (more than 1,300,000 in 1970); and Accra, Addis Ababa, Algiers, Giza, and Ibadan (each between 500,000 and 1,000,000 in the early 1970s).

In many countries a single large city, usually the capital, often attracts the vast majority of rural–urban migrants, as well as people from smaller towns. In some countries this has resulted in a situation similar to that of Malawi, where in 1966 there were no medium-sized towns (*i.e.*, of 100,000 to 200,000 inhabitants), and the whole urban population (*i.e.*, those living in localities of 20,000 or more inhabitants)—save for a few small towns with populations below 20,000—was concentrated in Blantyre.

International migration. Most movement occurs across uncontrolled borders and between people of the same tribal groups. Much is, in any case, seasonal and is restricted to labourers and nomadic herdsmen. Controlled immigration and emigration are generally negligible—the only recent examples being the forced emigration of Asians from East Africa and the emigration of people from neighbouring West African states caused by the enforcement of the Alien Compliance Order of 1969 in Ghana. Movement of refugees is quite substantial in some parts of Africa.

Substantial minority groups

In all African countries the majority of the population is of indigenous origin. In South Africa there are large numbers of Europeans (about 20 percent of the population). Other countries with significant European populations are: Rhodesia (243,000 in 1970), Zambia (43,000 in 1969), Angola (250,000 in 1966), South West Africa (91,000 in 1970), Mozambique (66,000 in 1960), Kenya (40,000 in 1969), and Senegal (about 47,000 in the later 1960s). Other substantial minority groups are Asian peoples (chiefly in South and East Africa), Arabs (in West and East Africa), and people of mixed origin.

Population growth and its effects. Birth rates have mostly remained high—around 45 per 1,000 people—with the average number of children per woman being between five and six. A sharp decline in death rates, particularly among infants and small children, is due to the spread of medical services. Population growth rates are in consequence often high. In several countries the population is already increasing at rates of about 3 percent per annum, and almost nowhere is the annual increase less than 2 percent. (Rates in Europe are at about 1 percent.)

High fertility and falling mortality considerably influence the age distribution of the population. In most African countries, about 45 percent of the population is less than 15 years of age, compared to about a quarter of the population in several European countries. People in the over-60 age group constitute a further 6.5 percent of the population, so that the ratio of the nonworking population to those of working age (generally from 15 to 59 years old) is slightly greater than 1:1, as compared to 1:1.4 in developed countries. Thus, theoretically, each member of the working population has to produce enough to support at least one nonworking member.

The labour force. It is estimated that as much as 60 percent of the working population is still engaged in subsistence agriculture and a further 20 percent, mainly in production of cash crops. In the sub-Saharan countries the total percentage dependent on agriculture is likely to be nearer 90.

The remainder of the working population is divided mainly between government service (including the police and armed forces), personal services (domestic services, laundry, hairdressing, etc.), and an increasing number of mining and industrial projects. Underemployment, particularly in the agricultural sector, is widespread, and unemployment is rising, especially in urban areas.

Women in the labour force

Participation in labour by women varies considerably from country to country. Women in paid employment are generally fewer than men, though a large proportion of women in sub-Saharan countries are engaged in subsistence agriculture—if only part of the time. Women are also employed in the civil service, in trading (especially in West Africa), in domestic service, and to an increasing extent in light industry.

FUTURE TRENDS

Current demographic patterns in Africa have, in general, evolved over a long period of time, although some are of more recent date. Future changes are expected to be much more rapid.

Predicted growth rates. With further improvements in public-health facilities and with wider education in such fields as hygiene, nutrition, and child care, the mortality rates are likely to decrease steadily for some time. Fertility, however, is expected to remain constant or even to increase slightly, with further improvements in the health situation of the general population and with the settlement of nomadic peoples. In consequence, the annual rates of population growth for most countries of the region might by the end of the century even rise above 3 percent. Should this be so, Africa's growth rate will be the highest in the world, surpassing even that of Latin America.

This high rate of growth will inevitably place a very heavy burden on the resources of the developing African countries, particularly in terms of housing, health services, education, and employment and, in some countries, on arable land and food supplies. Some 3 percent or more of the gross national product will be absorbed by the population growth; this will mean that these countries will at best keep the average standard of living at its present level.

Population policies. In several countries, the realization of the effects of rapid population growth has led to official policies aimed at reducing birth rates and at slowing down (though not eliminating) population growth. The implementation of these policies has taken the form of full-scale family-planning programs (including the weakening of child-bearing incentives, such as family allowances and paid maternity leave), the establishment of maternal- and child-health services (including contraceptive advice and supplies), and, in certain cases, the legalization of abortion.

The countries that had adopted such official policies by the end of the 1960s were Ghana, Kenya, Mauritius, Morocco, Tunisia, and Egypt. Of these, the only ones showing a decline in the birth rate by 1970 were Mauritius, Tunisia, and perhaps also Egypt. In addition, many other countries have voluntary agencies that provide family-planning advice and services with active government support; there are others that allow such agencies to function unofficially. A few countries, such as Gabon and Cameroon, still have policies encouraging an increase in the birth rate.

Prognostications of population

If birthrates in the African countries remain high and if mortality continues to fall, the population will almost double by 1990 and will be more than three times its 1960 size by the end of the century. The continuation of these vital rates will also further increase the proportion of the population in the younger age groups and thus their dependency on the working population.

The rate of growth of urban populations is expected to continue to accelerate, so that by 1980 more than 100,000,000 persons (22 percent of the population) will be living in towns of 20,000 and more. In large towns in particular, growth rates are expected to vary from 6 to 10 percent. By the end of the century it is likely that between 30 and 40 percent of the population will be settled in towns of 20,000 or more, and—including the smaller towns—at least half the population will be living in an urban environment. (R.K.A.G.)

BIBLIOGRAPHY

General works: PAUL BOHANNAN, *Africa and the Africans* (1964), the best introductory work on modern Africa; GEORGE H.T. KIMBLE, *Tropical Africa*, 2 vol. (1960); R.M. PROTHERO (ed.), *A Geography of Africa* (1969); ALFRED T. GROVE, *Africa South of the Sahara*, 2nd ed. (1970); WILLIAM A. HANCE, *The Geography of Modern Africa* (1964); M.F. THOMAS and G.W. WHITTINGTON (eds.), *Environment and Land Use in Africa* (1969); L. DUDLEY STAMP and W.T.W. MORGAN, *Africa: A Study in Tropical Development*, 3rd ed. (1972).

Geology: RAYMOND FURON, *Géologie de l'Afrique*, 3rd ed. (1968); LUCIEN CAHEN and N.J. SNELLING, *The Geochronology of Equatorial Africa* (1966); S.H. HAUGHTON, *Geological History of Southern Africa* (1969); EBERHARD MACHENS, "Das Präkambrium von Afrika," in FRANZ LOTZE (ed.), *Handbuch der stratigraphischen Geologie*, vol. 13, pt. 2, pp. 414–551

(1968); *Lexique stratigraphique international*, vol. 4, *Afrique* (1969). (*Special studies*): *Bulletin de la Société géologique de France*, vol. 11, no. 2 (1969), issue devoted to the geology of Mauritania; *Annales de la Faculté des Sciences de Clermont-Ferrand*, no. 41 (1969), report of the 5th symposium on African geology; MAURICE LELUBRE, "Chronologie du précambrien au Sahara central (Chronology of the Precambrian in the Central Sahara)," in *Age Relations in High-Grade Metamorphic Terrains*, pp. 27–32 (1969). (*Comprehensive geological maps*): *Structural Outline of Africa*, 1:10,000,000, 1 sheet (1958); *International Geological Map of Africa*, 1:5,000,000, 9 sheets (ASGA–UNESCO, 1963); *International Tectonic Map of Africa*, 1:5,000,000, 9 sheets (UNESCO, 1968).

Physical geography: (*Climate*): B.W. THOMPSON, *The Climate of Africa* (1965), provides background information. Also useful, although somewhat dated, is BERNHARD HAURWITZ and J.M. AUSTIN, *Climatology*, pp. 325–354 (1944). GLENN T. TREWARTHA, *The Earth's Problem Climates* (1961), gives a more advanced account of the major climatic anomalies in Africa (pp. 91–148). (*Relief*): There are several descriptive works on the relief or physiography; one of the most comprehensive of these is LESTER C. KING, *The Morphology of the Earth*, 2nd ed., pp. 241–309 (1967). (*Drainage*): There are numerous works on drainage patterns. A coherent account of drainage, mainly in West Africa, is given in M.F. THOMAS and G.W. WHITTINGTON (eds.), *Environment and Land Use in Africa*, pp. 83–102 (1969). H.E. HURST, *The Nile* (1952), is a classic, and is still the best account of the regime of the Nile and its tributaries. For Southern Africa, the most detailed textbook account is found in J.H. WELLINGTON, *Southern Africa*, vol. 1, pp. 339–508 (1955). The basins of the continent's remaining major rivers and lakes are adequately described in the handbooks, published by the Naval Intelligence Division of the British Admiralty, on *The Belgian Congo*, pp. 11–35 (1944); and *French Equatorial Africa* and *Cameroons*, pp. 39–44 (1942). (*Soils*): H.L. SHANTZ and C.F. MARBUT, *Vegetation and Soils of Africa* (1923), was the first systematic account of major soil groups. Far more modern and useful is S.J. SCHOKALSKAJA, *Die Böden Africas* (1953; orig. pub. in Russian, 1948), which is a basic text. For tropical Africa, particularly, R.P. MOSS (ed.), *The Soil Resources of Tropical Africa* (1968), presents a discussion of the framework within which tropical soils may be described. Among other books and reports for certain regions, PETER M. AHN, *West African Soils* (1970), may be mentioned.

Vegetation and animal life: Most works on African vegetation and animal life are regional, and do not embrace the whole continent. The best broad summary of vegetation south of the Sahara is R.W.J. KEAY'S annotated *Vegetation Map of Africa South of the Tropic of Cancer* (1959), which gives a highly condensed account. There is no good modern handbook of African mammals as a whole, but GLOVER M. ALLEN, *A Checklist of African Mammals* (1939), is valuable, even if out of date and inaccurate. RICHARD LYDEKKER, *The Game Animals of Africa*, 2nd ed. rev. (1926); P.L. DEKEYSER, *Les Mammifères de l'Afrique noire française*, 2nd ed. (1955); and JEAN DORST and PIERRE DANDELOT, *A Field Guide to the Larger Mammals of Africa* (1970), give good accounts of the larger mammal species, and are useful as references to past and present abundance. Bird faunas are discussed in a number of regional textbooks; the most useful are R.D. ETCHECOPAR and FRANCOIS HUE, *Les Oiseaux du Nord de l'Afrique, de la Mer Rouge aux Canaries* (1964; Eng. trans., *The Birds of North Africa, from the Canary Islands to the Red Sea*, 1967); C.W. MACKWORTH-PRAED and C.H.B. GRANT, *African Handbook of Birds*, 6 vol. (1952–), covering respectively East and northeast Africa, the southern third of Africa, and West Africa; and AUSTIN ROBERTS, *The Birds of South Africa*, rev. by G.R. MCLACHLAN and R. LIVERSIDGE (1966). R.E. MOREAU, *The Bird Faunas of Africa and Its Islands* (1966); and B.P. HALL and R.E. MOREAU, *An Atlas of Speciation in African Passerine Birds* (1970), are essential for students of distribution. Butterfly distribution has been discussed along the same lines by R.H. CARCASSON, "A Preliminary Survey of the Zoogeography of African Butterflies," *E. Afr. Wildl. J.*, 2:122–157 (1964). S.H. SKAIFE, *African Insect Life* (1953), is a classic. V.F.M. FITZSIMONS, *Snakes of Southern Africa* (1962), is useful to herpetologists. Two general nature histories, approaching fauna from somewhat different standpoints although complementary to one another, are LESLIE BROWN, *Africa: A Natural History* (1965); and ARCHIE CARR, *The Land and Wildlife of Africa*, together with the editors of *Life* (1964).

Natural resources: (*Hydrology*): JEAN ARCHAMBAULT, *Les Eaux souterraines de l'Afrique Occidentale* (1960), a manual for engineers and technicians dealing with the hydrogeological characteristics of West African states of French expression; ROBERT P. AMBROGGI, "Water Under the Sahara," *Scient. Am.*, 214:21–29 (1966); F.M. COSTER, *Underground Water in Tanganyika* (1960). (*Mineral resources*): UNITED STATES BUREAU OF MINES, *Mineral Facts and Problems* (published every five years), is a standard reference work on mineral commodities in the world. Recent mineral maps of Africa include the *Mineral Map of Africa*, 1:10,000,000 (ASGA-UNESCO, 1968) and explanatory note (1969); and S.A. MUSYLEV, *Map of Mineral Resources of Africa*, 1:10,000,000 (Ministry of Geology of the U.S.S.R., 1966) and explanatory note (1970). The ECONOMIC COMMISSION FOR AFRICA (ECA), *Survey of Economic Conditions in Africa* (1967) and the *Yearly Statistics of Industrial Production;* the INSTITUTE OF GEOLOGICAL SCIENCES, *Statistical Summary of the Mineral Industry* (HMSO, annual); MINING JOURNAL, *Mining Annual Review* (annual); UNITED NATIONS, *Statistical Yearbook* (annual); and the UNITED STATES BUREAU OF MINES, *Minerals Yearbook* (annual), are all periodical works describing new developments in exploration, mining, processing, marketing, and technology, together with statistical data about the mining industry throughout the world. General reviews of the mineral resources of the continent may be found in RAYMOND FURON, *Les Resources minérales de l'Afrique*, 2nd ed. (1961); NICOLAS DE KUN, *The Mineral Resources of Africa* (1965); and UNESCO, *A Review of the Natural Resources of the African Continent* (1963). Reports on individual commodities are published and brought up-to-date by many organizations, of which the four following works are good examples: ECA, *East Africa and the Aluminum Industry* (1965); OECD–ENEA, *World Uranium and Thorium Resources* (1965); UNITED NATIONS, *The World Market for Iron Ore* (1968); UNITED STATES BUREAU OF MINES, *Copper: Materials Survey* (1965). In addition, maps for specific commodities are published, such as the ECA–ASGA, *Coal Map of Africa*, 1:10,000,000, 2 sheets and explanatory note (1966); and the ASSOCIATION OF AFRICAN GEOLOGICAL SURVEYS, *Iron Ore Deposits in Africa*, 1:10,000,000 (1970).

Human resources: A.L. ADU, *The Civil Service in Commonwealth Africa: Development and Transition* (1969); K.M. BARBOUR and R.M. PROTHERO (eds.), *Essays on African Population* (1961); LEONARD BARNES, *Africa in Eclipse* (1971); K.A. BUSIA, *Africa in Search of Democracy* (1967); ROBERT CHAMBERS, *Settlement Schemes in Tropical Africa: A Study of Organizations and Development* (1969); BASIL DAVIDSON, *The Africans: An Entry to Cultural History* (1969) and *The African Genius: An Introduction to African Cultural and Social History* (1970); RUTH FIRST, *The Barrel of a Gun: Political Power in Africa and the Coup d'Etat* (U.S. title, *Power in Africa*; 1970); WALTER FITZGERALD, *Africa: A Social, Economic, and Political Geography of Its Major Regions*, 10th ed. (1967); B.W. HODDER and D.R. HARRIS (eds.), *Africa in Transition: Geographical Essays* (1967); ELLIS AYITEY KOMEY and EZEKIEL MPHAHLELE (eds.), *Modern African Stories* (1964); P.C. LLOYD (ed.), *The New Elites of Tropical Africa: Studies Presented and Discussed at the Sixth International African Seminar at the University of Ibadan, Nigeria, July 1964* (1966); P.J.M. MCEWAN (ed.), *Twentieth Century Africa* (1968), and ed. with ROBERT B. SUTCLIFFE, *The Study of Africa* (1965); W.B. MORGAN and J.C. PUGH, *West Africa* (1969); W.T.W. MORGAN (ed.), *East Africa: Its Peoples and Resources* (1969); A.M. O'CONNOR, *The Geography of Tropical African Development* (1971); R.S. WASSING, *The Arts of Africa* (1970; U.S. title, *African Art*, 1968).

Exploration and settlement: SAMUEL WHITE BAKER, *The Albert Nyanza* (1866); HEINRICH BARTH, *Travels and Discoveries in North and Central Africa* (1857–58); VERNEY LOVETT CAMERON, *Across Africa* (1877); HUGH CLAPPERTON, *Journal of a Second Expedition into the Interior of Africa* (1829); DIXON DENHAM and HUGH CLAPPERTON, *Narrative of Travels and Discoveries in Northern and Central Africa in the Years 1822, 1823, 1824* (1826); ROBERT JAMESON, *Narrative of Discovery and Adventure* (1830); MARY KINGSLEY, *Travels in West Africa* (1897); RICHARD and JOHN LANDER, *Journal of an Expedition to Explore the Course and Termination of the Niger* (1832); DAVID LIVINGSTONE, *The Last Journals of David Livingstone* (1874); MUNGO PARK, *Travels in the Interior Districts of Africa* (1800); JOHN HANNING SPEKE, *Journal of the Discovery of the Source of the Nile* (1863); HENRY MORTON STANLEY, *How I Found Livingstone* (1872) and his *Exploration Diaries* (1961)—all narrative accounts of African exploration and travel by the explorers. ANDREW BOYD and PATRICK VAN RENSBURG, *An Atlas of African Affairs* (1962), concise notes on African historical, economic, and political developments; JAMES L. GIBBS, *Peoples of Africa* (1965), a good, comprehensive recent study on sub-Saharan Africa; LEONARD FRANK HOBLEY, *Opening Africa* (1955), a good, concise history of African exploration from early times; HENRY HERSCH HART, *Sea Road to the Indies* (1950), an account of Portuguese voyages to the East Indies; SIR HARRY

HAMILTON JOHNSTON, *History of the Colonization of Africa by Alien Races* (1899), a comprehensive earlier study of the history of colonization in Africa; ALAN MOOREHEAD, *The White Nile* (1960) and *The Blue Nile* (1962), together a most absorbing study of the Nile's exploration in the 19th century; HEINRICH SCHIFFERS, *Wilder Erdteil Afrika* (1954; Eng. trans., *The Quest for Africa*, 1957), accounts of 2,000 years of exploration of Africa.

Commerce: UNITED NATIONS, *Yearbook of International Trade Statistics* (annual), a basic document for statistical data on international trade; ECA, *Africa's Trade: Trends, Problems and Policy Issues* (1968), a brief but concise outline of problems of Africa's external trade.

Transport: E.W. BOVILL, *The Golden Trade of the Moors*, 2nd ed. rev. (1968), describes the influence of the trans-Saharan trade routes on the peoples north and south of the desert. The IMPERIAL ETHIOPIAN GOVERNMENT HIGHWAY AUTHORITY, *Ethiopia: Progress in Highway Transport* (1967), is a concise description of road building and road transport industry in Ethiopia. The AFRICAN DEVELOPMENT BANK, *Survey of African Regional Transport Studies* (1969), is an inventory, covering the period 1960–68, of transport studies that affect more than one African country.

Resource development: UNITED NATIONS, *World Energy Supplies*, Series J, no. 11, 12, and 13 (1968–70); NEVILLE RUBIN and WILLIAM WARREN, *Dams in Africa* (1968); ECA, *Energy Development in the Countries of the East African Sub-Region* (1965); *Energy Development in the Countries of the West African Sub-Region* (1966); and *Energy Development in the Countries of the Central African Sub-Region* (1967).

Demographic patterns: ECA, *Demographic Handbook for Africa* (1968), is a comprehensive volume containing data and analysis of demographic levels and trends. See also the UNITED NATIONS, *Demographic Yearbook* (annual) and *World Population Prospects As Assessed in 1963* (1966), which contains population projections and projections of mortality and fertility trends; ECA, *Size and Growth of Urban Population in Africa* (1968), which assesses levels and trends of urbanization in the different African countries and subregions; *Europa Yearbook*, containing the latest population estimates of African cities and towns; and the publications of the INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES, a major source for data on French-speaking African countries. Other works include JOHN C. CALDWELL and C. OKONJO (eds.), *The Population of Tropical Africa* (1968), a compilation of papers presented at a conference in Ibadan, Nigeria in 1966; WILLIAM A. HANCE, *Population, Migration, and Urbanization in Africa* (1970); and WILLIAM BRASS *et al.*, *The Demography of Tropical Africa* (1968), a comprehensive volume analyzing demographic data.

(D.S.H.W.N./R.L.C.F./Ro.W.St./
K.B.D./L.H.B./F.S.A./R.K.A.G.)

African Languages

Although no definitive count exists, the number of separate and distinct African languages, as estimated by various authorities, ranges from 800 to more than 1,000. Thus, Africa is unquestionably the world's most polyglot continent. The linguistically most homogeneous area of Africa is the northern part, in which Arabic predominates from Egypt to Mauretania, albeit in a number of sharply distinguished dialects; the most important dialect cleavage is between the Egyptian–Sudanese varieties and those of the Maghrib (from Libya westward). Intermingled with the Maghrib dialects are the Berber languages, concentrated principally in Algeria and Morocco. They range as far east as the Siwa Oasis in western Egypt, as far west as the Senegalese–Mauretanian border area, and as far south as the southern rim of the Sahara. The nomadic Tuaregs speak a Berber language. Nubian, a totally different language, is spoken as far north as southern Egypt along the Nile, but its links are clearly with the languages of sub-Saharan Africa.

Sub-Saharan languages

In sub-Saharan Africa the linguistic picture is far more complex. Except for the Khoisan (Bushman and Hottentot) languages of the extreme south, approximately the entire southern third of Africa is occupied by the relatively closely interrelated Bantu languages. They extend eastward from roughly the Nigerian–Cameroon border to the Indian Ocean. Bantu and non-Bantu languages are considerably interspersed around the northern Bantu border, running through Cameroon, slightly south of the northern boundary of Zaire, and through Uganda and Kenya. There are Bantu enclaves as far north as Somalia and non-Bantu enclaves in northern Tanzania. It is the area north of the Bantu and south of the Sahara that is linguistically the most diverse; there the languages are most numerous and their interrelationships most remote and difficult to establish with certainty.

This article is divided into the following sections:

I. General considerations
  Language classification
  Similarities and correlations
    Common features of the African languages
    Linguistic, racial, and cultural interrelations
II. Niger-Congo languages
  Classification
    West Atlantic subgroup
    Mande subgroup
    Voltaic (Gur) subgroup
    Kwa subgroup
    Adamawa-Eastern subgroup
    Benue-Congo subgroup
  Adoption of foreign words and sounds
III. Chari-Nile and Nilo-Saharan languages
  Classification
    The larger Nilo-Saharan group
    Languages of the Chari-Nile group
    Controversies concerning classification
  Linguistic characteristics
    Phonology
    Grammar
    Writing
IV. Khoisan languages
  Classification
  Linguistic characteristics
    Phonology
    Grammar
    Vocabulary
    Writing and texts

I. General considerations

African lingua francas

Because probably not more than 30 African languages have more than 1,000,000 native speakers, a number of lingua francas have developed in various areas as a means of coping with this enormous linguistic diversity. Arabic, in addition to having the largest number of native speakers of any language of the continent, is sometimes used as a lingua franca, in its literary form by educated non-Arab Muslims, and in a number of colloquial varieties by much of the non-Arab population of The Sudan and Chad. Swahili, a Bantu language heavily influenced by Arabic, is official in Tanzania and Kenya and is used as a lingua franca throughout most of East Africa including eastern Congo. Other important lingua francas include Lingala, also a Bantu language, in western Congo; Fanagalo (vulgarly known as "kitchen Kaffir" or "mine Kaffir"), a pidginized form of Zulu with many English and Afrikaans loanwords, in South Africa, particularly in the mines; and Sango, a pidginized form of Ngbandi (included in the Adamawa-Eastern subdivision of Niger-Congo by some scholars; see below) with many French loanwords, spoken in the Central African Republic.

In addition, there are Bambara-Maninka, a Mande language used in Mali, Guinea, and Ivory Coast; Hausa, spoken in northern Nigeria and neighbouring areas; Amharic, used throughout Ethiopia, where it is the official language; Wolof, the language of Dakar, widely spoken throughout Senegal and The Gambia; and Kongo (Kikongo), a Bantu language that is used in the area of the mouth of the Congo River and is also spoken in a pidginized variety known as Kituba.

Standard European languages, particularly English and French and, to a lesser extent, Portuguese and Italian, also serve as lingua francas, generally among the better educated, in areas in which they are or were the colonial languages and where they enjoy or have enjoyed official status. Localized varieties of European languages are also spoken in numerous parts of Africa, such as Pidgin English, a widely used lingua franca of Cameroon and West Africa; Krio, a creolized form of English, differing slightly from Pidgin English and used by the local population of Freetown, Sierra Leone; various forms of Creole Portuguese, spoken in the Cape Verde Islands,

the islands of São Tomé and Príncipe, and Portuguese Guinea; and Afrikaans, an outgrowth of Dutch, which serves along with English as an official language of South Africa.

LANGUAGE CLASSIFICATION

Early attempts at classification

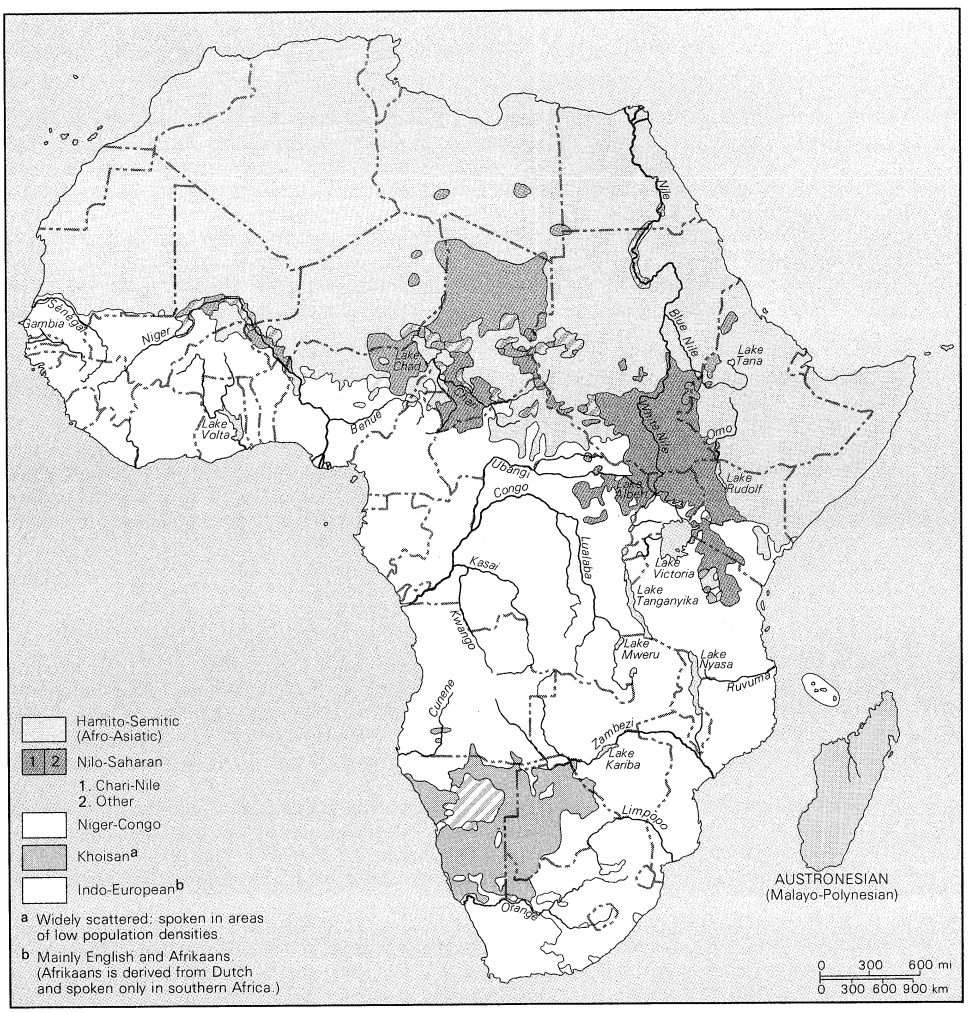
Although word lists of a number of sub-Saharan languages were recorded by Europeans as early as the 15th and 16th centuries, and the first grammar of a sub-Saharan language (Kongo) was published in 1659, extensive study of African languages and systematic attempts at grouping and classifying them did not begin until the 19th century. In 1854 Sigismund Koelle, a missionary living in Freetown, Sierra Leone, was able to collect substantial word lists of roughly 150 languages, principally of West Africa but also including some belonging to the Bantu family. He established 11 groups, leaving approximately 40 languages unclassified. His groups are generally small, rarely exceeding 15 members, and he made no systematic attempt to explore wider relationships. He laid the groundwork for establishing the Mande family, which has been universally accepted ever since. (The term "family" is used loosely in this article, sometimes referring to large groupings of languages and sometimes referring to subgroups or branches of these groupings. Niger-Congo, for example, is usually called a language family, but its subgroups [*e.g.*, Mande] and its sub-subgroups [*e.g.*, Bantu] are often referred to as families as well.)

The unity of the Bantu family, although suggested by a number of earlier writers, was first established in 1862 by the German scholar Wilhelm Bleek, who gave it its name, the word for "people" in most languages of the family. During approximately the same period, the work of several scholars suggested that Ancient Egyptian, the Berber languages, certain languages of northeastern Africa called Cushitic (*e.g.*, Galla, Somali, Beja, and Afar [Danakil]), and perhaps others (*e.g.*, Hausa) were remotely related to one another and to the Semitic family. These languages were called Hamitic at the suggestion of a French Semitist, Ernest Renan, who was also the first to propose the name Cushitic, because the peoples speaking these languages coincided roughly with the descendants of Ham (one of whom was Cush), according to the biblical account in Genesis, just as the people speaking the Semitic languages coincided with the descendants of Shem (Sem).

The first attempts at an overall classification of African languages were those made by the Austrian linguist Friedrich Müller in 1876–88 and by the German Egyptologist Karl Richard Lepsius in 1880. Müller, who classified all the world's languages on the basis of a correlation of linguistic and racial groupings (the latter determined largely by hair type), set up the following divisions within Africa: (1) Bushman-Hottentot, as a subdivision of the languages of the tufted-haired peoples (the Papuan languages of New Guinea formed the other subdivision); (2) Bantu and (3) Negro, both subdivisions of the languages of the fleece-haired (*vliesshaarig*) peoples; (4) Hamitic, (5) Semitic, and (6) Nuba-Fulah, all three subdivisions of the languages of the wavy-haired (*lockenhaarig*) peoples. Linguistically, Nuba-Fulah was completely heterogeneous and was abandoned by later scholars.

Lepsius' classification

Lepsius proposed an alternative classification of the languages and peoples of Africa, of which the three major divisions were (1) Semitic; (2) Hamitic, including Hausa, which he grouped with the Berber or Libyan branch, and Hottentot, neither of which had been classed as Hamitic



Major language groups in Africa.

by Müller; and (3) the Negro languages, consisting of two subdivisions, Bantu and Mixed Negro, composed of many subgroups, and containing all the indigenous languages of Africa not included in the other divisions.

During the early part of the 20th century, a revision of this classification emerged from the work of the two leading German Africanists, Carl Meinhof and Dietrich Westermann. Meinhof added a number of languages to Lepsius' Hamitic grouping, namely Fulani, a major West African language, and a group of East African languages, including Masai and Bari. Westermann attempted to show the unity of virtually all the non-Bantu, non-Hamitic, and non-Semitic languages of Africa; he designated them Sudanic. Thus, a threefold division was established: Bantu, Hamitic, and Sudanic. The Bushman languages of southern Africa were sometimes treated as a separate group and sometimes linked with Sudanic, and Semitic was generally linked to Hamitic, though not included within it. Not everyone accepted this classification. French scholars, in particular Maurice Delafosse and Lilias Homburger, restricted Hamitic to Ancient Egyptian, the Berber languages of North Africa, and the more heterogeneous Cushitic group of languages of northeastern Africa. They connected Hottentot with the Bushman languages while grouping the other languages classed by Meinhof as Hamitic in the very general Negro-African or Sudano-Guinean family, to which they also believed the Bantu languages to be affiliated.

Westermann himself revised his views of the earlier classification scheme and came to regard the great majority of the languages of West Africa (*i.e.*, the Western Sudanic languages) as more closely related to Bantu than to the languages farther east, previously regarded as Sudanic. Sir Harry Johnston, a British Africanist, likewise pointed out similarities between Bantu and many West African languages that he termed Semi-Bantu.

Joseph Greenberg's classifications

No single scheme was generally accepted in 1955, when Joseph Greenberg, a United States linguist, presented the earlier version of his major classification of the languages of Africa. In 1963, Greenberg made public a substantially revised version of his work that consolidated previously unclassified languages and smaller language families into four large linguistic stocks—Niger-Kordofanian, Nilo-Saharan, Khoisan, and Afro-Asiatic (Hamito-Semitic). They included all the indigenous languages of Africa except Malagasy, the language of Madagascar, which belongs to the Austronesian (Malayo-Polynesian) family and is thus unrelated to the languages of the continent. The Niger-Kordofanian group consists of two branches, Niger-Congo, which is very extensive, and the much smaller Kordofanian group, so named because all of its members are located in Kurdufān province, The Sudan. Kordofanian contains a number of subgroups, of which Kadugli-Korongo is most distantly related to the others. In fact, its inclusion within the group is open to serious question. Niger-Congo consists of Westermann's Western Sudanic group plus Bantu, whose interrelationship had earlier been advocated by Westermann. Greenberg differed from Westermann, however, in that he treated Bantu as a subgroup of one of the branches of Western Sudanic rather than an independent unit related to Western Sudanic as a whole. Greenberg also argued strongly for the inclusion of Fulani in this family. (Fulani had been regarded as Hamitic by Meinhof, but some French scholars had already advocated its relationship to other West African languages, such as Serer of Senegal.) Furthermore, Greenberg extended Niger-Congo eastward to include languages of Central Africa, such as Ngbandi of the Central African Republic and Azande (Zande) of Zaire, with which Westermann had not dealt.

The Nilo-Saharan group consists of six branches, of which Chari-Nile is by far the most heterogeneous. This family is the least adequately substantiated of the four. The Khoisan languages comprise the Bushman and Hottentot languages of southern Africa and two languages of Tanzania, Sandawe and Hadza (Hatsa). Afro-Asiatic, which is Greenberg's name for Hamito-Semitic, consists of five branches: (1) Semitic, represented in Africa by Arabic and the Ethiopic languages, of which the most important are Amharic, the official language of Ethiopia; Tigrinya, the predominant language of Eritrea; and Ge'ez, the liturgical language of the Ethiopian Church, which is no longer spoken; (2) Ancient Egyptian and its daughter language, Coptic, now extinct; (3) the Berber languages of northern Africa, including Kabyle of Algeria, Riffian and Shluh (Shilha) of Morocco, and Tuareg, or Tamashek, of the Sahara; (4) the extensive and diversified Cushitic group of northeastern Africa, whose most important members are Somali and Galla; (5) Chadic, spoken in Chad, Cameroon, and northern Nigeria, although Hausa, its best known and most widely spoken member, is used throughout much of West Africa.

Branches of Afro-Asiatic (Hamito-Semitic)

Hamito-Semitic languages also are spoken in the Middle East. For this reason, they will be treated in a separate article (see HAMITO-SEMITIC LANGUAGES).

Because of the inexact nature of the procedures for establishing genetic classification and also because of incomplete data on the majority of African tongues, differences of opinion exist on various aspects of Greenberg's classification. Some scholars, such as the Hungarian linguist István Fodor and the English Bantuist Malcolm Guthrie, insist that genetic interrelationship of a group of languages can be established only by demonstrating regular phonological correspondences among them, as have been shown among the Bantu languages, for example. Greenberg and his supporters, however, maintain that similarities in both sound and meaning between two or more languages, which are too numerous to be dismissed as mere coincidence and are found in that part of language least susceptible to borrowing (*e.g.*, grammatical elements, body parts and functions, common natural objects and phenomena, and lower numerals), must be attributed to genetic relationship. Nevertheless, there are no rigorous procedures for determining how similar two items must be in sound and meaning to be considered evidence for genetic relationship or how many items are needed to rule out other explanations.

SIMILARITIES AND CORRELATIONS

Common features of the African languages. Although African languages are extremely diverse in structure, certain phonological and grammatical features are widespread throughout much of the continent or at least within extensive and well-defined areas.

Phonological features. Tone—*i.e.*, the use of pitch to distinguish words and grammatical forms that otherwise would be phonetically identical—occurs in the overwhelming majority of sub-Saharan languages. The famous click sounds are largely restricted to two language groups of southern Africa: the Bushman and Hottentot (Khoisan) languages and the southern Bantu languages (*e.g.*, Zulu, Xhosa, and Sotho), which borrowed them from the Bushman- and Hottentot-speaking peoples. Clicks are also common to three East African languages, Sandawe, Hadza, and Sanye, but nowhere else on the continent or outside it. The consonants *kp* and *gb*, called coarticulated labiovelar consonants, are, in Greenberg's words,

Tones and clicks

> distributed over a wide belt of languages of diverse genetic affiliations from the Atlantic Ocean almost to the Nile Valley. Outside of Africa these sounds are only known from a restricted area of Melanesia. . . .

Other phonological features common in Africa but rare in the rest of the world are word-initial combinations of a nasal sound plus another consonant (*e.g.*, *mb* or *nd*), and implosive consonants.

Grammatical features. In some African languages, words generally consist of a root only, with grammatical affixes rare and even nonexistent in a few languages. These are often referred to as isolating languages. Other African languages have words that are composed of many elements. Those languages are subdivided into the agglutinative languages (in which each element of a word has a distinct and separate meaning and form) and inflectional languages (in which the various elements of meaning may be fused into such forms as prefixes or suffixes or the entire word may change internally to indicate grammatical relationships).

In some of these languages, a single word might be equivalent to a whole sentence in English or other European languages. Thus, in Swahili, a Bantu language, "we will not hit him" is expressed in a single word, *hatutampiga*, in which *ha* is the negative marker, *tu* the verb subject "we," *ta* the future-tense marker, *m* the verb object "him," and *piga* the verb "hit." The verb likewise may undergo modifications in meaning through the addition of suffixes; thus *pigwa* means "(to) be hit," and *pigana* "(to) hit one another."

Noun classes and gender

A widespread feature among African languages employing words composed of more than one element is a system of noun classes (see below *Benue-Congo subgroup*), particularly marked in the Bantu family but also occurring in a wide variety of West African languages, as well as in certain languages of Kurdufān province, The Sudan. All of these languages have been grouped by Greenberg in his Niger-Kordofanian family, although others in this family lack this feature altogether.

Grammatical sex gender is not nearly as widespread as noun classes. It is characteristic of the Semitic and Hamitic languages (although lacking in some members of the Chadic branch) and of other languages such as Hottentot and the closely related Naron Bushman language of South Africa and of the much more distantly related Sandawe and Hadza languages of Tanzania. Gender is likewise found in East Nilotic but not South Nilotic (these two comprise what was earlier known as "Nilo-Hamitic") or West Nilotic (previously known simply as Nilotic). Gender plays an even more restricted role in other African languages, such as Bongo, a language of the southwestern Sudan in which different pronouns are used for masculine and feminine 3rd person singular. The great majority of African languages, however, do not make this distinction, but many of them do distinguish between animate and inanimate personal pronouns.

Linguistic, racial, and cultural interrelations. It is difficult to find strict correlations between linguistic, cultural, and racial groupings, either in Africa or elsewhere in the world. Groups speaking identical or closely related languages may be strikingly different racially, culturally, or both; while groups that are very similar racially and culturally may speak different and even unrelated languages. Thus the Damara (or Bergdama) of South West Africa, who speak the language of the Nama Hottentot, are sharply distinguished from the latter both racially and culturally. The Damara have much darker pigmentation and other more Negroid physical features and are predominantly nonpastoralists. Likewise, the Arusha of northern Tanzania, who speak the language of the pastoralist Masai, are horticultural (*i.e.*, agricultural without use of the plow). On the other hand, the forest-dwelling Pygmies, who are spread over much of central Africa, form a distinct racial and cultural group and are exclusively hunters and food-gatherers, yet they have no distinctive language. Instead, they speak the languages of their various neighbours, such as Bantu, Moru-Mangbetu (grouped by Greenberg in the Central Sudanic branch of Chari-Nile), and reportedly Sere-Mundu (grouped by Greenberg in the Adamawa-Eastern branch of Niger-Congo).

Broad correlations between race, language, culture

Some broad correlations can nevertheless be made between language groupings and the variables of race and culture, as long as these are not interpreted too rigidly. The Bushmen of southern Africa form a distinct racial, cultural, and linguistic group, and, though the Hottentot are akin to them linguistically and racially, they show important cultural differences, most notably that the Hottentot are pastoralists, not hunters and gatherers. The Berber-speaking peoples of North Africa likewise form a relatively homogeneous racial, cultural, and linguistic group; they are Caucasoid and pursue a mixed agricultural and pastoral way of life. Only the Sahara-dwelling Tuareg are exclusively pastoral nomads and caravaners. The Cushitic peoples of northeastern Africa, who like the Berbers speak languages of the Hamito-Semitic family, represent varying degrees of Caucasoid and Negroid elements. Some of the more prominent among them, such as the Beja or Bedawiye of The Sudan and the Saho-Afar and Somali of the Horn, are nomadic pastoralists. The Galla of southern Ethiopia, the most numerous of all Cushitic peoples, include some pure pastoralists, but the great majority practice agriculture and husbandry. Their neighbours to the north, the Agau of central Ethiopia, do likewise, though they have largely assimilated culturally to their other Semitic-speaking neighbours, while retaining their Cushitic form of speech.

The unusually tall and slender Nilotes of The Sudan, such as the Dinka and Nuer, are sometimes considered a subrace of the Negroid; but it has also been suggested that their physical peculiarities are at least to some extent the result of differences in diet rather than heredity. Herding of cattle and other livestock is predominant among the majority of Nilotic peoples (including the so-called "Nilo-Hamites"), but nearly all practice some horticulture, with the striking exception of certain groups of Masai. Certain Nilotic peoples, however, are predominantly sedentary horticulturalists, such as the Anuak of The Sudan and the Luo, principally in Kenya, who are the most numerous of the Nilotic peoples.

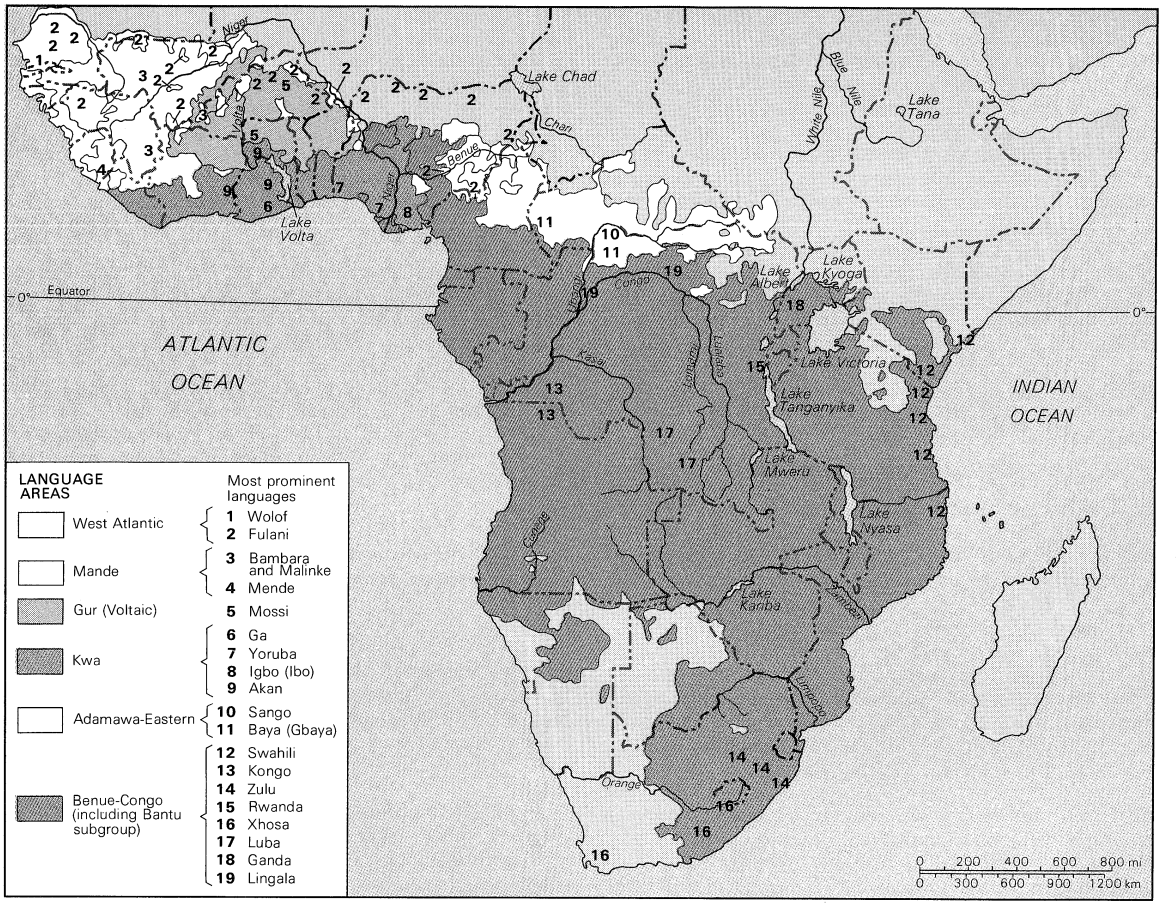
The overwhelming majority of the Bantu and the peoples of western and central Africa are sedentary horticulturalists. Likewise, most of these peoples belong linguistically to Greenberg's Niger-Congo family, although some fall within Nilo-Saharan, and others within the Chadic branch of Hamito-Semitic; furthermore, they comprise the overwhelming majority of the Negro race, which, however, is no more homogeneous a physical group than the peoples of Europe. The most important group of pastoralists in West Africa are the cattle-herding Fulani, who are widely distributed over the savanna belt just south of the Sahara. They are physically as well as culturally distinct from their neighbours, but linguistically their closest affinities are with the sedentary horticultural Negroes of Senegal. Cattle raising likewise plays an important role among certain Bantu peoples of eastern and southern Africa, such as the Ganda, Nkole (Ankole), Sotho, and Nguni peoples. Among some of the East African Bantu (*e.g.*, the Nkole, Rwanda, and Rundi) cattle ownership has been the prerogative of a privileged group probably of Nilotic origin and known by various names, such as Hima (Panda) and Tutsi; these people are physically and culturally distinct from their agricultural neighbours, although totally assimilated to them linguistically. (M.F.Go.)

II. Niger-Congo languages

The Niger-Congo family is the largest family of languages in sub-Saharan Africa, with 890 known member languages. All are considered to be distinct languages and not dialects. The named dialects of these languages number in the many thousands; if the variant names for the same language or for the same dialect are to be counted, the figure must be further expanded. For example, for Swahili, there are 17 named separate dialects, 15 additional variant names for some of the dialects, and at least four "nicknames" for popularized or debased forms of the language.

The classification of such a large and diverse family was first accomplished in 1949 by Joseph Greenberg, who demonstrated the genetic unity of this group on the basis of an examination of documentation of all the languages of Africa. The primary sources for such a massive comparison consist of published descriptions or vocabularies of tribal languages made by missionaries or colonial officers; these were written in European languages because practically none of the African languages had any written tradition of their own. Usable documents stretch back for only about 125 years, and there are still indigenous languages that either have not been reported at all or lack sufficient detail or accessibility to be included in any final count of Niger-Congo languages. Thus the figure given here may have to be revised upward, especially as there is no evidence that the indigenous languages of sub-Saharan Africa are dying out. On the contrary, of a list of such languages collected in the mid-19th century, only one is known to have become extinct.

The area in which the Niger-Congo languages are



Distribution of the Niger-Congo languages.

Location of the Niger-Congo languages

spoken is fairly easily demarcated (see map) and includes the entire sweep of sub-Saharan Africa with the exception of the Horn of Ethiopia. All indigenous peoples of sub-Saharan Africa west of the Nile are speakers of Niger-Congo languages with the exception of a few to the extreme south (Bushman–Hottentot peoples) and a more substantial number along the northern rim of the area (speakers of Chadic or Nilo-Saharan languages).

A conservative estimate of the total number of speakers of Niger-Congo languages is 181,000,000. Such a figure is obtained by adding the populations (as of the middle 1960s) of the 32 countries of tropical and southern Africa and then subtracting the number of European and Asian speakers reported for South Africa. (It is assumed that the number of speakers of Niger-Congo languages in the countries to the north of the tropical belt omitted by this method—especially in Mali and Upper Volta—would be balanced out by the number of indigenous non-Niger-Congo speakers enumerated in the included populations—especially in Nigeria and Uganda.)

Table 1: Niger–Congo Languages with More Than Three Million Speakers

	subgroup	estimated number of speakers*
Fulani (Fulbe)	West Atlantic	7,360,000
Malinke (Mandingo), *et al.*	Mande	3,010,000
Mossi	Voltaic (Gur)	3,800,000
Yoruba	Kwa	6,580,000
Igbo (Ibo)	Kwa	6,430,000
Akan group	Kwa	3,300,000
Rwanda (Banyaruanda)	Benue–Congo	4,880,000
Kossi (Kosa)	Benue–Congo	3,470,000
Makua	Benue–Congo	3,460,000
Zulu	Benue–Congo	3,110,000

*Figures rounded off to the nearest 10,000.

CLASSIFICATION

The Niger-Congo family is divided into six genetic subgroups. Viewed on a geographical continuum from northwest to southeast, these are:

1. West Atlantic, 43 languages
2. Mande, 26 languages
3. Voltaic (Gur), 79 languages
4. Kwa, 73 languages
5. Adamawa-Eastern, 112 languages
6. Benue-Congo, 557 languages

This subclassification results from very fundamental and very old divisions in the language family, and each subgroup represents the results of separate developments that can be estimated to be about 5,000 to 10,000 years old at the least. Obviously, such a deep division has yielded widely separated languages—separate not only in geographic distribution but in elements of the sound systems, grammars, and vocabularies as well. It is surely safe to say that there is no one feature common to the entire group and that the recognition of their relationship depends upon a truly global view of the evolution of a few separate ancestral languages into the thousands of descendant languages spoken throughout the world today. Because this linguistic evolution is neither random nor chaotic, it is possible by means of a very careful plotting of common clusters of features to establish the approximate degrees of relationship even when neither the intermediate nor earliest stages of the protolanguages concerned have been reconstructed. Greenberg used this method to include the nearest relatives of the Niger-Congo languages with them in a superfamily that he designates as "Niger-Kordofanian." This grouping contains two branches: all the languages here considered, namely the Niger-Congo family, and a coordinate branch of the Kordofanian family of languages, spoken in the Nuba Hills (Jibāl an-Nūbah) of Kurdufān province, The Sudan.

Niger-Kordofanian

West Atlantic subgroup. The centre of gravity of the 43 West Atlantic languages would seem to lie in the Senegal–Guinea area, in which are found the two most important languages—Wolof and Fulani (the latter also

known as Fulbe, Fula, and Peul). The area is also the locus of the great majority of the other languages. Fulani, the language of a highly mobile pastoralist people numbering as many as 6,000,000, is also found in a scattered distribution as far east as Chad. The immense extension of this language and its association with cattle and Islām led earlier researchers to give it a special place outside the other West Atlantic languages—in fact, outside what is now recognized as Niger-Congo. It was first isolated in 1883 by an Englishman, Robert Needham Cust, who based his work on that of Friedrich Müller, an Austrian linguist. The German Karl Richard Lepsius classified it as one of the "mixed Negro" languages not fully sub-Saharan or Bantu (1880). Subsequently, it was included as a "Hamitic" language in the German Africanist Carl Meinhof's influential classification (1912), which was adopted by many scholars writing in English.

A feature of Fulani that is shared systematically with some of the other West Atlantic languages is an "alternation" whereby both the beginnings and the endings of words go through parallel changes according to grammatical considerations. This feature is found in its greatest elaboration only in Fulani; it is represented in either vestigial or undeveloped form in most of the other West Atlantic languages. This fact made Fulani appear more different from the others than it actually is and contributed to its misclassification. It is of interest to speculate why this most "vigorous" of the West Atlantic languages should be the most difficult or complex in its grammatical system. ("Vigorous" is used here in the sense that it has several millions of speakers, and also in the sense that it is associated with a dynamic expansionism). By comparison, Latin indicated grammatical case, gender, and number only by the endings of words, and English has virtually eliminated case and gender and has largely simplified number to a single suffix.

Bambara and Mende languages

Mande subgroup. The Mande subgroup includes 26 languages, of which the most prominent are Bambara of the Mali–Guinea area and Mende of Sierra Leone. In contrast to the classification of the West Atlantic subgroup, there has been little doubt as to the identification of the languages in the Mande subgroup. This certainty results from the subgrouping conforming closely to the "classic" expectation of language differentiation, in which a small number of contemporary languages are clearly seen to share a number of striking features not found elsewhere. Such features presumably provide evidence of early segregation of the ancestor language of the subgroup and of a continued effective isolation. Examples of such key features in the Mande languages are the systematic distinction between free and dependent nouns (see below) and the form and use of numerals.

As would be expected, however, the results of a longer period of independent development have also deflected the attention of many researchers to the differences of Mande from the Niger-Congo group—*i.e.*, to the common negative evidence of relationship with the rest of the Niger-Congo family. This negative evidence led Maurice Delafosse, in 1924, to propose two criteria for the identification of this subgroup: (1) a complete absence of the noun-class system that characterizes other West African languages and (2) a complete absence of tonal contrasts. Investigators who view Mande languages in terms appropriate to the group itself have shown that this is a prejudiced view and that a more positive formulation of these criteria is possible. Such a restatement would run somewhat as follows: (1) Instead of the particular distinctions among nouns reflected in the grammatical classes of the other subgroups of Niger-Congo, the Mande subgroup has elaborated a grammatical distinction between free and dependent nouns—nouns referring to "alienable" or "transferable" possessions, as opposed to "inalienable" possessions such as kinsmen or parts of the body. (2) The role of tone is tied to features other than those of semantic reference (*i.e.*, the meaning of the word taken in isolation).

Whereas Chinese and other tone languages use changes of pitch to distinguish meaning in words with the same configuration of consonants and vowels (when pronounced by themselves), the Mande languages reveal a different use of tone, one that also differs from practically all the other Niger-Congo languages. In the Mande languages, a difference in meaning based on tone seems to demand a contrast of complete utterances, which are usually made up of more than one word. This contrast of entire utterances places Mande at one end of a continuum of distinctions based on tone, because linguists are still finding many variations in the situation of "meaningful semantic tone" or "contrastive tone patterns in isolation" that seem to demand some sort of context in order to be operative. In this respect, it is of interest that most of the differences in meaning between complete utterances that are identical except for tone in the Mande languages are more of a grammatical nature than of a semantic one. It must be stressed, however, that the Mande languages *are* tone languages and that the contrasts so produced are discrete and absolute and thus not of the "sliding" nature of speech melody. In form, they are closer to the question intonation of English, but in function they perform grammatical work internal to the sentence rather than coextensive with it.

Role of tone in the Mande languages

It is striking that several independent writing systems based on the syllable (which is the unit that bears the tone) have been reported for Mande languages. The best known of these is the one devised for the Vai language, although Mende, Toma (Loma), and Kpelle also have indigenous syllabic writing systems. This type of writing is somewhat reminiscent of the writing system of Chinese (in which the syllable is often both the tone-bearing unit and the meaning-bearing unit) and of Cherokee, an American Indian language.

Voltaic (Gur) subgroup. The Voltaic (Gur) subgroup includes 79 languages spoken in Upper Volta and the upper halves of Ghana and Ivory Coast. Mossi is the most prominent of them. The term Gur was used as early as 1886 to designate the languages centred around Ouagadougou in what is now Upper Volta and northern Ghana. Languages with such names as Gurma or Grunshi (Gurunsi) were later related, and the first syllable of the names was taken as a name for the group. In 1911 Delafosse used the designation "Voltaic" for the same group, probably because many of the Gur languages are spoken in the Volta River Basin.

The study of the Voltaic languages as a group is important because they do not partake strongly of any entrenched typological or grammatical "manifestations" (such as the "manifestations" of sex gender distinctions in Indo-European, triconsonantal roots in Semitic, and so on). A wide range of possibilities is present, and, if any "manifestation" is at all apparent, it is a tendency toward simplification and serial constructions. This means that there are a very small number of closed grammatical classes and a greater reliance on constructional markers. Nevertheless, it is important, when considering any of the 1,000 or more languages surveyed in this article, to emphasize that each language within the subgroup is also important for its own sake. Each language reveals a particular way of communicating about the world, a way that is fully developed and elaborated in its own terms. These spoken languages are certainly not "underdeveloped" or primitive. In fact, the very complexity of analyzing the Voltaic languages (or any other group of natural languages) in terms and categories sufficiently "universal" to be operational both for those who already speak the languages and those who want to learn something about them reflects their developed nature.

Typological significance of the Voltaic languages

The most widely known characteristic of the Voltaic (Gur) languages is the realization of the Niger-Congo noun classes in the shape of parallel prefixes and suffixes, in which the suffixes appear to be dominant. In languages like Mõõre and Dagbane, for instance, nouns for a "person" end with the suffix *-a* (*-u* in Kasem), which is replaced by *-ba* in the plural. This feature is somewhat reminiscent of Fulani of the West Atlantic subgroup but differs in that it typically appears in the form of separable syllables in the Voltaic languages, rather than as an alternation or mutation of the beginning and ending consonants of the root itself, as in Fulani. The partial classifi-

cation of human beings, animals, and liquids into separate grammatical classes provides evidence for the original nature of such classes in the ancestral Niger-Congo language. So does the fact that this classification governs the participation of variant forms of the detachable suffixes (or suffix–prefix frames) in sentence constructions calling for pronoun-like behaviour—referential, possessive, demonstrative, relative, and so on. Thus, Bargu has the suffix *-a* for the "person" class, a suffix that always "governs" the occurrence of *yé* as independent pronoun, *yù* as object marker, and *yè* as one demonstrative.

Ga, Yoruba, Igbo: prominent Kwa tongues

Kwa subgroup. The Kwa subgroup includes 73 languages, many of which are prominent (*e.g.*, Ga, Yoruba, Igbo). This grouping includes some of the most well known of the Niger-Congo languages and some with the greatest number of speakers. The Kwa languages are spoken in the belt of tropical forest through the southern section of the West African bulge. They are the languages of the former great kingdoms of tropical West Africa, such as Ashanti (Akan or Twi language), Dahomey (Ewe), Oyo (Yoruba), and Benin (Bini language). In addition, the large decentralized group of Ibo (Igbo) also speak a Kwa language. Each of these languages has millions of speakers and a growing written literature, and one of them, Yoruba, boasts the earliest published dictionary and grammar written by a native speaker. It is symptomatic of the position of these languages in modern life, however, that these Yoruba works were written in English by an African Anglican churchman, Bishop Samuel Crowther. English or French remains the national language of the countries in which Kwa languages are spoken, and interest in employing the local languages is largely restricted to religious or academic institutions based on European models.

Like the Voltaic subgroup, the Kwa languages are diverse in their structure and characteristic features, and examples of this diversity are found in the roles of tone and vowel harmony. The continuum from clear tones in isolated dictionary citations to modification in sentence context referred to in connection with the Mande languages is particularly striking in the Kwa languages. Thus, some investigators heard consistent pitch levels that could be noted on a musical staff for any one speaker (in at least one Kwa language, Grebo of Liberia, there are four such levels). Others, such as those investigating Akan, heard just as clearly that there is a complicated system of pitches that are varied according to transitions between syllables or interrelations among syllables. It is as yet unclear just which, if either, of these systems is historically earlier: whether the clear tone levels are a crystallization of the distinctions imposed by contextual reactions with other tones or whether the languages that have almost lost all independent lexical tone are a newer reinterpretation of a previous system of fixed pitches.

Vowel harmony in the Kwa tongues

Vowel harmony involves a system of mutual selection or restriction among the vowels that can go into the makeup of a word. The vowels in a word are said to be in "harmony," with high vowels usually occurring with high vowels, front vowels with other front vowels, and so forth, in such a way that a constant feature seems to be realized throughout the word. Vowel harmony of a type peculiar to sub-Saharan languages is called "cross-height" vowel harmony by the British scholar John M. Stewart. In this situation, the higher of two varieties of a vowel selects the matching higher variety of another vowel and vice versa. Thus, *u* (as in "rule") would occur with *o* (as in "bone"), and the lower sound ʊ (as in "pull") with the correspondingly lower ɔ (as in "law"). Stewart also summarized evidence for contrast in the formation of the sounds in terms of the advancement of the root of the tongue. Under conditions of change of a vowel system, this leads to reinterpretations that cloud the nature of earlier contrasts and add to the variety found in the Kwa languages.

It is very difficult to provide a generalized "profile" of the Kwa languages that would give an idea of what is typical of a language of this subgroup. A stereotype would be that the words are mostly short and the nouns begin with vowels. Proper names often appear to be quite long, however, because they are compounds or phrases made up of several words. Many words of Kwa origin are found in the Americas, as is the pantheon of Yoruba gods in South America and the Caribbean.

Adamawa-Eastern subgroup. The Adamawa-Eastern subgroup comprises 112 languages spoken in the Central African Republic and northern parts of Cameroon and Zaire. These geographically remote languages are the least known of the Niger-Congo family. The subgroup's most prominent members are Sango and Baya (Gbaya). The name Adamawa-Eastern implies a combination of two recognized sub-subgroups: Anamawa and Eastern. None of the language names of this subgroup is well-known to outsiders, with the possible exceptions of the ethnographically noted Azande (Zande) and the trade language Sango.

Sango is clearly simplified from a dialect of one of the Adamawa-Eastern languages. It shows the result of contact with French as well as the usage of many river tribes in trade along the Ubangi River. It is the lingua franca of the Central African Republic and is extremely widespread as a second language in Central Africa. Many other pidginized trade languages based on an African vernacular are also spoken in this central area of the great rivers; *e.g.*, a pidgin dialect of Swahili, Kituba (a pidgin form of Kongo [Kikongo]), and Lingala.

Benue-Congo subgroup. The Benue-Congo subgroup comprises 557 languages, spoken from Nigeria to South Africa. Its most prominent members include Swahili, Kongo (Kikongo), and Zulu. In terms of numbers of speakers and geographical extension, Benue-Congo and Bantu are practically the same entity. That is, compared with the many millions of speakers of Bantu, the speakers of non-Bantu languages are only a tiny minority of Benue-Congo speakers. Furthermore, the Bantu speakers are spread out over most of middle and southern Africa, whereas the remainder of Benue-Congo speakers occupy only a corner of the northwest extension, mostly in Nigeria. The only two major non-Bantu Benue-Congo languages are Efik and Tiv, found in eastern Nigeria.

Bantu, largest Benue-Congo subgroup

The following description sacrifices a consideration of the numerous but less important non-Bantu languages of the subgroup for a still superficial view of the Bantu languages. The most notable feature of the Bantu languages is their system of noun classes (concord). This has also been called "alliterative" concord because it often entails a repetition of the same syllable as prefixes at the head of successive words in the sentence. An example, drawn from Swahili, provides an illustration: *wa-tu wa-le wa-mefika* (noun–demonstrative–verb) "those people have arrived." Such prefixes as *wa* represent an intersection of two systems: one syntactic (function in the sentence) and the other paradigmatic (mutually exclusive replacements in the same position in the sentence). Syntactically, the shape of the prefix can vary according to the part of speech to which it is attached—a prefixed syllable may be different in the noun, adjective, possessive, locative, or verb. That there is agreement (*i.e.*, a concord element is present) with virtually all parts of speech creates an aesthetically pleasing system by means of an exhaustive display of marked relationships in the sentence.

Paradigmatically, the system of concord is most simply viewed as being based upon the governing noun in its singular or plural form. That is, the noun belongs to one of a number of noun classes that determine the proper set of concording elements to be used throughout the part of the sentence dominated by that noun. These sets (which can be thought of as replacement sets) come in pairs according to whether the governing noun is singular or plural; *e.g.*, in Swahili, the singular form *m-tu yu-le a-mefika* "that person has arrived" contrasts throughout with the plural form *wa-tu wa-le wa-mefika* "those people have arrived." This alternation of singularity and plurality is, of course, more closely related to the semantics of the noun than to its purely grammatical function. The paradigmatic system goes even further in this direction by making a rather loose use of other possible semantic contrasts (especially: person–nonperson and countable–noncountable) in assigning nouns to noun

Noun concord in Swahili

classes. In no language, however, is there a strict relation between meaning and noun class. Nevertheless, the tendency in this direction again stimulates a sense of completeness of aesthetic form and has added to the fame of Bantu languages as abstract figures.

Correlates of these noun classes (in both form and function) are found in all branches of Niger-Congo except Mande and only slightly in Kwa. This led some early researchers to combine all of them as one subgroup and later led to a typological distinction between noun-class and non-noun-class languages. This distinction is not very useful, however, because there are a number of languages with rudimentary or partial noun-class systems, and it even seems possible for a language of entirely different ancestors to adopt noun prefix alternation.

Swahili remains the best known Bantu language. It is the national language of Tanzania and is used as a lingua franca throughout East Africa. The use of Swahili for this purpose is facilitated because it no longer makes use of tone—it is an isolated non-tone language in the Bantu area. There is another phenomenon related to tone in the Bantu languages besides its disappearance in Swahili that should be mentioned: there exists an area of "tonal reversal," a geographic area centred in the western Congo in which the tonal realizations in words and sentences are the mirror images of what they are elsewhere in Bantu languages; *i.e.*, high tones become low tones, and vice versa.

Area of tonal reversal

Another distributional characteristic is the presence of a prefix form made up of two syllables as opposed to the form with only one. Following out the implications of the geographic distribution as well as the function of this feature is of help in mapping the prehistory of the Bantu languages. As in all other areas of the world, however, it is the distribution of certain items of vocabulary, or of vocabulary replacements, combined with the application of laws of sound change, that produce the most valuable historical conjectures. An impressive amount of work has been done on the reconstruction of ancestral Bantu word roots and the sound system used to realize them. As early as the 1890s, Carl Meinhof postulated an original sound system and started working out the shifts that it had been subjected to in order to produce the various Bantu languages as they are spoken today. Meinhof called his systematic reconstruction Ur-Bantu, and most subsequent historical research (and much actual description of contemporary languages) has been based on his surprisingly trustworthy descriptions. *Comparative Bantu* (4 vol., 1967–70), by Macolm Guthrie of the University of London, is a much fuller and more painstakingly detailed compendium of the entire geographical distribution of the Bantu languages.

Although textbooks and grammatical lessons have been published for many Niger-Congo languages, those of the Bantu area are by far the most numerous. In part, this availability reflects the influence of the South African linguist Clement Doke and the department of Bantu studies at the University of the Witwatersrand in South Africa, which was in direct touch with a huge area of Bantu-speaking peoples.

ADOPTION OF FOREIGN WORDS AND SOUNDS

All the subgroups of Niger-Congo contain languages that show evidence of intensive contact with each other or with languages outside the group to such an extent that special mechanisms are in operation for the inclusion of borrowed words (such as the assignment of nouns to a particular noun class). This does not make them especially unusual from the standpoint of the world's languages, but the borrowing of the very difficult click consonants from the neighbouring Bushman–Hottentot languages into the Bantu languages of South Africa is definitely remarkable. Clicks are especially widespread in Zulu and Xhosa—the first consonant of even the name Xhosa itself stands for a click. This dramatic exception to the theory of language change that states that languages alter in order to simplify or economize the effort of sound production seems to require a special explanation. Possibly in-marrying Bushman or Hottentot wives used their native click sounds as disguised forms of words that would be taboo in Zulu or Xhosa. (D.W.C.)

Borrowing of click sounds

III. Chari-Nile and Nilo-Saharan languages

The term Chari-Nile refers to a group of languages presumed to be genetically interrelated, or descended from a common ancestral language. Joseph Greenberg proposed the Chari-Nile family as part of his overall classification of African languages. At first, he designated the group by the name Macro-Sudanic (1955), but in 1963 adopted the term Chari-Nile. The name Chari-Nile is geographically descriptive in that virtually all of the languages in the family are located in the watersheds of the Nile and the Chari rivers or in the areas in between them.

CLASSIFICATION

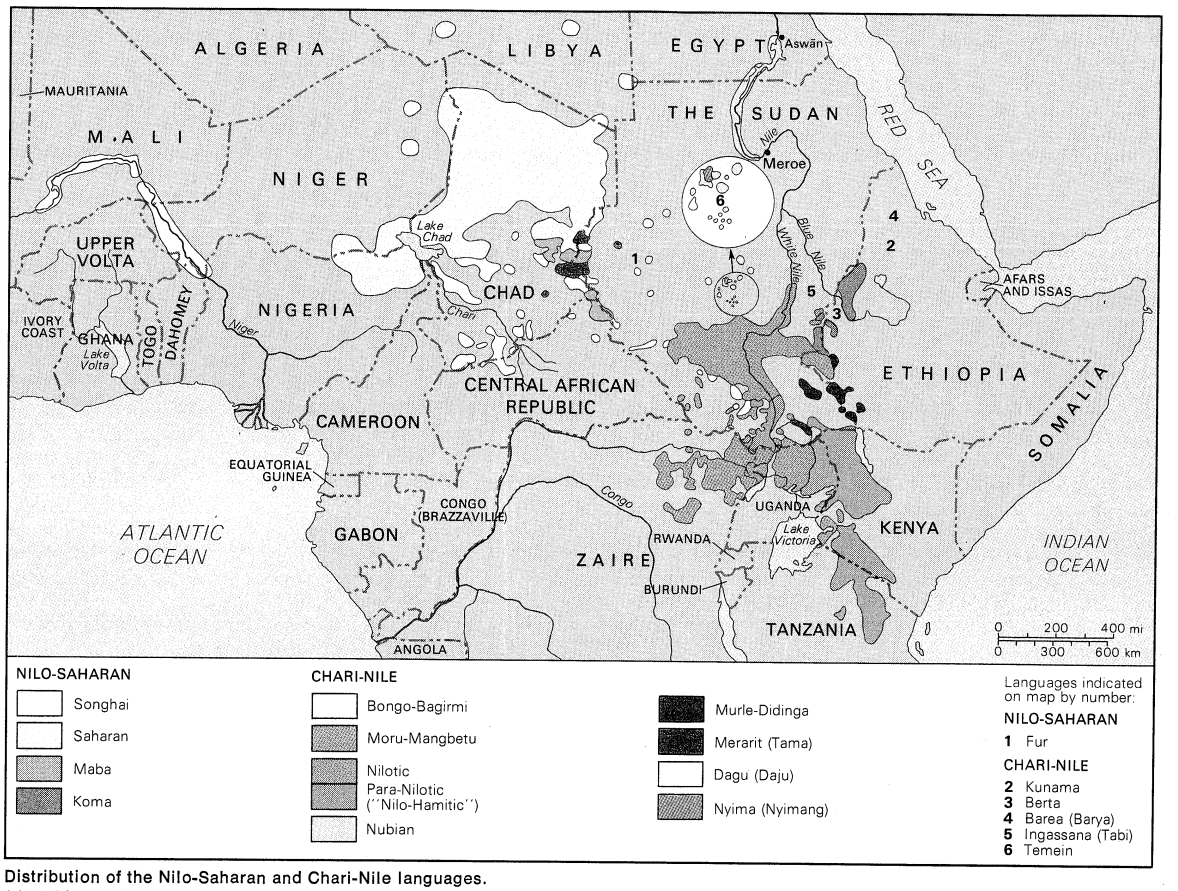
The larger Nilo-Saharan group. Greenberg, as of 1963, regarded Chari-Nile as only a branch of a larger family, Nilo-Saharan, of which it is by far the most heterogeneous. The other branches of Greenberg's Nilo-Saharan family are: (1) Songhai, an important language with no close relatives, which consists of a number of dialectal varieties (*e.g.*, Zerma) spoken along the Niger River in Mali and Niger; (2) Saharan, a language group including Kanuri, the major language of northeastern Nigeria, as well as its relative Teda and its more distant relative Zaghawa, both of which are spoken to the east of Kanuri, principally in Chad, but also in The Sudan; (3) Maba, a group of interrelated languages of Chad; (4) Koma (also written Coman), a group of interrelated languages of the Ethiopia–Sudan border area; and (5) Fur, a language of Darfur province, The Sudan.

Languages of the Chari-Nile group. Greenberg considered Chari-Nile to be composed of four branches. The first, Kunama, is a language of northern Ethiopia (Eritrea), contiguous to Barea (Barya), an Eastern Sudanic language (see below); Berta, the second branch, is spoken in the Ethiopia–Sudan border area. Central Sudanic, a large and heterogeneous group of languages, extends over northwestern Uganda, the southern Sudan, northeastern Zaire, Chad, and the Central African Republic, although it is interspersed with languages of other groups. The British Africanists Archibald N. Tucker and Margaret A. Bryan (1956) regarded Central Sudanic as two distinct unities, which they hesitated to group together; one group is Bongo-Bagirmi, of which Bagirmi, Sara, and a number of other languages are found in Chad and the Central African Republic, while Bongo and its closest relatives are concentrated in the extreme south of The Sudan near the Zaire and Uganda boundaries. The other division is Moru-Mangbetu (like Bongo and Bagirmi, these are the names of individual languages), which is concentrated in northeastern Zaire but overlaps into Uganda. In addition, it includes Efe, Lendu, and Madi, which is linked to Moru. The fourth and most heterogeneous branch of Chari-Nile is Eastern Sudanic.

The 10 subgroups of Eastern Sudanic

Eastern Sudanic consists of 10 subgroups: (1) Nilotic extends from southern Sudan through Uganda, Kenya, and northern Tanzania, although it is interspersed with languages of other groups, and includes the so-called "Nilo-Hamitic" languages. (This name is controversial; see below *Nilotic and "Nilo-Hamitic."*) The Nilotic subgroup includes Shilluk, Dinka, and Nuer in the southern Sudan, Acholi in northern Uganda, and Luo, in western Kenya along Lake Victoria but overlapping into Tanzania. The Nilo-Hamitic languages include Bari and Lotuko (Lotuho) in the extreme south of The Sudan; Karamojong in northern Uganda (just to the west of Acholi) and its neighbour and close relative Turkana across the border in Kenya; Nandi and Suk, just to the south of Karamojong; and Masai, extending from southern Kenya into northern Tanzania.

(2) Nubian is spoken along the Nile from Aswān, Egypt to Meroe, The Sudan (*i.e.*, Nile Nubian), and a number of dialect enclaves in the Nuba Hills of Kurdufān, The Sudan (*i.e.*, Hill Nubian), as well as in two separate enclaves, Midobi and Birked, in Darfur, The Sudan. The



Distribution of the Nilo-Saharan and Chari-Nile languages.
Adapted from A.N. Tucker and M.A. Bryan, *Linguistic Analyses: The Non-Bantu Languages of North-Eastern Africa* (1966); Oxford University Press (for International African Institute), London

relationship of Midobi and Birked to one another and to the other varieties is not close. Currently, Birked is giving way to Arabic. Nile Nubian consists of three languages, of which the northernmost, Kenuzi (Kenzi), and the southernmost, Dongolese, are closely related, almost to the point of mutual intelligibility; Mahas, located between the two, is a more distant relative and includes a subdialect, Fadidja, spoken to the north.

(3) Murle-Didinga designates a group of closely interrelated languages spoken on both sides of the Sudan–Ethiopia boundary; Murle and Didinga are individual languages within the group. (4) Barea (or Barya), a language of northern Ethiopia, is spoken just to the north of Kunama. (5) Merarit (also written Mararit), spoken on both sides of the Chad–Sudan border, comprises a group of closely interrelated languages; actually, Merarit is only one language of the division, which also includes Tama, for which the group is sometimes named. (6) Ingassana, also known as Tabi, is found in the eastern Sudan just north of Berta. (7) Dagu, or Daju, is the name of a group of closely related languages spoken in widely separated enclaves in Chad and The Sudan; most of them are called Dagu or some variant of the name.

Recent additions to the Eastern Sudanic branch

Three subgroups of Eastern Sudanic were added by Greenberg in 1963; the first two had previously (1955) been treated as entirely independent language families, while the third had not been mentioned. These are: (8) Nyangiya, also known as Teuso, a group of dialects or closely related languages of northern Uganda, whose inclusion in Eastern Sudanic Greenberg regards as only tentative (this is omitted on the map); (9) Temein, a language of Kurdufān, The Sudan; and (10) Nyima, or Nyimang, also found in Kurdufān. Another language, Afitti, spoken about 100 miles (160 kilometres) to the north, evidently should be grouped with Nyima.

Controversies concerning classification. There has been much disagreement about the classification of various groups of African languages and considerable confusion in terminology.

Nilotic and "Nilo-Hamitic." Some scholars have objected to grouping the so-called "Nilo-Hamitic" languages together with Nilotic. This question has been at issue since the earliest attempts at African linguistic classification. One of the first proposed systematic classifications (1877–78) established the very heterogeneous Nuba-Fula group, named for two of its members, Nubian (an Eastern Sudanic language, according to Greenberg) and Fula, or Fulani (a West African language). The group also included Masai, subsequently classified as "Nilo-Hamitic," but not the other Nilotic or "Nilo-Hamitic" languages. This classification, made in the 19th century by Friedrich Müller, was based mainly on a simplistic correlation with racial categories determined largely by hair type. Müller recognized a close affiliation between Bari (subsequently classed as "Nilo-Hamitic") and Dinka, a Nilotic language. Dinka, Shilluk, Nuer, and Bari were grouped under the heading Nile languages (*Nil-Sprachen*) within the Negro group, as distinct from the Nuba-Fula group. (Later Müller admitted the close affinities between Masai and Bari, previously pointed out by Karl Richard Lepsius, but attributed them to a mixture of two distinct elements in Bari.)

Objections to grouping Nilotic with "Nilo-Hamitic

In 1880, Lepsius proposed another of the early classifications of African tongues. As a result of his views that Masai and Bari were related, the genetic affiliation of Masai, Bari, Nuer, Dinka, and Shilluk was generally accepted, and languages showing clear affinities to them were indicated as they became known (*e.g.*, Lotuko, Nandi-Suk, Luo).

In 1902 Sir Harry Johnston divided these languages into two groups, one of which he called Nilotic, including Dinka, Shilluk, Acholi, Alur, Lango, and Luo. He proposed no specific name for the second group, which later came to be known as "Nilo-Hamitic," but included in it Masai, Turkana, Lataka (*i.e.*, Lotuko), Nandi, Suk, and Karamojong. He posited a distant relationship between the two groups but also noted some striking lexical similarities between Somali (a Hamitic language) and the sec-

ond, or "Nilo-Hamitic," group that led him to conclude that the latter resulted from a mixture of one group of people related racially and linguistically to the Somali and another related to the Nilotes.

Most investigators of this period grouped these languages in much the same way as Johnston; for example, the English diplomat and scholar Sir Charles Eliot stated:

> East Equatorial Africa is the home of a group of tribes which cannot be called Bantu or Hamitic, and have received no satisfactory designation. The group contains at least two subgroups. One of these, including the Shilluk, Dinka, Bari, Acholi, and Jaluo [*i.e.*, Luo], inhabits the banks of the Nile and the shores of Lake Victoria; the other has its headquarters in the highlands of . . . East Africa. . . . Its chief members are the Masai, Nandi, Turkana, and Suk. . . . To these should probably be added Latuka [*i.e.*, Lotoho] but our information about it is slight. . . . Karamojo is not well known but is apparently closely allied to Turkana.

Correspondence of Sir Charles Eliot's classification with Nilotic and "Nilo-Hamitic"

Except for Bari, which was put in the first group for reasons more geographic than linguistic, but which is actually closely related to Masai, Eliot's two divisions correspond exactly to Johnson's—namely, Nilotic and what later became known as "Nilo-Hamitic."

A sharply different view was introduced in 1912 by Carl Meinhof, who denied the linguistic unity of the two groups suggested by Johnston and Eliot. Meinhof regarded the term Nile (or Nilotic) merely as a geographic concept. He clearly considered Bari and Masai to be Hamitic languages (*i.e.*, related to Somali and Galla, among others) and also, presumably, Lotuko, Turkana, Nandi, and Suk. (He was not, however, explicit about these.) Dinka and its close relatives—Shilluk, Nuer, Luo, and others—were classified as Sudanic; the resemblances between the two groups (in particular those between Dinka and Bari, which had been pointed out by Müller) he attributed to borrowing. Meinhof's most important reason for allying Masai and its close relatives to the Hamitic languages was that in both groups nouns have grammatical gender and other inflectional complexities, which are absent in Nilotic languages such as Shilluk, Luo, and their close relatives.

Much earlier, Lepsius also had noted the presence of grammatical gender in Masai, a trait to which he attributed great importance for genetic classification, but he specifically ruled out any Hamitic affiliation because gender is expressed very differently in the Hamitic languages. Because of Meinhof's great prestige and preeminence among Africanist linguists, his views were widely accepted. The scholar Dietrich Westermann (in 1912), however, regarded all of the above as Nilotic languages, dividing them into the Niloto-Hamitic (Nilo-Hamitic) and Niloto-Sudanic subdivisions, which correspond to the groups delineated by Johnston. On the other hand, contrary to general practice, he also included among the Niloto-Sudanic languages another group, Moru-Madi, of western Uganda and eastern Congo. Greenberg subsequently included Moru-Madi within the Central Sudanic family, which he considered a branch of Chari-Nile and, thus, ultimately but distantly related to Nilotic.

In 1935, Westermann claimed that the relation between Niloto-Sudanic and the Sudanic languages of West Africa was not as close as the relation of the latter to Bantu, whose interrelationship he had sought to demonstrate in 1927. Greenberg was to designate this family, Bantu included, as Niger-Congo. Westermann likewise expressed doubt as to the affiliation of "Nilo-Hamitic" to the Hamitic languages and reaffirmed the unity of Nilotic and "Nilo-Hamitic." Greenberg also adopted this position and argued persuasively against the relationship of "Nilo-Hamitic" to the Hamitic languages and, consequently, against the misleading nature of its name, which he changed to Great Lakes and which, together with Nilotic, formed the southern branch of his Eastern Sudanic family.

Doubtful affiliation between "Nilo-Hamitic" and Hamitic

The German Africanist Oswin Köhler proposed, in 1955, the name Nilotic for the entire southern branch and divided it into three subdivisions: West Nilotic (including Nuer, Dinka, Shilluk, Luo), East Nilotic (including Bari, Masai, Lotuko, Karamojong), and South Nilotic (including Nandi, Suk, Tatoga). (The last two had usually been grouped together as Nilo-Hamitic, while the first had simply been designated Nilotic.) In 1963 Greenberg subsequently accepted Köhler's grouping and nomenclature. Tucker and Bryan, while still maintaining the unity of "Nilo-Hamitic " and hesitating to group it unequivocally with Nilotic (in the older restricted sense, equivalent to Köhler's West Nilotic), proposed a new name for the group, Para-Nilotic, thus avoiding any implication of Hamitic affiliation.

The wider relationships of the Nilotic family. The question of the relationship of Nilotic (in Köhler's extended use of the term) to other African languages and language families has been raised at various times. In 1920 the English anthropologist G.W. Murray noted similarities between Bari and Nubian and, in fact, between Nubian and the Nilotic and "Nilo-Hamitic" languages as a whole. He also linked these languages to Kunama, Barea, and the recently discovered Tabi (also known as Ingassana). Westermann suggested in 1935 that Didinga might be a distant branch of Nilotic and had earlier (1912) made a similar claim for the Moru-Madi group. All these languages were subsequently classed by Greenberg either as Eastern Sudanic or within the wider grouping of Chari-Nile.

Apart from Eastern Sudanic, the most heterogeneous subgroup of Chari-Nile is Central Sudanic. Its unity was first suggested in 1940 by Tucker, who included it in a group of languages that he called Eastern Sudanic, a term that he admitted was essentially geographic because it included another group of different and distinct languages, among them Ndogo-Sere and Azande. (Greenberg subsequently incorporated these into Niger-Congo.) Greenberg introduced a terminological confusion here by calling part of Tucker's Eastern Sudanic group Central Sudanic and using the name Eastern Sudanic for an entirely different group of languages. Although Tucker hesitated to unite the Moru-Madi and Bongo-Bagirmi languages—which together comprise Greenberg's Central Sudanic—into a single family, the evidence that he himself presented appeared to support such a grouping.

Some scholars doubt that adequate evidence has been presented for regarding the whole of Eastern Sudanic as a valid subdivision of Chari-Nile or even for regarding the whole of Chari-Nile as a valid subdivision of Nilo-Saharan. Although Nilo-Saharan and the more heterogeneous subdivisions within it are far from conclusively established, no convincing alternatives have been suggested.

Opposition to Greenberg's groupings

LINGUISTIC CHARACTERISTICS

The Chari-Nile languages are a very heterogeneous group whose structural features, both phonological and grammatical, are in no way uniform or easily characterized. Furthermore, the great majority are not adequately described, and their comparative study is still in its infancy.

Phonology. Nearly all Chari-Nile languages seem to be tonal (*i.e.*, words and grammatical forms are differentiated by pitch), as are the overwhelming majority of sub-Saharan languages. Some varieties of Nubian may be exceptions, but the question has yet to be fully investigated. Central Sudanic languages generally have only syllables that are open (*i.e.*, end in vowels); some final vowels in the Bongo-Bagirmi languages of this division, however, are semi-mute or whispered. Other Chari-Nile languages have syllables that are open as well as closed (*i.e.*, end in consonants). Central Sudanic languages likewise differ from other Chari-Nile languages in that they are the only ones with the labiovelar consonants *kp* and *gb* (articulated with both the lips and the soft palate), which are common in much of West and Central, but not East, Africa. As in other African language families, a large number of Chari-Nile languages distinguish dental from alveolar or retroflex stops, among them most Central Sudanic languages, Nyima, Temein, Murle-Didinga, Hill Nubian, and West Nilotic. (The dental stops, *t* and *d*, are pronounced with the tongue tip against the back of the upper teeth; the alveolar and retroflex stops are pro-

nounced with the tongue tip farther back along the roof of the mouth.) Some Chari-Nile languages—most Central Sudanic languages, Dagu, the Murle-Didinga group, and Bari (East Nilotic)—distinguish implosive and explosive voiced stops; the former are pronounced by drawing the air into the mouth, the latter by expelling it. The vowel systems of most Chari-Nile languages are fairly rich, the great majority seeming to distinguish seven or more vowel qualities. Apparent exceptions are Dagu, Nile Nubian, and possibly Kunama, which have only five vowels. Many languages make distinctions in the length of vowels (*e.g.*, Nile Nubian, Nilotic, Dagu, Kunama), but not all (*e.g.*, Central Sudanic, Nyima). In Nilotic languages a contrast exists between so-called breathy versus hard vowels, sometimes analyzed as close versus open. Nasal vowels occur in Sara-Mbai of the Bongo-Bagirmi group within Central Sudanic but in no other Chari-Nile language.

Vowel systems of the Chari-Nile languages

Grammar. No Chari-Nile language has concordial noun classes like those of Bantu (see above *Benue-Congo subgroup*) and many of its more distant relatives in West Africa, but grammatical gender is distinguished in three very different Chari-Nile groups. It is prominent in East Nilotic (*i.e.*, "Nilo-Hamitic" minus the Nandi-Suk-Tatoga group) but absent in the other Nilotic languages. There are two genders of nouns, normally indicated by a determiner (in Masai, these are *l* for masculine, *n* for feminine) with various grammatical functions. On the other hand, gender is not distinguished in the personal pronouns, but in Dagu (grouped by Greenberg in Eastern Sudanic) there are three genders (masculine, feminine, neuter) distinguished in the 3rd person singular pronoun ("he," "she," "it") but not in the noun. The same usage occurs in Bongo but in no other Central Sudanic language.

A few Chari-Nile languages distinguish exclusive forms from inclusive forms of the 1st person plural pronoun. The exclusive pronoun excludes the hearer (*e.g.*, "he and I" = "we"); the inclusive includes him (*e.g.*, "You and I" = "we"). Among such languages are Lendu (alone among Central Sudanic languages), Dagu, most West Nilotic languages, and Teso (alone among East Nilotic languages). Virtually all Chari-Nile languages except Berta and Nyima formally distinguish plural from singular nouns, some by means of prefixes (*e.g.*, the Moru-Mangbetu languages within Central Sudanic, and Temein) but most with suffixes (*e.g.*, the Bongo-Bagirmi languages within Central Sudanic, Kunama, Barya, Dagu, Ingassana, Murle-Didinga, Nyangiya, and East and South Nilotic, which together comprise "Nilo-Hamitic"). West Nilotic utilizes suffixes sometimes, as well as change of tone, vowel length, or vowel quality. These processes can also be found to some extent in other branches of Nilotic. Affixes are sounds or groups of sounds that are added to the beginning (prefixes) or end (suffixes) of a word or inserted in the middle (infixes). In Chari-Nile, it has been noted that two pairs of singular–plural affixes are widespread, though not universal; these are also used to some extent outside this family. The affixes are *t* and *k*, confined principally to nouns, and *n* and *k*, principally used with pronominal elements. The *t* and *k* forms occur in fewer languages than the *n* and *k* ones, and only in languages in which the *n* and *k* are also found. The significance of these grammatical features for the historical and comparative study of the Chari-Nile languages remains to be investigated.

Widespread use of specific affixes in Chari-Nile

Writing. Only one Chari-Nile language, Nubian, has a written tradition of any antiquity. Dating from the 8th to the 11th centuries, the script was employed by Christianized Nubians in much the same area as that occupied by modern Nile Nubian speakers. Generally called Old Nubian, this written language most closely resembles Mahas among the modern dialects. It was abandoned after the Nubians adopted Islām and survives only in ancient manuscripts and inscriptions that were not deciphered until 1906. (Since the adoption of Islām, Nubian has occasionally been written in Arabic script.) Old Nubian utilized a script derived from that of Coptic, which in turn was adapted from the Greek alphabet. Coptic added seven non-Greek letters of Egyptian demotic origin to represent non-Greek sounds. Nubian retained three of these letters, in addition to adding three new ones from cursive forms of Meroitic letters.

Meroitic, the extinct language of an even more ancient Sudanese civilization (*c.* 300 BC to AD 100), may be related to the Chari-Nile languages, but too little is known to establish any conclusive tie. It is not closely allied to any existing language. Surviving only in inscriptions, Meroitic was written in a consonantal alphabet derived from Egyptian hieroglyphic writing and in a cursive form partly of Egyptian demotic and partly of indigenous origin. In the first decade of the 20th century, the script was deciphered by the English Egyptologist Francis Griffith.

Since the colonial period, a number of Chari-Nile languages, particularly in Kenya, Uganda, and The Sudan, have been written in derivatives of the Latin alphabet and are used for religious and educational purposes. Nearly all of the languages are Nilotic. Of these, Luo, a West Nilotic language and one of the major indigenous languages of Kenya, has the largest vernacular literature. It also has the largest number of speakers of any Nilotic tongue (and, in fact, of any Chari-Nile language). Other West Nilotic languages, however, such as Dinka and Nuer of The Sudan and Acholi of Uganda, have some vernacular literature, as does Kalenjin (Nandi-Kipsigi), a South Nilotic language of Kenya. (M.F.Go.)

Languages using Latin-based alphabets

IV. Khoisan languages

The Khoisan languages are click languages spoken in southern Africa. The term Khoisan was created to subsume the related Bushmen and Hottentots (*i.e.*, the Khoisanid peoples) under a common name and has become increasingly accepted since its creation in 1928. The word is a combination of *khoi* and *sa;* the first part is taken from Khoi-Khoin, the name of the Nama Hottentots, and the second is the name tor the Bushmen in Nama.

The languages of the group are now usually divided into three groups—North, Central, and South Khoisan. They are spoken by remnants of a yellow pre-Bantu population, the so-called Bushmen, who are hunters and collectors living in and around the Kalahari and who number between 35,000 and 40,000. A Khoisan language is also used by the so-called Hottentots, pastoralists surviving with about 40,000 Nama speakers in South West Africa, where the Nama language has also been adopted by the Bergdama, a non-Khoisanid people (about 53,000), and by the Hai-//'om Bushmen (about 3,000 speakers). The Khoisan family comprises about a dozen languages and dialect clusters such as Nama, !Kora, Naro, /Kham, !Khung, Kxoe, and others. (The unusual symbols in the names of the languages stand for clicks; they are listed in Table 2 and are explained below in the section on phonology.)

The assumption of a genetic relationship between all the languages of the Khoisanids and the application of the term Khoisan to this language "family" are problem-

Table 2: Notation of Clicks

	characters used in text	International Phonetic Alphabet	Zulu conversion
Influxes			
(ingressive parts of clicks)			
Dental	/	ǀ	c
Alveolar	≠	ʘ	
Alveopalatal (or retroflex)	!	ǃ	q
Lateral	//	ǁ	x
Bilabial	·		
Effluxes			
(egressive part of clicks)*		k	

*A click consists of two types of articulation, influx and efflux, produced simultaneously. In general orthography (names, etc.), the efflux part of the click is usually represented by k. In actual pronunciation, this k is a velar closure that may be accompanied by aspiration, velarization, or glottalization.

Problems of Khoisan genetic relationship

atic. Criteria brought forth against the point of view of a genetic relationship are the cleavage in morphology and vocabulary between sex-gender languages (Central Khoisan) and nongender languages (North Khoisan and South Khoisan), on the one hand, and a similar cleavage in morphology and vocabulary between North Khoisan and South Khoisan, on the other. These arguments are countered by the fact that there are a number of common words and a few particles, traceable by comparison of sound and meaning, in two or more of the three groups of Khoisan. These permit the establishment of a strong hypothesis for a genetic, though remote, relationship. Regular sound correspondence, however, has not been found in all common items; this is the most difficult part of the Khoisan problem.

The Khoisan hypothesis is built on: (1) common special features of the phonological system—*e.g.*, clicks are regarded as inherent to the languages; (2) widespread and common patterns of root formation, combined with special patterns of consonant distribution; (3) the occurrence of some probably related particles in more than one group; and (4) the occurrence of related words (sound-meaning units) in two of the groups, but more rarely in all three groups—*e.g.*, the terms for "chin," "lungs," "throat," "wound."

In addition to Khoisan proper, there are two click languages, Sandawe and Hadza (Hatsa), spoken by peoples in Tanzania; these languages possess a few words, affixes, and particles that justify the assumption of a distant relationship with Khoisan. The link between Central Khoisan and Sandawe is supported by the racial affinities of the Sandawe and Hottentots. The relations between Khoisan and Hadza are more remote. All studies on this subject are still in the initial stages. The U.S. linguist William E. Welmers proposed the tentative extension of the Khoisan family into Macro-Khoisan, including Sandawe and Hadza. In 1962 Joseph Greenberg subsumed Khoisan proper and Sandawe and Hadza under the title Khoisan.

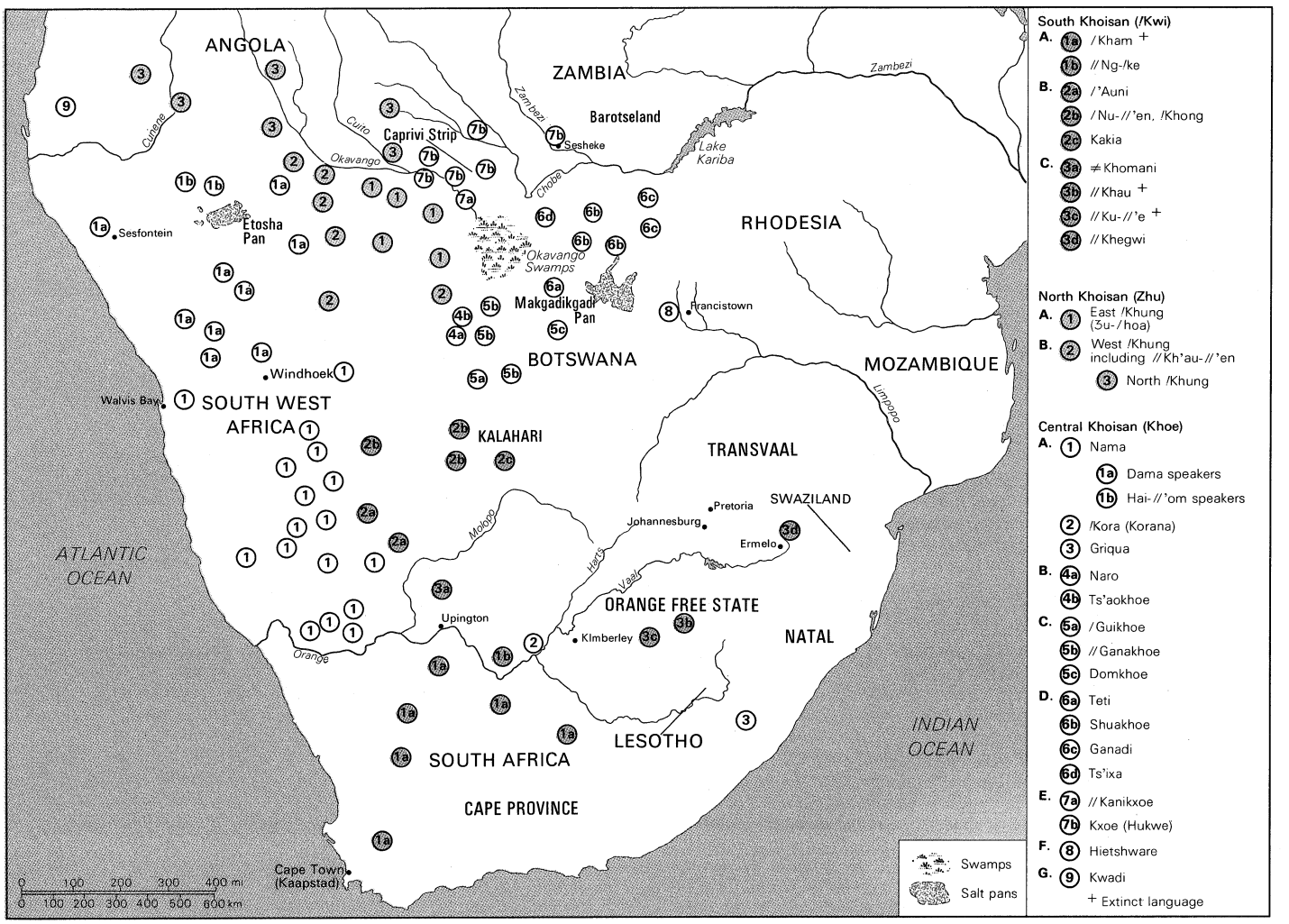
Some Central Khoisan gender affixes were at one time considered as possibly related to the Hamitic languages. Today, the basis for the Hottentot-Hamitic hypothesis is generally regarded as too narrow for conclusiveness and the hypothesis has been given up for lack of evidence.

CLASSIFICATION

All classifications of Khoisan are based on a general comparison of morphology and vocabulary. Lack of sufficient language material limits systematic comparative studies. While the division into groups is fairly consistent, the classification within the groups must be regarded as preliminary. The first classification of these languages, made by the German linguist Wilhelm Bleek in 1858, distinguished between the Hottentot and Bushman languages according to the gender–nongender dichotomy. This remained the traditional model until 1927, when Bushman was subdivided into Northern, Central, and Southern groups. These groups were maintained in D.F. Bleek's *Bushman Dictionary* (1956), in which the Hadza language was included in the Central Group.

Modern theories of classification

Since 1950, several attempts at a reclassification of the languages of the Khoisanids have been proposed. The contribution of Greenberg to the Khoisan hypothesis was to include Hottentot in the Central group and, as mentioned above, to extend Khoisan to include the more closely related Sandawe and the more distantly related



Tentative distribution of the Khoisan languages.

Hadza. Greenberg's main aim was to trace far-spread common elements in Khoisan as evidence for the genetic relationship of the groups. The method is based on sound-meaning comparison. The South African scholar E.O.J. Westphal separated the languages of the Bushmen from the gender languages that he interpreted as of Hottentot origin. Later, he divided the languages of the Khoisanids into five genetic groups, and in 1963 he further revised his classification of the non-Hottentot tongues by establishing four genetically unrelated groups, to which he gave the status of language families (1965); in 1971 he reduced these to three families. A new classification of Central Khoisan was put forth by Oswin Köhler in 1962, which showed the special position of the Nama and !Kora dialects of the Hottentots within the Central Khoisan group (see below).

There is no general agreement on the division of the South Khoisan languages. Some special phonological and morphological features characterize South Khoisan as a coherent and genetically related group of languages whose divergences must go back to a much older separation than that in Central and North Khoisan. South Khoisan speakers were probably the first Khoisanid migrants to reach South Africa.

The following is a classification, proposed by Köhler, of the languages subsumed in the three major Khoisan groups (a comma between languages indicates a close relationship; a semicolon designates a more distant relationship):

South Khoisan: /Kham, //Ng-!ke; /'Auni, Xatia, /Nu-//'en, !Khong, /Namani, Kakia; ≠Khomani, /Nhuki, //Khau, //Ku-//'e, Seroa, //Khegwi (Batwa), !Gãne.

North Khoisan (dialects): East !Khung (ɜu-/hoa), West !Khung (called //Kh'au-//' en or Auen in its southern area), North /Khung (!'O-!Khung in Angola), closely related to West !Khung.

Central Khoisan: Nama, spoken by the Bergdama and by the Hai-//'om; !Kora (Korana); Griqua; Naro, Ts'aokhoe; /Guikhoe, //Ganakhoe, ≠Heβa-khoe, Domkhoe; Teti, Danisa, Hura, Shuakhoe, Ts'ixa; //Kanikxoe, Kxoe (Hukwe, Kwengo, Bumakxoe, including Bugakxoe); Hietshware, /Haitshuari, Mohisa; Kwadi (tentatively classified as Central Khoisan by Köhler in 1968 and Westphal in 1971).

Sandawe has similarities to Central Khoisan and may be regarded as distantly related. Hadza shows some relationships to Khoisan and relatively few connections to Sandawe. Its inclusion in Khoisan is problematic and assumes a very high, abstract level of classification, presupposing that the languages separated a long time ago and that a great number of cognates and common features that normally attest to a common origin were lost over an extended period. This high level of classification is also characteristic of all three Khoisan groups.

LINGUISTIC CHARACTERISTICS

Phonology. Most characteristic of the Khoisan languages are the click sounds, which form a subsystem of the consonants. Clicks are produced by simultaneous closure of the tongue with two areas in the mouth, one at the soft palate (the velum), and the other at such prevelar locations as the teeth, alveolar ridge, or palate. After the closures are made, the body of the tongue is moved down and back, so that the enclosed part of the mouth cavity becomes larger. This causes a lowering of the air pressure in the cavity. The click noise occurs when the tip of the tongue is lowered, breaking the partial vacuum and allowing air to rush into the cavity.

Click sounds as a subsystem of the consonants

There are two types of release of clicks. In one the bilabial, dental, and lateral closures are released by an imploding affricate (a sound beginning as a stop and ending with more or less distinctly audible friction); the alveolar, alveopalatal (or retroflex) closures are released by a hard implosive sound. Both releases are called ingressive, or influx, releases, meaning that the air rushes into the mouth. The second type is called egressive, or efflux, release; it results from the velar closure. In this case the air rushes out of the mouth. A click plus efflux (*i.e.*, velar release, *k* alone, *k*+*h*, *k*+*x*, *k*+*q*, *k*+', *k*+*x*') are considered as a single sound rather than as a consonant cluster. In the South Khoisan languages there is also a bilabial click with labial closure; this is like the sound of a kiss. The Khoisan click mechanisms also combine with other phonetic features—*e.g.*, voicing (vibration of the vocal cords), nasalization—to produce additional sounds. The clicks occur only in initial position in word roots (also in compound roots), rarely in affixes and particles. Some of the Khoisan click sounds were adopted into Bantu languages, notably the Nguni group that includes Zulu and Xhosa (see Table 2 for click notations).

In the Khoisan languages, there are intricate systems of tones. The main vowels are *a, e, i; o,* and *u;* ɛ as in "bed" and ɔ as in "all" do occur, but not in all languages, and the ə sound, as the *a* in sofa, is relatively rare. Nasalized *ã, ĩ,* and *ũ* are most common. Pharyngealized (or pressed) vowels and vowels with intermittent glottalization occur in North Khoisan and, to a much lesser degree, in South Khoisan but seldom in Central Khoisan. (A pharyngealized vowel is one produced with constriction of the pharynx, the area at the back of the mouth before the esophagus.) A characteristic feature of the consonants is their use in clusters with final velarization (articulation of the back of the tongue with the velum, or soft palate), aspiration (an accompanying puff of breath), or glottalization (accompanying closure of the glottis). This occurs in all three Khoisan groups.

Grammar. Because the grammatical features of the various Khoisan languages are so divergent and lack overall uniformity, they will be discussed in three groups —South Khoisan, North Khoisan, and Central Khoisan.

South Khoisan. The languages of the South Khoisan group are characterized by a great number of different grammatical elements and processes. For example, the plurals are formed not only by the addition of suffixes but also by reduplication (*i.e.*, the repetition of an element to indicate plural, as "bird-bird" for "birds"), by partial change of the stems, and by completely different stems, called suppletive forms. In the formation of tenses and moods, a variety of particles occurs before the verb. Significantly, there is no class or gender distinction in the South Khoisan languages, except for a few traces in the /Kham language. The normal sentence order is subject–predicate–object.

Types of plural formation

North Khoisan. The North Khoisan dialects are characterized by an almost complete absence of affixes (endings, prefixes, and so on). The relations of the words to one another and their function in particular utterances are designated by particles (similar to English prepositions such as "to," "for," or "with") and by word order. The tense of roots functioning as verbs is shown by an optional preceding particle for present and future time, by nothing for the past tense, or by an expression describing time ("now," "tomorrow," etc.) if used in a sentence without context. As in the South Khoisan languages, there are suppletive (different) stems for some plurals; *e.g.*, *!hũ* means "to kill one," but *!'oa* means "to kill many." The North Khoisan dialects, also in common with the South Khoisan tongues, lack a passive construction and have the basic sentence order of subject–predicate–object.

Central Khoisan. In the treatment of the Central Khoisan languages, the Kxoe language of the west Caprivi Strip (South West Africa) will serve as representative. The Central languages have a highly developed morphology. Kxoe adds suffixes that indicate gender—masculine, feminine, and common in the dual and plural forms, and masculine, feminine, and neuter (*i.e.*, undetermined, with affix *a* or no affix) in the singular—to roots that take noun (nominal) endings. The same gender suffixes are added to the 3rd person pronouns (originally demonstratives). In grammatical gender, masculine implies "long, narrow, strong," and feminine suggests "short, round, broad, weak." For example, */u* or */u.ma* "canoe" is masculine, while */u.hɛ* "ferry" is feminine. Gender distinction is also found in the 1st persons (except singular) and 2nd persons.

Masculine and feminine gender

There are six tenses for the verb, with the present tense also expressing whether the subject is standing, sitting, or lying. Unlike the other Khoisan groups, the Central Khoisan languages have a passive form, which is frequently used. In sentence order, subject–object–predicate is the most usual, but subject–predicate–object is also common, and object–subject–predicate occurs if the subject is a pronoun or a nominal.

Vocabulary. Khoisan vocabulary is highly adapted to the needs of life in a poor natural environment, especially to the hunt, to the gathering of food, and to all skills serving the preservation of life. As a result of contact with other cultures, loanwords have been adopted—*e.g.*, from European languages into Nama, from Bantu into Kxoe. Native Khoisan creations that reflect non-native innovations also occur; *e.g.*, Kxoe */am-mũ-'o-xo* "clock, watch" is literally "sun-see-on-thing." The use of words in a figurative sense to label new objects is another process for word-formation. Examples include Kxoe *yɛ-≠'am* "bridge," which is literally "hole upper side," and *kuru* "to drive [a car]," which derives from "to press the bellows." Gender symbolism also plays a role in the formation of words for objects in the contact cultures; *e.g.*, the Nama masculine term */nū-b* means "leg," while the feminine counterpart */nū-s* means "wheel."

Basic numbers exist for 1 and 2 and, in some languages, for 3; higher numbers, however, are circumscribed. For example, the number 4 in Kxoe is literally "lick hand bone quantity"—that is, "the finger with which one licks out a pot." The numbers in the languages of the Bushmen seldom go beyond 10; this contrasts greatly with the practice in Nama, and in Hottentot in general, in which there is a decimal system with noncompound numerals from 1 to 10.

In general, the unit of the vocabulary equivalent to the word in European languages is a root whose category as a noun or verb is determined by the context and by the use of noun and verb affixes or particles. Compound words are often used to extend meaning, as English derived words do (*e.g.*, in Kxoe, *kx'â-xò*, literally "drink [verb]-thing" means "beverage").

Writing and texts. About 10 Khoisan languages have been recorded with various degrees of intensity by missionaries and linguists. The best known are Nama, Kxoe, !Khung, and /Kham, followed, to a lesser degree, by !Kora, ≠Khomani, //Khegwi, Naro(n), and /'Auni. The old individual methods for recording clicks were superseded by the Lepsius system (1854), which is, with some modifications, still in use. Studies in Kxoe, !Khung, !Khong, /Guikhoe, and Sandawe have been carried out in recent times, with current research concentrating on !Khung. In view of the difficulties of research among vanishing bands of peoples, the prospects of gathering sufficient material in the remnant Khoisan languages are not good.

Folklore collections

The largest collection of folkloristic texts was taken down in the now extinct /Kham language of the Cape of Good Hope. Next to /Kham, the most comprehensive collection of texts on history, folklore, and traditional and modern life was recorded in the Kxoe tongue. Texts on folklore and history also exist in Nama and Korana and to some extent in !Khung. Literacy is found only in the Nama language, among the Nama and Bergdama. In !Khung, the first two primers were printed in 1969.

(O.R.A.K.)

BIBLIOGRAPHY

General works: ISTVAN FODOR, *The Problems in the Classification of the African Languages: Methodological and Theoretical Conclusions Concerning the Classification System of Joseph H. Greenberg* (1959); JOSEPH H. GREENBERG, *Studies in African Linguistic Classification* (1955), Greenberg's first classification of African languages, includes articles that originally appeared in the *Southwestern Journal of Anthropology*; "Africa As a Linguistic Area," in W.R. BASCOM and M.J. HERSKOVITS (eds.), *Continuity and Change in African Cultures*, pp. 15–27 (1959); *The Languages of Africa* (1963), a revision of Greenberg (1955), representing the most up-to-date version of the author's classification; SIR HARRY H. JOHNSTON, *A Comparative Study of the Bantu and Semi-Bantu Languages*, 2 vol. (1919–22), a pioneering study showing many significant similarities between Bantu and many West African languages, both in structure and vocabulary; RICHARD LEPSIUS, *Nubische Grammatik, mit einer Einleitung über die Völker und Sprachen Afrika's* (1880), a grammar of Nubian with an extensive introduction containing one of the earliest overall classification of African languages; CARL MEINHOF, *Die Sprachen der Hamiten* (1912), discusses a selection of languages considered by the author to be Hamitic, including Masai, and the linguistic characteristics that link them; FRIEDRICH MUELLER, *Grundriss der Sprachwissenschaft*, 4 vol. (1876–88), contains one of the earliest attempts at an overall classification of African languages; GEORGE PETER MURDOCK, *Africa: Its Peoples and Their Culture History* (1959); THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics*, vol. 7, *Linguistics in Sub-Saharan Africa* (1971); ARCHIBALD N. TUCKER and MARGARET A. BRYAN, *The Non-Bantu Languages of North-Eastern Africa* (1956), the most detailed survey of all the Chari-Nile languages and language groups, plus many others; *Linguistic Analyses: The Non-Bantu Languages of North-Eastern Africa* (1966), linguistic sketches of the languages included in Tucker and Bryan (1956); DIEDRICH WESTERMANN, *Die Sudansprachen, eine sprachvergleichende Studie* (1911), an attempt to demonstrate the interrelationship of five West African languages (Twi, Ga, Ewe, Yoruba, and Efik) and three from East Africa (Kunama, Nubian, and Dinka), thereby seeking to show the unity of all the Sudanic languages; *Die westlichen Sudansprachen und ihre Beziehungen zum Bantu* (1927), weakens the case for the interrelationship of all the Sudanic languages by showing a closer interrelationship between the Western Sudanic and Bantu languages; "Charakter und Einteilung der Sudansprachen," *Africa*, 8:129–148 (1935), revises the author's earlier views on the Sudanic languages; and with MARGARET A. BRYAN, *The Languages of West Africa* (1952), the most detailed survey of the languages of West Africa.

Niger-Congo languages: Current Trends in Linguistics, vol. 7, *Linguistics in Sub-Saharan Africa* (*op. cit.*), is the most authoritative source on the Niger-Congo languages. This volume includes, among others, articles on West Atlantic, Mande, Voltaic, Kwa, Adamawa-Eastern, Benue-Congo (exclusive of Bantu), and Bantu. Significant light is thrown on the history and current research in these languages, and the "Check-List of African Language and Dialect Names" by WILLIAM E. WELMERS is an indispensable guide to the problem of variant names for languages. A summary of selected facts about the languages may be found in the series called *Handbook of African Languages*, sponsored by the International African Institute, especially the two volumes on Bantu by C.M. DOKE (1945 and 1954), the two on the same subject by MALCOLM GUTHRIE (1948 and 1953), and the one by DIEDRICH WESTERMANN and MARGARET A. BRYAN on the *Languages of West Africa* (*op. cit.*). The most important comparative study of the Bantu languages is presented in MALCOLM GUTHRIE, *Comparative Bantu: An Introduction to the Comparative Linguistics and Prehistory of the Bantu Languages*, 4 vol. (1967–70). The following periodicals are devoted entirely to articles on African languages, of which the overwhelming majority are concerned with Niger-Congo: *Bantu Studies*, *Journal of African Languages*, *Journal of West African Languages*, *Afrika und Übersee*, *Studies in African Linguistics*, *African Language Studies*, and the *African Language Review*.

Chari-Nile languages: MERVYN W.H. BEECH, *The Suk: Their Language and Folklore*, with introduction by SIR CHARLES ELIOT (1911); MARGARET A. BRYAN, "The T/K Languages: A New Substratum," *Africa*, 29:1–21 (1959); "The *N/*K Languages of Africa," *J. Afr. Lang.*, 7:169–217 (1968); MORRIS GOODMAN, "Some Questions on the Classification of African Languages," *Int. J. Am. Linguistics*, 36:117–122 (1970); OSWIN KOEHLER, "Geschichte der Erforschung der Nilotischen Sprachen," *Afrika und Übersee*, vol. 28 (1955), a history of the investigation of the Nilotic languages, including "Nilo-Hamitic," and suggesting a new subgrouping; G.W. MURRAY, "The Nilotic Languages: A Comparative Survey," *Jl. R. Anthrop. Inst.*, 50:327–368 (1920), an attempt to prove the common ancestry of Nubian, Bari, Masai, and Shilluk; P.L. SHINNIE, *Meroe: A Civilization of the Sudan* (1967), includes a discussion of the Meroitic language; BRUCE G. TRIGGER, "Meroitic and Eastern Sudanic: A Linguistic Relationship?" *Kush*, 12:188–194 (1964); ARCHIBALD N. TUCKER, *The Eastern Sudanic Languages* (1940), compares, among others, the languages renamed Central Sudanic by Greenberg, and concentrates on the Moru-Madi group and Lendu; DIEDRICH WESTERMANN, *The Shilluk People: Their Language and Folklore* (1912), discusses the classification of the Nilotic languages.

Khoisan languages: (Classification): E.O.J. WESTPHAL, "A Re-classification of Southern African Non-Bantu Languages,"

J. Afr. Lang., 1:1–8 (1962): "The Click Languages of Southern and Eastern Africa," *Current Trends in Linguistics*, 7:367–420 (1971); and "The Linguistic Prehistory of Southern Africa: Bush, Kwadi, Hottentot, and Bantu Linguistic Relationships," *Africa*, 33:237–265 (1963), with maps; L.W. LANHAM and D.P. HALLOWES, "Linguistic Relationships and Contacts Expressed in the Vocabulary of Eastern Bushman," *Afr. Stud.*, 15:45–48 (1956); for a comprehensive list of classified languages, see D.F. BLEEK, *A Bushman Dictionary* (1956). (*Distribution*): OSWIN KOEHLER "Die Khoe-sprachigen Buschmänner der Kalahari: Ihre Verbreitung und Gliederung," *Kölner Geogr. Arbeiten. Festschrift Kurt Kayser*, pp. 373–411 (1971), with map. (*Phonology*): C.M. DOKE, "An Outline of the Phonetics of the Language of the Chũ: Bushmen of North-West Kalahari," *Bantu Stud.*, 2:129–166 (1925); D.M. BEACH, *The Phonetics of the Hottentot Language* (1938). (*Grammar*): OSWIN KOEHLER, "Noun Classes and Grammatical Agreement in !Xũ," *Actes du VIIIe Congrès Intern. de Linguistique Africaine*, pp. 489–522 (1971); J.W. SNYMAN, *An Introduction to the !Xu (!Kung) Language* (1970); D.F. BLEEK, "Bushman Grammar: A Grammatical Sketch of the Language of the /xam-ka-/k'e'," *Zeitschrift für Eingeborenen-Sprachen*, vol. 19–20 (1928–30); L.F. MAINGARD, "The ≠Khomani Dialect of Bushman," in J.D.R. JONES and C.M. DOKE (eds.), *Bushmen of the Southern Kalahari* (1937); F. RUST, *Praktische Namagrammatik* (1965), based on material of H. Vedder and J. Olpp. (*Comparative studies*): OSWIN KOEHLER, "Studien zum Genussystem und Verbalbau der zentralen Khoisan-Sprachen," *Anthropos*, 57:529–546 (1962), deals with gender and verb system of Central Khoisan; D.F. BLEEK, "A Short Survey of Bushman Languages," *Zeitschrift für Eingeborenen-Sprachen*, vol. 30 (1939–40). (*Vocabulary*): D.F. BLEEK, *A Bushman Dictionary* (1956), a compilation of all language material available, except Hottentot; J.G. KROENLEIN, *Wortschatz der Khoi-Khoin-Namaqua-Hottentoten* (1889), on Nama. (*Oral literature*): W.H.I. BLEEK and L.C. LLOYD, *Specimens of Bushman Folklore* (1911), texts in /Xam with translation and annotations. (*General information*): I. SCHAPERA, *The Khoisan Peoples of South Africa* (1930), deals also with distribution and classification. (*Eyasian languages*): O. DEMPWOLFF, *Die Sandawe* (1916), with a chapter on language and vocabulary; D.F. BLEEK, "Traces of Former Bushman Occupation in Tanganyika Territory," *S. Afr. J. Sci.*, 28:423–429 (1931).

(M.F.Go./D.W.C./O.R.A.K.)

African Peoples, Arts of

This article is concerned primarily with the arts of sub-Saharan Africa, which is often called black Africa, from the predominant skin colour of its peoples, some of whom, the Bushmen and Pygmies, are more properly described as yellow; but the prehistoric art of the Sahara is also included as an extension of The Sudan (the belt of open country stretching across Africa from east to west, just south of the Sahara). The arts of settlers (from Europe and Asia) of the colonial period are not included except insofar as they have influenced the arts of black Africa.

Subdivisions of black Africa

The continent can be subdivided in a variety of ways. In this article it is divided, for the sake of convenience and consistency from section to section, into five areas, corresponding primarily to geographical regions: the open grasslands of The Sudan, the woodlands and forests of West Africa, the basin of the Congo River, the East African savannas, and those of South Africa. These divisions do not necessarily correspond to other possible divisions, such as linguistic or religious ones. Nor do they reflect homogeneous artistic production. The western part of The Sudan, for example, is an area rich in sculptural styles and production; in the eastern part, however, sculpture has been little developed, partly because of the less settled way of life; and artistic expression manifests itself mainly in music and oral literature. Again, colonial experiences in French-speaking and English-speaking Africa have given to contemporary literature a unity of theme and language that corresponds not to geographical regions but to the way the European colonial powers divided the continent.

Influences from outside have also subdivided the continent on a basis of religious tradition. Christians in black Africa form subgroups of traditional cultural units, their artistic expression being often totally alien, though sometimes expressing the imported values through traditional modes, as in the music and dance of the Aladura churches of Nigeria or the Oye-Ekiti sculptures, with Christian subjects. The primary concern in this article, however, is with peoples who have their own religious system—often called pagan animist—which their arts reflect. Islām, however, for about a millennium, has been converting black Africans, with the result that in many areas it has become so closely integrated into the culture of the people that it must be regarded as their traditional way of life; the arts of these peoples, too, will therefore be considered.

This article is divided into the following sections:

I. Theoretical considerations
The African conception of art and the artist
Backgrounds of African arts
General characteristics of African arts
African arts in the contemporary world
II. Literature
Oral traditions
Modern literatures
III. African music
Music and society
The structures of African music
African musical instruments
IV. Performing arts
General characteristics
African dance
African drama
V. Visual arts
Types of African visual art
Sudanic cultures
West African cultures
East African cultures
Southern African cultures
Central African cultures

I. Theoretical considerations

THE AFRICAN CONCEPTION OF ART AND THE ARTIST

In attempting to understand the arts of Africa, non-Africans must free themselves from a conceptual framework that separates factors in a way not always relevant for Africa. The West sets both artists and the arts apart from the rest of life: there are special buildings for the performing arts, and artists expect to be able to support themselves entirely from the practice of art. In Africa the artist may be professional in the sense of having served a long apprenticeship to acquire a specialized skill for which he is paid, but it is only those sculptors who work mainly for sale to Europeans or those musicians performing in large cities who can support themselves exclusively from their art. Although artists are sometimes thought by their countrymen to possess a special temperament, as among the Gola, or a special prestige gained by their skill, as among the Baule, which sets them apart, they nevertheless remain integrated members of the community and must still cultivate their farms.

The African artist as a "professional"

Skill in one or more of the performing arts is often a prerequisite to high rank in an African society: a chief or a priest may have to be a skilled dancer, for example, among the Yoruba and Ashanti; and in order to understand messages sent to him by his drummer he must also have mastered that art. A good deal of the respect such chiefs enjoy accrues from their artistic skills.

The Western distinction between artists and craftsmen is not usually relevant in Africa; the sculptor does not consider it demeaning to produce domestic utensils of wood, and in consequence everyday objects tend to be well made and well designed. The West, with its emphasis on specialization, has had to relearn the value of functionalism; yet this is a value that modern man's peasant ancestors, like Africans in traditional society, took for granted.

Field investigations have revealed that in African societies there are generally accepted standards by which an artist's performance or product is evaluated. A sculpture, for example, that does not reach the minimum standard will be rejected at once. Basically, this means that its form must comply to the norm established by the tradition for the class of object to which it belongs; for

instance, a mask required for a cult must possess features that distinguish it from the masks of other cults. The application of critical standards may vary to some extent from one art form to another within the society: a fine old sculpture, because of termite damage, may be replaced by a new one of inferior quality, but a poor dancer can never replace a good one. The reason for this is that the effectiveness of the ritual depends directly on the skill of the dancer, whose creation exists only at that moment in time, whereas the carving used in the ritual is generally seen as no more than a vehicle in which a spiritual force is brought into action during the performance of the ritual. Exceptionally, the Rega (Lega) judge sculptures on the basis of their effectiveness in use.

Integration of African arts

An important factor, which is only beginning to be appreciated by non-Africans, is that Western academic disciplines have separated out (as has been done in this article) artistic phenomena that in Africa are conceived as a unit. Because an African mask can be enjoyed as a work of sculpture when torn from its context and isolated, immobile, in a museum case, Westerners tend to equate "African art" with sculpture. Africans, however, do not consider the mask a discrete entity; it is part of a costume, and it is of no importance until it is danced in, to appropriate music and lyrics. When an Ibo sculptor was asked to carve a mask that could be taken to America as an example of his work, he produced a carving of a human figure wearing the mask, explaining that the full beauty of the masquerade would not be seen in the mask alone.

For an African, in fact, the sculpture is likely to be the least important element; costumes for the masquerade, for example, are considered more important than the mask, which is sometimes dispensed with completely. The *egungun* (ancestor) cult of the Yoruba may employ masks or elaborate sculptured headpieces or, with equal effect, merely a costume that hides the body entirely. In some societies the dancing is also considered more important than the mask: the Igbira, when watching a masquerade, criticize the dancing rather than the sculpture. Consequently, the dancer or costume maker enjoys greater prestige than the carver in these societies. Often the musician is the highest ranking artist, while the respected *griots* of the western Sudan are musician, singer, poet, and historian all at the same time.

The arts in Africa for too long have been observed through Western eyes; the data have been recorded and interpreted ethnocentrically, with unfortunate results. For a long time it was thought, for example, that the African artist was anonymous and that he worked in a style which, in some mysterious way, expressed the collective volition of his community. (Such ideas are still being expressed.) In fact, African artists were anonymous only because investigators had not bothered to ask for their names; the anonymity was only a reflection of Western ignorance. The situation is comparable to that in medieval Europe, in which the alleged anonymity of the artist simply reflects the incompleteness of written records; for with more careful examination of those that do survive, the names of many artists have been revealed. Thanks to field studies, the names of hundreds of African sculptors are now known throughout the rest of the world. The greater the sculptor's skill, the greater his fame among his countrymen, and, in consequence, the more widely spread are his sculptures. A performing musician similarly achieves individual fame, which is closely associated with his repertory. Even the composer of a piece that others perform is often remembered by name, though perhaps as often only the clan to which he belongs or his village is remembered as having originated it.

Freedom of the artist

Any art product in a traditional African society is required to conform to the norms for the class of object; but provided these minimal requirements are met, the artist has considerable freedom to express his own ideas. For example, among the Mende the *bundu* mask of the Sande (women's initiation society) has a closely prescribed shape (Figure 1), but the artist is free to represent the hair as elaborately as he wishes. Similarly, the masks of the Gelede society of the Yoruba have a prescribed



Figure 1: Mende *bundu* mask of the Sande (women's initiation society), wood, Sierra Leone. In the Manchester Museum. Height 38.1 cm.
Frank Willett

form for the cap mask itself, but the carver is encouraged to be inventive in the superstructure; for the purpose of the dance is to please the witches—that is, potentially all women in the community—so the more unusual his idea, the more effective the ceremony will be. Thus, the set of masks carved for a 1969 series of ceremonies at Meko represented a kerosene pressure lamp; a couple riding a motorcycle; two hunters with a boar; and a carpenter sawing wood in his workshop.

BACKGROUNDS OF AFRICAN ARTS

For an understanding of the African arts, a knowledge of their context in the society in which they originate is required. The carver's use of masses and of open space may produce an aesthetically successful sense of weight or of lightness, but the degree of the artist's technical success can be fully appreciated only if one knows the function the piece was intended to have in the society. Similarly, the rhythm for which African music is praised in the West often has subtleties that escape the untutored ear; and the combination of rhythm and tune played by so-called "talking drums" is subtly related to the tones and rhythms of the drummer's spoken language, which has to be understood before the drumming itself can be appreciated or analyzed.

The arts and the structure of society. All African arts reflect the structure of the society in some degree. This is especially clear in societies that have a hierarchical organization. For example, in Ashanti and among the Yoruba, there are special dances reserved for the king, others specifically for chiefs, and others for priests. Again, sculptured objects—for example, ceremonial staffs and swords—are the prerogative of specific ranks in the society. Among the Yoruba, bead-embroidered objects have a very restricted distribution, confined mainly to the king and his family; though beadwork emblems may be allowed to his chiefs, as is the practice at Ife. In Benin, castings in brass and carving in ivory (Figure 2) were restricted to the king and the palace altars, though permission could be given for small (and, very exceptionally, large) castings to be made for lower ranking members of the community. No such concession ever appears to have been made in the use of ivory. The king's palace is usually decorated with motifs that echo his position in society; the palace of the kings of Dahomey, for example, is

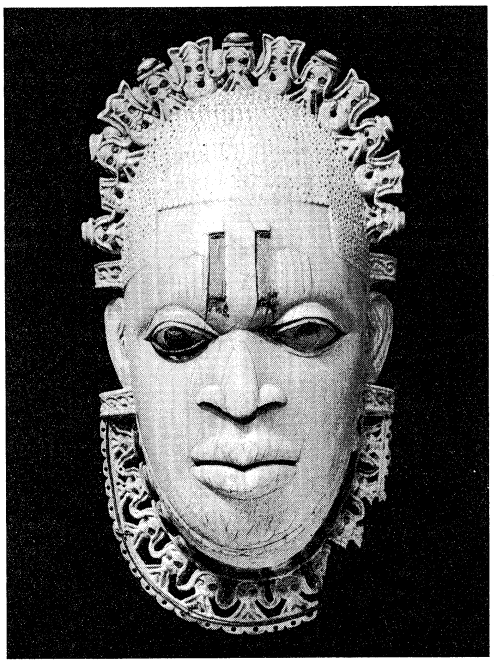


Figure 2: Benin ivory regalia mask, Nigeria. In the Museum of Primitive Art, New York. Height 23.8 cm.
By courtesy of the Museum of Primitive Art, New York; photograph, Charles Uht

decorated with painted low-relief sculptures of royal emblems and events of their reigns. Even in societies whose social stratification is simply one of age groups, there are dances belonging to each individual group. In some societies, the playing of one-note parts in musical consort may express the greater importance attached to the group than to the individual. Occupations, too, are reflected in the arts. In some societies each occupational group has its own dances, as well as its own festival; and some have special ceremonial emblems. The head of the smiths in a Yoruba town has a curved iron rake with a decorative brass handle that is both an emblem of his position and a piece of ritual equipment.

The arts and religion. One of the most widespread religious ideas regularly finds expression in the arts; namely, the cult of the ancestors. The community is commonly seen as composed of the unborn, the living, and the dead. Life is conceived of as cyclical: each individual goes through these three stages, and it is the ancestors who send children. The Fang express this continuity in the figures that guard the bones of their ancestors (Figure 3; right), for these figures combine the proportions of the newly born—large head, long body, and short limbs—with the marks of maturity, adult sexual organs. Because the ancestors are concerned in the affairs of their descendants, they are made manifest, often by means of masks, in the *rites de passage*, during which an individual member of the community changes his position in the cycle of existence. Thus, the Dogon of the Niger Bend employ masks when the young men are initiated into full membership in the community, at funerals, and, later, to send the spirits of the deceased to join the other shades. Among the Dogon, too, everyday objects such as stools and doors (Figure 3; left) are decorated with the figures of ancestors to invoke their protection. The entire Dogon people reassert their unity every 60 years by making a Great Mask for each separate community in turn.

The role of masks

Art, environment, and way of life. The arts also reflect both the environment and the way of life. Wood sculpture is naturally most highly developed where suitable wood can be easily obtained, but a settled agricultural existence is also necessary. Because nomadic hunters such as the Pygmies cannot carry sculptures around with them, they express themselves instead in the arts of music, dance, and oral literature. In open country lacking timber, both hunters and pastoralists may express themselves visually, through painting (for example, Bushman paintings and those in the Tassili), as well as in the non-

By courtesy of (left) the Museum of Primitive Art, New York, (right) Pierre Arman; photograph, (left) Charles Uht

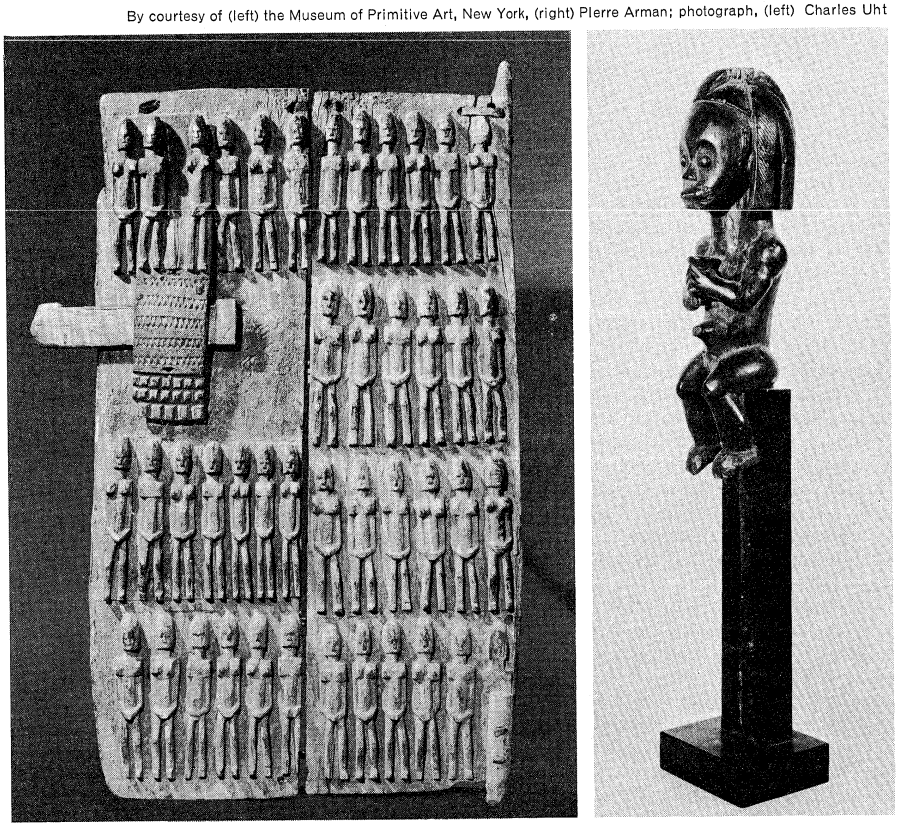


Figure 3: *Ancestor cult objects.*
(Left) Dogon wood granary door decorated with figures of ancestors, Mali. In the Museum of Primitive Art, New York. Height 92.7 cm. (Right) Fang wood figure (*bieri*) from a reliquary of ancestor's bones, Gabon. In the Pierre Arman Collection, Paris. Height 47 cm.

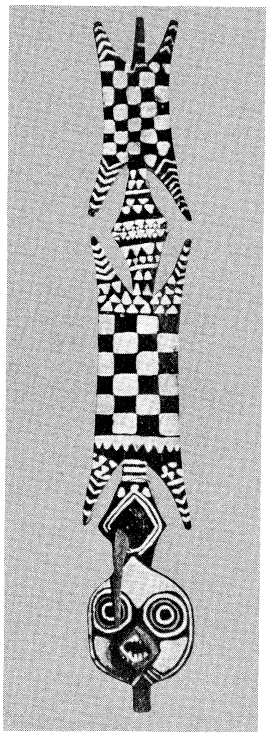


Figure 4: Bobo-Ule dance mask, painted wood, Upper Volta. In the Museum of Primitive Art, New York. Height 1.83 m.
By courtesy of the Museum of Primitive Art, New York; photograph, Lisa Little

visual arts. Architecture, too, shows adaptations to the environment—for example, in providing insulation in hot climates; to social structure, in providing separate huts within the family compound for each wife and her children; and to the type of existence, as in settled communities, where special places are often reserved for musicians and dancers. Dances often enact ritually the means of food production, as in dances of planting and harvest; and the music that accompanies planting among the Dogon accelerates at the end, as if to symbolize growth.

The environment is important to the historian of the arts, for in most parts of black Africa it is exceptional for wooden objects—sculpture and musical instruments—to be preserved. The history of sculpture has to be written largely in terms of stone, metal, and pottery; the history of musical instruments through their representation in these media.

Intertribal and intercultural relations. In attempting to break away from the idea that the African arts are "primitive," when in fact they are highly sophisticated, the term tribal arts has been suggested. Unfortunately, this term attempts to define the form of the society in which the arts are produced rather than the characteristics of the arts themselves. Furthermore, a "tribe" is by no means a fixed and defined social unit. At different times and in different circumstances an individual would identify himself with a social grouping of greater or lesser size, usually seen in opposition to another group of comparable size. Thus, in the Cameroon grasslands, groups based on a single large settlement have been regarded as separate tribes, though they share the same culture; and the 10,000,000 or so Yoruba-speakers have often been regarded as a single tribe despite the fact that they live in large, separate settlements, which are veritable city-states. The assumption of the reality and permanence of tribal boundaries has led to serious misunderstandings about African art. One critic, for example, has suggested that "every tribe is, from the point of view of art, a universe to itself. . . . The tribe is an exclusive 'in-group,' which uses art among many other means to express its internal solidarity and self-sufficiency, and conversely, its difference from all others." There are cases where this is in some sense true. The Dogon, for example, use the *sigui* ceremony and the Great Masks carved for it to express their unity; yet the basic form of the mask, with a very simplified human face surmounted by a tall, flat panel, is shared with the neighbouring Bobo (Figure 4) and Mossi. On the one hand, cult groups, together with their sculptures, overlap "tribal" boundaries—for example, the Poro society is found among peoples in the modern countries of Guinea, Sierra Leone, and Liberia; on the other hand, individual sculptors carve masks for the society's rites in a variety of "tribal" styles as different in character as the Dan and the Ngere modes. The attempt to refine the precision with which one can locate the place where sculptures were carved is further complicated by the fact that sculpture has been traded to and used by totally unrelated groups; for example, Ibibio sculptures are found among the Edo, who live 150 miles away, on the other side of the Niger River. Some similarities of form, however, may be due not so much to diffusion among peoples as to sharing a common heritage. It has been suggested, for example, that the wide distribution of the so-called heart-shaped face, found across the Congo Basin from the Lega to the area of the Ogowe River, is an indication of great antiquity and that the form was spread by the dispersing Bantu-speaking peoples. That it has also been found, however, among the Ibo, who do not speak a Bantu language, may perhaps imply diffusion of the form rather than migration of its users. Reconstructing history from present-day distributions is thus a dangerous pursuit, justified only in the total absence of archaeological and historical data.

Difficulty of defining "tribe"

Implicit in the study of African cultures is the assumption that they were without change before the first contacts with Europeans. This assumption seems to arise not from positive but from negative evidence; that is, the fact that documentation of past changes is lacking, whereas recent changes can easily be documented: Western musical instruments, for example, are commonly employed; and an army gas mask is used in a traditional Ibo masquerade. Yet change seems to be an integral part of the human experience: no society is or ever has been totally unchanging, though the pace of change in Africa might be faster now than it used to be. Thus, scholars may hope one day to write a history of African visual arts. Sufficient evidence has already been found to demonstrate change with time in the visual arts of the Tassili, of the Bushmen, and of Ife and Benin. Because of their impermanent nature, it may never be possible to write the history of African music or dance. Yet scholars will not too easily abandon the hope of doing so; evidence of dance in Meso-America between 500 BC and AD 1500, for example, has been deduced from sculptures and painted pottery.

The role of change in African art

Probably the greater the freedom an artistic tradition allows to the artist, the greater the receptivity to outside influences, particularly when the arts have a secular as well as a religious function. Indeed masquerades in modern Ghana and the so-called fairy masquerades in Sierra Leone, both essentially secular, seem to have been taken over from European papier-mâché masks used for Mardi Gras and Christmas celebrations. The changes under European influence, however, were exchanges: as a result of African influence via the Americas, Western popular dance now emphasizes rhythm, whereas, formerly, as in traditional country dances, or the minuet, space was much more important. Similarly, Western sculptors learned not only to simplify natural forms in their creations but also to enclose space within their masses.

GENERAL CHARACTERISTICS OF AFRICAN ARTS

Westerners still have a long way to go in understanding the arts of Africa and the cultures that produced them. While the importance of studying African arts as an integrated whole is now recognized, for example, relatively little has been done in the field to discover values that apply across Western disciplinary boundaries. One critic who has made the attempt points out that soft materials

such as clothing are represented in Yoruba sculpture as hard and polished; and he compares this to the penetrating sound of rattles and hand clapping and the high pitch of bells and gongs that characterize Yoruba music. Similarly, in singing, a woman soloist's voice often stands out by reason of a high-pitched, penetrating metallic timbre. Perhaps these reflect the same value placed on clarity of line in sculpture. The same critic has also drawn attention to the importance of a "coolness" of expression in the faces of Yoruba sculptured human figures, a quality that avoids emotional expression and that is also expected of dancers in certain cults from which violent dance gestures, too, are excluded. This composure is also an ideal in general human behaviour, not only among the Yoruba but also, as another critic has shown, among the Gola of Liberia, where "the ability to be nonchalant at the right moment . . . to reveal no emotion in situations where excitement and sentimentality are acceptable" is much admired. Among the Gola, too, the dancer should not look emotionally involved.

Avoidance of emotional expression

The arts in Africa are rich in overtones of meaning, echoing the world view of the community; but their investigation is fraught with difficulty. The investigator, for example, may inadvertently lead his informant to produce, simply to please him, the answer he seeks. Moreover, the investigator seeks out the person he judges to be the most knowledgeable member of the community and writes a description of the cosmogony or of the symbolism of the society from that person's evidence. Probably very few other members of the community would share in the subtlety of his interpretation; and in many cases the structure of the society, with successive grades of initiation, reflects this uneven distribution of insight. Thus, what might be presented as the cosmogony or symbolism of an African people may not represent knowledge shared by the entire population. Some of the apparent contradictions in information available to scholars may well arise from the fact that different individuals in a society may have different depths of understanding. Moreover, variation is to be expected from place to place. Even a people like the Yoruba, many of whose cults are widespread, have other cults of very localized distribution; generalizations about the symbolism or cosmogony of the Yoruba people, therefore, must be made circumspectly.

A further complication is the common belief that artists are the best sources of critical information about the arts of their society. In fact, African artists usually refuse to judge their own work and are unwilling to assess that of others, whereas critical judgments of the most important art in a society (for example, the dance) may be obtained quite freely from any of the onlookers.

AFRICAN ARTS IN THE CONTEMPORARY WORLD

In the second half of the 20th century, the arts in Africa are changing more rapidly than ever before. As society changes, the occasions when traditional art forms are required become less frequent; but they are replaced by demands of a new kind, to which African artists are responding. Since colonial times, many potential artists have received advanced Western training, first at European institutions and, more recently, at African universities. Attendance at these institutions has removed the artist from his own society, which would have provided a traditional training; but it has exposed him to Western techniques, values, and traditions, thus broadening his horizons so that he has been able to draw on any tradition that interests him. Thus, Ayo Bankole has composed a Christmas oratorio in Yoruba for soloists, chorus, organ, and drum ensemble; comparable eclectic experiments are conducted in the other arts. Some visual artists as well as musicians have rejected the religion of their ancestors and have found inspiration in Christianity instead.

Traditional literature in Africa was oral, and written literature is an entirely new African art form; it has already produced some writings worthy of a place in world literature. Much of the writing reflects the problems of a changing society (*e.g.*, Chinua Achebe's trilogy), both in looking back to the value system of the past and in seeking new values for the future—an involvement in society that reflects the traditional position of the African artist.

The multiple roles of artists in traditional society are continued also into Western forms: the best known African film producer, Sembene Ousmane, is also a novelist; Demas Nwoko not only sculpts and paints but also designs stage settings; and Wole Soyinka, who built his reputation as a playwright and poet, has written one of the best African novels.

Skills in Western forms, however, seem to have encouraged an interest in the corresponding traditional arts, and modern artists often consciously employ traditional forms. For example, Vincent Kofi's bronze sculpture "Awakening Africa" (Figure 5) is a reclining nude figure, a subject and an attitude more familiar in Western art,

Ulli Beier

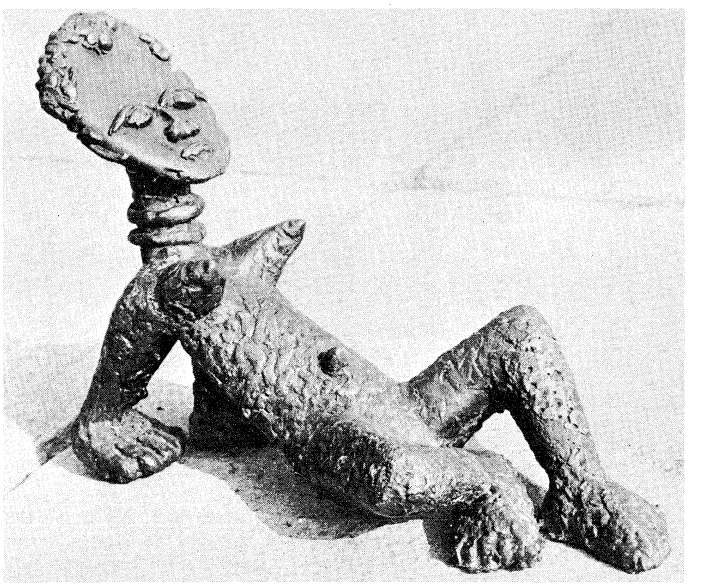


Figure 5: "Awakening Africa," bronze sculpture by the Ghanaian sculptor Vincent Kofi.

with the head and neck drawn from a traditional form of Ashanti sculpture in wood and terra-cotta.

The intermingling of African and Western traditions is currently producing a striking richness in the arts. One aspect of this is the development of what might be regarded as folk art: artists without formal training, either traditional or Western, have been offered the opportunity through workshops in Salisbury (Rhodesia) and Oshogbo (Nigeria) to express themselves in both traditional and Western media. Oshogbo artists, such as the sculptor in cement Adebisi Akanji and the graphic artist Jimoh Buraimoh, have stayed rooted in their community and have found patrons there in a way that Western-trained artists have yet to do. Many visual artists of Oshogbo have drawn inspiration from the writings of Amos Tutuola, who himself can be regarded as a folk artist; for his novels recombine traditional folktales into new stories that are recounted in a form of English closer to that of the Nigerian streets than to the Standard English of the other Nigerian writers. *The Palm-wine Drinkard* has been translated into Yoruba, adapted for the stage by Kola Ogunmola, and is one of the most successful items in the repertory of popular theatre in western Nigeria. Similarly, there has grown up since 1962 the Yoruba Folk Opera, associated particularly with the name of Duro Ladipo, employing traditional music to tell traditional stories in an overall form of presentation that is essentially Western. It was the need to find scene painters for Ladipo's productions, in fact, that led to the establishment of the Oshogbo artists' workshop (known as Mbari Mbayo).

Folk art

There is at present a richness of self-confident creativity in all Nigerian arts that is reminiscent of Elizabethan England. Here clues may be sought to the future of the arts

on the whole continent. Artists with a predominantly Western training have drawn their patronage from outside their own people. Their horizons, wider than those of their ancestors, permit them a wider range of expression, usually entirely personal. The work of painters and sculptors is often African in subject matter but not especially African in style. If they continue to be eclectic, their contribution may well be the valuable one of reducing the emphasis on national and continental characteristics; but they may turn increasingly to their own artistic traditions and eventually develop new African forms of art. For the present, however, it is the folk artists, serving primarily their own communities, who are developing new and specifically African art forms. What is certain is that the present richness of production in all African arts augurs well for the future. (F.Wi.)

II. Literature

The term African literatures is intended to cover both the oral traditions and the mainly 20th-century writings of the Negro, Bantu, and Nilotic peoples living in the part of the African continent south of the Sahara. Though it is not possible to claim a genuine cultural unity for the several thousand peoples living in this area, certain cultural features can be traced throughout black Africa. Especially after World War II, a political consciousness of belonging together, expressed in a Pan-African movement, gained in strength. It is, moreover, convenient to treat the cultures of these peoples separately from that of the Arabs of North Africa, whose traditions, being Islāmic, are basically different, and from that of the European culture of white South Africa. (See ISLĀMIC PEOPLES, ARTS OF; see LITERATURE, WESTERN.)

With the slave trade, oral traditions spread from Africa into the New World, and many African myths, songs, and sayings have survived in the Caribbean and the Americas, where they have helped to give birth to a rich syncretic culture. (Literature by black writers in the Western Hemisphere is discussed in LITERATURE, WESTERN.)

The first modern African writers to be heard significantly in the West, the French-speaking poets of the Negritude movement (see below) of the late 1940s and 1950s, have generally been the most insistent on identifying themselves as African, or black, writers. The majority of English-speaking writers, at a conference in Kampala in 1962, made it clear that they wished to be judged on their merits as writers, not just as "African writers." By the mid-1960s, however, with the development of a larger body of writings by English-speaking Africans and with the growing distance of French-speaking writers from their colonial situation, the apparent differences between French-speaking and English-speaking African writers had begun to seem more verbal than real.

Despite the fact that modern literary activity in Africa exists in relation to vast illiterate populations with their own rich vernacular cultures and is conducted in three major European languages—French, English, and Portuguese—as well as in dozens of printed African languages, certain habits of expression, themes, and approaches to imaginative creation do appear to merit the discussion of this literature as a continental phenomenon. By the 1970s, however, the time was approaching when discussion of "African" literature would need to be accompanied by more precise examination of certain emerging national literatures, such as those of Nigeria, Senegal, and Cameroon. The body of published literature existing in the three European languages is certainly sufficient to be studied in its own right and not merely as a series of exotic extensions of French, Portuguese, or English literature. Furthermore, it is only by the study of African literature comparatively and in its continental context that the true features it has been displaying to the world since its emergence in the mid-20th century can be discovered and understood.

Emergence of national literatures

ORAL TRADITIONS

The poetic and narrative forms of oral tradition among the tribes living south of the Sahara are immensely rich and varied. Most tribes have creation myths and cosmologies; some have pantheons of gods and a body of myths to explain their relationships. The worship of these gods involves praise songs, oracular verse, and incantation. The courts of chiefs usually employ professional bards who are both praise singers to the chief and court historians of the tribe.

During initiation and puberty rites the young men of the tribe, in a period of seclusion in the forest, are made acquainted with tribal history and lore, though this practice is dying out under pressure from modern education and employment. Not all oral traditions are bound up with religion or tribal institutions. There are forms whose only purpose is entertainment, such as proverbs, riddles, animal and trickster tales, and dilemma tales. Even these, however, have a certain didactic function, since they help to establish social norms and to teach sophisticated use of the language.

An important feature of African oral traditions is their close link with music: poetry exists almost exclusively in chanted form or as song, and, among West African peoples with tonal languages (*e.g.*, the Akan and the Yoruba), much poetry is recited on the "talking drum" rather than spoken or sung.

Relation of poetry to music

Myths. African creation stories are perhaps the most varied and imaginative in the world. The Kono of Guinea believe that the original force in the world was Death, who existed before God; the Lozi of Zambia see God as retreating helplessly from the cruelty of man; the creator goddess of the Ijaw peoples of the Niger Delta allows man to choose his own fate before he is born. The Pangwa of Tanzania have a fantastic vision of the world as having been created from the excrement of ants. The Yoruba of Nigeria tell of a creator who got drunk on palm wine and so created cripples and albinos. The most detailed cosmology known, requiring seven days for its recitation, is that of the Dogon, a people living in what is now the West African republic of Mali. An unusually attractive creation myth is that of the Fulani of Mali, who, being herdsmen, have centred their mythology on the sacred substance milk:

> At the beginning there was a huge drop of milk.
> Then Doondari came and created the stone.
> Then the stone created iron;
> And iron created fire;
> And fire created water;
> And water created air.
> Then Doondari descended the second time. And he took the five elements
> And he shaped them into man.
> But man was proud.
> Then Doondari created blindness and blindness defeated man.
> But when blindness became too proud,
> Doondari created sleep, and sleep defeated blindness;
> But when sleep became too proud,
> Doondari created worry, and worry defeated sleep;
> But when worry became too proud,
> Doondari created death, and death defeated worry.
> But when death became too proud,
> Doondari descended for the third time,
> And he came as Gueno, the eternal one,
> And Gueno defeated death.
>
> (Eng. trans. by Ulli Beier in *The Origin of Life and Death*, 1966.)

According to nearly all African mythologies, God first agreed to give man eternal life, but his message was perverted through the stupidity or malice of the messenger. Several hundred variants of the myth of the perverted message are known, distributed over almost the whole of Africa south of the Sahara.

The most elaborate pantheons of gods are probably those found among the Yoruba of Nigeria and the Fon of Dahomey. These deities (*orisha* or *vodun*) are often seen as being, at one and the same time, legendary kings and founders of cities and supernatural spirits and controllers of the elements.

The story of a god must often be gleaned from cryptic references embedded in his praise names (*oriki*, or *mlenmlen*). These are curious mixtures of praise, description, joking abuse, and prayer; and, although the phrases have been handed down for centuries, the singer is free to add

new ones. Thus, Shango (the Yoruba god of thunder and lightning) is compared to the power and rumbling noise of the railway.

The most widely used poetic form

Poetry. The praise name is probably the most widely used poetic form in Africa. It is applied not only to gods but to men, animals, plants, and towns. Most important in many African communities are the praise names of chiefs and war leaders, as, for example, those of the great Zulu chieftain Shaka:

> He is Shaka the unshakeable,
> Thunderer-while-sitting, son of Menzi.
> He is the bird that preys on other birds,
> The battle-axe that excels over other battle-axes.
> He is the long-strided pursuer, son of Ndaba,
> Who pursued the sun and the moon.
> He is the great hubbub like the rocks of Nkandla
> Where elephants take shelter
> When the heavens frown . . .
> (Eng. trans. by Ezekiel Mphahlele)

But there are numerous other poetic forms: the Yoruba distinguish between praise names (*oriki*); the poetry of lineages and towns (*orile*); oracle verse (*odu*); hunters' songs (*ijala*); the poetry of masqueraders (*iwi*); incantations (*ofo*); songs (*orin*); and improvisations (*rara*). Incantations play an important part in all African traditions. The Ibo diviner, for example, invokes truth before consulting his bones:

> What will it be today?
> Success or failure? Death or life?
> Ha! The flood cannot run uphill.
> What is the evil spirit that throws his shade
> between me and truth?
> I hold my sacred staff against it.
> Here is the East, there is the West;
> Here the sun rises—
> See the truth come riding on the rays of the sun.
> (Eng. trans. by Romanus Egudu, in *Black Orpheus*)

A common poetic form is the magic formula, in which the meaning of the words is often obscure. These formulas are sacred combinations of words the correct repetition of which, accompanied by the proper ritual, is believed to be effective both in curing and cursing.

Probably the most elaborate body of poetry is that of the Ifa oracle among the Yoruba. Even the most learned priest is not expected to know it all, and the recital of the most important *odu* takes a whole night. The poems are accompanied by stories that constitute the mythological or historical precedents by which the diviner judges his client's case.

Every situation in African life is accompanied by poetry and song. The herdsman praises his bull ("My bull is dark like the raincloud in a storm" [Dinka]); the young warrior sings of his bride ("Neither her heel nor her palm are rough, but sweet to the touch like liver" [Fulani]); children invent a song to comment on an important event ("Europeans are little children, at the riverbank they shot an elephant, its blood became a canoe and it sank" [Nyasa]); the widower mourns the death of his wife ("What are your wares that they sold out so quickly?" [Akan]).

Animal-trickster stories

Folktales. The best known type of African folktale is the animal-trickster story. In Bantu Africa (East, Central, and South Africa) and the western Sudan, the trickster is the hare; in West Africa, the spider (Ghana, Liberia, Sierra Leone) or the tortoise (Ibo, Benin, Yoruba). The dominant characteristic of the trickster is his ingenuity, which he often uses to gratify his greed and satisfy his malevolence. But, despite his faults, he enjoys the audience's sympathy because by his trickery he usually defeats bigger and stronger animals. The Yoruba consciously poke fun at their own faults when they tell stories of the tortoise-trickster. Sometimes the tortoise's cunning defeats itself, as, for example, in the delightful tale in which the tortoise steals from the gods a calabash that contains all the wisdom in the world. He hangs it around his neck and is so eager to get home with it that, when he comes to a tree trunk lying across the road, he is unable to cross it because the calabash gets in his way; and in his anxiety he fails to think of putting it on his back. Frustrated, the tortoise smashes the calabash, and so, ever since that day, wisdom has been scattered all over the world in tiny pieces. Anansi (the spider-trickster, of whom stories are told among the Akan of Ghana and, as Anancy, in Jamaica) often appears as a mythological figure, the fiendish opponent of the sky god who steals the sky god's stories or tricks him into allowing disease to enter the world. In this function he shows some similarity with the Yoruba trickster god, Eshu, who consistently opposes the other gods and thwarts their intentions. There are a number of African tribes that also have story cycles about human tricksters—for example, the stories of Yo in Dahomey, the Lay cycle of the Iraqw in Tanzania, and the Zulu Uhlakanyana stories.

A variant of the trickster story is the escape story, in which the hero extricates himself from an impossible task by imposing an impossible condition. One such story tells how a cruel king of Benin ordered his subjects on pain of death to build a new palace but to start at the top and build downward. All were in despair until one wise old man went to the King and said that they were now ready to begin and would he, according to tradition, come and lay the foundation stone?

Dilemma tales

Stories of another kind, told for entertainment, usually on moonlit nights, are dilemma tales, in which the audience is invited to supply the ending or solution. An example of a dilemma tale told among the Wolof (Senegal, The Gambia) is as follows. Three brothers journey to a strange land and are all married to the same girl, with whom they sleep in turn. One night she is murdered by a robber, and the eldest brother, with whom she is sleeping, is condemned to death on suspicion. He begs leave to visit his father before he dies. When he is late in returning, the second brother offers to die in place of him. As he is about to be executed, the third brother steps forward and "confesses" that he is the murderer. But, at that moment, the eldest brother rides in, just in time to embrace his fate. Which of the brothers (the listeners are asked) is the most noble?

In another group of tales, prominent features of animals are explained (*e.g.*, ". . . so that is why the tortoise has a broken back").

Proverbs and riddles. "Proverbs are the palm oil with which words are eaten," say the Ibo. The art of conversation and argument depends, in fact, on their use. By them the speaker shows his learning, and often only the first half of a proverb is spoken. Thus, when a man says, "We know whom we love . . .," the audience knows that what is implied is ". . . but we don't know who loves us." Use of proverbs also enables the speaker to attack an opponent obliquely, without mentioning his name or the subject of the dispute. Proverbs express not only a people's inherited wisdom and code of behaviour ("If a child washes his hands, he will eat with kings" [Ibo]) but also their imagination and sense of humour ("If the earthworm does not dance in front of the cock, he will still be eaten, but at least the cock cannot say that he was provoked" [Yoruba]). The largest collection of African proverbs, made by a Swiss Lutheran missionary, J.G. Christaller, and published at Basel in 1879, contains 3,600 in the Twi language of Ghana.

Riddles usually take the form of a statement, not a question. So, in the riddle "People run away from her when she is pregnant, but they rejoice when she has delivered" (the answer is a gun), the question "What is it?" is understood. Often the riddle is an exercise in metaphorical speech, intended to display the questioner's imagination rather than to test the cleverness of the audience (*e.g.*, the Yoruba "We tie a horse in the house, but its mane flies above the roof, to which is answered "fire" and "smoke").

Riddles used as metaphors

Survival of oral tradition. By the mid-20th century many oral traditions, particularly those associated with traditional ritual, were disappearing, and poetic forms were gaining new content and application (*e.g.*, praise names, applied to politicians; songs of abuse during elections, often taking the form of incantations). Nationalism and higher education tended to make Africans more conscious of their cultural heritage, and the collection and

conservation of oral traditions was no longer left to foreign anthropologists. Some of the best, most authentic collections are those by African writers and scholars that began to be published in the 1930s and that increased in number after World War II.

African folklore has come most powerfully alive in modern writing in the works of Amos Tutuola. *The Palm-wine Drinkard* (1952) and *My Life in the Bush of Ghosts* (1954) are inspired re-creations of the world of Yoruba folklore. Tutuola, whose long rambling tales are improvisations on traditional themes, is a born storyteller. Better than any exact translation, his books convey the rich inventiveness of traditional Yoruba folktales, the rather grim humour of which is often mingled with fear, pain, and other extremes of sense or impression.

MODERN LITERATURES

Though the written literatures in Swahili, in the main Bantu languages—Zulu, Xosa, and Sotho—and in Yoruba antedate in origin those in European languages, they are discussed after the more considerable modern literatures in French, Portuguese, and English—the so-called metropolitan languages of Africa. Ironically, African nationalism and anticolonialism had to find expression in the languages of Africa's colonial rulers.

French. The first contemporary written and printed literature to emerge was born as a protest against French rule and the policy of assimilation. Its leading figure was Léopold Sédar Senghor, who in 1960 was elected first president of the Republic of Senegal. In Paris during the 1930s he met Negro writers from the French Caribbean such as the poets Aimé Césaire from Martinique and Léon Damas from French Guiana. Together, they began a critical examination of Western values and a reassessment of African culture, and they founded *Présence Africaine* (1947–), Africa's leading literary journal. Senghor's anthology *Anthologie de la nouvelle poésie nègre et malgache* (1948; "Anthology of the New Negro and Madagascan Poetry") was an important influence in the formation of the idea of a Negritude (a term first used by Césaire) that should include poets outside Africa—in the French Caribbean territories and in Madagascar, for example. Senghor's poems are sometimes regarded as examples of 20th-century French poetry in the manner of Paul Claudel or St. John Perse, but they are, in fact, essentially African: his love poetry, in particular, is intensely so in structure and tempo. His themes are those of Negritude: he attacks the soullessness of Western civilization ("no mother's breast, but only nylon legs") and proclaims that African culture alone has preserved the mystic warmth of a life that could still revive "the world that has died of machines and cannons." This culture, says Senghor, gains strength from its closeness to nature and constant contact with "the ancestors"; Western culture is out of step with the world's natural and ancient rhythm. Therefore, he proclaims, "[We are] the leaven that the white flour needs." In his long rhapsodic lines he has tried to make the French language swing and dance in the rhythms of his native Serer language.

Poetry of Senghor

Another Senegalese, Birago Diop, has similarly explored the mystique of African life in a volume of poems, *Leurres et lueurs* (1960). His fellow countryman David Diop wrote the most violent and full-blooded protest poetry to be produced by the Negritude movement: "When civilization kicked us in the face, when holy water slapped our cringing brows. . . ."

Negritude novelists

The most important Negritude novelists are the Cameroonians Mongo Beti (pseudonym of Alexandre Biyidi), who wrote *Le Pauvre Christ de Bomba* (1956; *The Poor Christ of Bomba*, 1971), and Ferdinand Oyono, author of *Une Vie de boy* (1956; *Houseboy*, 1966) and *Le Vieux Nègre et la médaille* (1956; *The Old Man and the Medal*, 1967). All three novels aim to explode the French colonial myth of France Outre-Mer: that the French West African possessions were not really colonies and that educated Africans are thus simply "black Frenchmen." Both Oyono's main characters—in the first novel, the houseboy; in the second, the old man—live under the illusion that they are the white man's friends and protégés; both learn through bitter experience that the white man is merely exploiting them. Beti's novel is similarly a novel of disillusionment, but here it is the white man, Father Dumont, whose ideals are shattered. After many years as a missionary in Africa, he is finally forced to realize that under a colonial system the missionary becomes the involuntary servant of the administration and so, whatever his personal convictions, is inevitably part of the colonial setup.

The second generation of French African writers was less concerned with the public rhetoric of Negritude. Thus, though the Congolese poet Tchicaya U Tam'si sometimes speaks of his people's sufferings ("My race remembers the taste of bronze drunk hot"), he does not claim to be the spokesman of his race. In *Le Mauvais Sang* (1955; "Bad Blood"), *Feu de brousse* (1957; *Brush-Fire*, 1964), *À triche-coeur* (1960; "A Game of Cheatheart"), *Épitomé* (1962), and *Le Ventre* (1965; "The Belly"), he explores his personal agonies in Surrealist poems in the dense texture of which mythological, Christian, and sexual imagery is juxtaposed.

Camara Laye became famous with his romantic autobiography *L'Enfant noir* (1953; *The Dark Child*, 1954), which draws a poetic, idyllic picture of life in a traditional African town. His most important work, however, is the novel *Le Regard du roi* (1954; *The Radiance of the King*, 1956), which describes a white man's quest for personal salvation in the mysterious atmosphere of the West African jungle. It is regarded as among the most imaginative novels to have come from Africa; in its haunting atmosphere and nightmarish intensity, it reminded some critics of the Czech writer Franz Kafka, though the debt to Kafka is limited, and the book is in many respects profoundly African and Islāmic. In a third novel, *Dramouss* (1966), Camara Laye, who in 1965 became a political refugee in Dakar, attacks the harsh methods of Guinea's ruling party.

Importance of *Le Regard du roi*

In the 1960s there was an important development of the philosophical novel in French-speaking Africa, notably by Sheikh Hamidou Kane in *L'Aventure ambiguë* (1961) and by Yambo Ouologuem in *Le Devoir de violence* (1968; "Bound to Violence"). Both writers belong to the Islāmic western Sudan and present their novels partly in the form of "dialogues," either between Islām and Western materialism or between traditional autocracy and Christian compassion.

Portuguese. The 20th-century poetry of Portuguese Africa, know chiefly through Mário de Andrade's *Antologia da Poesia Negra de Expressão Portuguesa* (1958), is mainly extremely militant. Aldo do Espirito Santo sings passionate songs of freedom: "And the blood of lives fallen/in the forests of death,/innocent blood/drenching the earth/in a silence of terrors/shall make the earth fruitful,/crying for justice." Both Mário de Andrade and the more important Agustinho Neto became actively engaged in the Angolan liberation movement.

José Craveirinha, an *assimilado* (*i.e.*, assimilated to Portuguese culture and Roman Catholicism), like all the African writers who have published work in Portuguese Africa, is yet deeply concerned with the problems of race and discrimination: "All feel/uneasiness/at the undoubted whiteness of my bones/white as the breasts of Ingrids or Marias/in Scandinavian lands. . . ." Lourenço Marques, the centre of a considerable literary and artistic activity, has also produced Luís Bernardo Honwana, one of Africa's outstanding short-story writers.

English. Although a genuine African literature in English did not emerge until the 1950s, writing in English by Africans goes back to the 18th century. Perhaps the most remarkable of the works published in England by West Africans sold into or born in slavery and later freed is *The Interesting Narrative of the Life and Adventures of Olaudah Equiano or Gustavus Vassa, the African* (1789; 3rd ed., enlarged, 1790), the first account by an African of his native country. In the 19th century, with the spread of missionary schools, several Africans published prose works in English, among them Samuel Adjai Crowther (see below *Yoruba*), Africanus Horton, and Edward Blyden; in the early 20th, the Ghanaian

18th- and 19th-century English writings

minister of religion Rev. Carl Christian Reindorf wrote one of the first works by an African to make use of oral traditions, *The History of the Gold Coast and Asante* (*c.* 1911); and Casely-Hayford, also a Ghanaian, wrote *Ethiopia Unbound* (1911), a strange mixture of novel, autobiography, and pamphlet.

In the 1940s great advances were made. Popular novelettes, known as Onitsha novels, from the name of the famous market where they were sold, began to be written, and a number of pioneer writers—among them Michael Dei-Anang and Raphael Armattoe of Ghana, Dennis Osadebay of Nigeria, and Edwin Barclay of Liberia—began to publish poems that, technically often crude and somewhat didactic in theme and style, prepared the way for later poets.

The institution in 1948 of university colleges at Accra and Ibadan caused a sudden increase in the number of West Africans receiving a literary education, and the upsurge of the 1950s and 1960s in poetry, the novel, and drama can be related to the emergence of this new generation. The most intense literary activity was in Nigeria, the Ibo proving particularly prolific. Cyprian Ekwensi, who began as an Onitsha pamphleteer, had had six novels published in London by 1969, the most successful being *Jagua Nana* (1961), the story of a Lagos prostitute. Onuora Nzekwu explored the relationship of the educated Ibo to his traditional culture in *Wand of Noble Wood* (1961), *Blade Among the Boys* (1962), and *Highlife for Lizards* (1965). Nkem Nwankwo created the comic figure of *Danda* (1964), the irreverent anti-hero of a novel perhaps inspired by traditional trickster tales. But the outstanding Ibo novelist was Chinua Achebe. All his novels present the conflict of emergent Africa: in *Things Fall Apart* (1958) he showed the impact of British rule on Ibo village life; in *No Longer at Ease* (1960) he analyzed the conflict in the mind of an African civil servant in Lagos whose life is torn between the social pressures of upper class urban life and the demands of his village union; in *Arrow of God* (1964) he again explored the breakup of traditional values and the struggle for political power in an Ibo village. His most committed novel, *A Man of the People* (1966), is a bitter, disillusioned story of political corruption and intimidation in independent Nigeria, ending—prophetically—with a military coup.

Literary activity in Nigeria

The development of Nigerian literature owed much to two related centres in Ibadan. The journal *Black Orpheus* (1957–), with Ulli Beier and Janheinz Jahn as its first co-editors, took its name from Jean-Paul Sartre's famous "Orphée noir" (1948), the introduction to Senghor's anthology (see above), and from Jahn's anthology, *Schwarzer Orpheus* (1954). Inspired by *Présence Africaine* and the First Congress of African Writers (Paris, 1956), *Black Orpheus* published the earliest efforts of African writers in English to find a literary mode of expression. A second influence on the establishment of an independent Nigerian culture was the foundation (1961) of the Mbari Club, a meeting place for new writers and artists that also published the work of many poets and playwrights; a third was the University of Ibadan, where many writers studied (and some later taught), thus, to some extent, avoiding the dichotomy set up in French African writers between their French education, with higher education abroad, and their African cultural heritage.

Role of the Mbari Club

Among poets published by Mbari was Christopher Okigbo, who, until his tragically early death in 1967 while fighting for Biafran independence, was the most polished and prolific Nigerian poet. In *Heavensgate* (1962) and *Limits* (1964), he developed a sensuous, esoteric style. His poetry reads like incantation: mysterious, often obscure, but always musical and intensely evocative.

Suddenly becoming talkative
like weaverbird
Summoned at offside of
dream remembered
Between sleeping and waking,
I hang up my egg-shells
To you of palm grove,
Upon whose bamboo towers hang
Dripping with yesterup wine
A tiger mask and nude spear . . .

(From G. Moore and U. Beier's *Modern Poetry from Africa*, 1963.)

Like Okigbo, John Pepper Clark graduated from Ibadan. His passionate *Poems* (1962) and *Song of a Goat* (1961) were first published by Mbari. Clark is the most urgent and immediate of the Nigerian poets. Less reflective than Soyinka, less meticulous that Okigbo, he works out each poem on the impetus of a single lyrical impulse. He has also written successful plays: *Song of a Goat*, which was performed in London at the 1965 Commonwealth Festival, is the tragedy of a woman, childless because her husband is impotent, who is advised by the priest to have a child by his younger brother. *The Ozidi of Atazi* (performed in the early 1960s; published, 1966) is a stage version of a traditional Ijo ritual play, which in his native village would take seven days to perform.

Works of John Pepper Clark

One of Nigeria's most sensitive poets is Gabriel Okara, whose novel *The Voice* (1964) is a fascinating linguistic experiment—he forces English into the straightjacket of Ijo syntax, so producing an extraordinary and archaic effect, often as beautiful as it is bizarre.

Africa's most famous playwright is a Nigerian, Wole Soyinka. By the early 1970s he had published nine plays. All are set firmly in contemporary West Africa. In his lighter mood he makes fun of pompous village teachers, as in *The Lion and the Jewel* (first performed Ibadan, 1959; published 1963), or of the cunning "prophets" of the upstart dissident African churches as in *The Trials of Brother Jero* (1962). More important, however, are his serious plays, revealing skepticism and disillusionment. In *A Dance of the Forests* (published 1963), written for and performed to celebrate Nigerian independence in October 1960, he criticized the myth of the glorious African past, rejecting the Negritude concept that the revival of African culture must be inspired by the African cultural heritage alone. *Kongi's Harvest*, which opened the World Festival of Negro Arts (Dakar, Senegal; April 1966), is a brilliant satire on African dictatorship. Though many of its "hero's" traits seemed to be borrowed from the Ghanaian political leader Kwame Nkrumah, Soyinka has made it clear that he had no one person in mind but was attacking the new trend in African nationalist politics.

Plays of Wole Soyinka

Soyinka's tragic view of life is most fully expressed in his play *The Road* (published 1965), first performed at the 1965 Commonwealth Arts Festival, and in his novel *The Interpreters* (1965). Publication in 1967 of *Idanre*, his first collection of poems, established him as a poet of distinction. Here, too, he explored his tragic sense of the difficulties and cost of human progress. He has always played an active part in public life, forming two theatre companies, 1960 Masks and the Orisun Players, and lecturing successively at Ife, Lagos, and Ibadan, where he headed the Drama School.

While Nigeria has dominated the literary scene in West Africa, other African nations have also contributed to the new literature. Among the most important poets in the late 1960s were a Gambian, Lenrie Peters, and a Ghanaian, George Awoonor-Williams (now known as Kofi Awoonor). The early work of both was published by Mbari: Peters' *Poems* and Awoonor-Williams' *Rediscovery and Other Poems* (both 1964). Peters brings an unusual scientific intelligence to bear in his poetry, while Awoonor-Williams is deeply concerned with his native Ewe poetic tradition and the means of extending this in English.

In Ghana Efua Sutherland sparked off a theatre movement inspired by folk drama, founding the Drama Studio in Accra. Her own principal plays are *Foriwa* and *Edufa* (both 1967). The most gifted Ghanaian playwright is the young Ama Ata Aidoo who, in *The Dilemma of a Ghost* (1965), explored the complex cultural conflict arising in a village in Ghana when a young man returns from his studies abroad with an Afro-American wife. Her second play, *Anowa* (1969), greatly extended her dramatic range with its study of a tragic misfit in traditional Ghanaian society.

In East Africa in the 1960s, literature was only just coming to birth; and the literary review *Transition* (Kampala,

Uganda; 1961–) played an important part in encouraging young writers. Already, however, James Ngugi had emerged as East Africa's leading novelist, with two novels—*Weep Not Child* (1964) and *The River Between* (1965)— in which he tried to do for his native people, the Kikuyu, what Achebe had done for the Ibo. In a third novel, *A Grain of Wheat* (1967), he reflected profoundly on the meaning of heroism and of national independence. But in the late 1960s the most original talent in East Africa was that of Okot p'Bitek, whose long poem *Song of Lawino* (1966) treats the conflict of cultures with freshness and imagination. The simple device of making it a lament spoken by an illiterate woman, who complains about the strange ways of her university-trained husband, provided p'Bitek with a new approach to a familiar theme and also helped him to find a new and vigorous language, closely based on his native Lwo.

Short stories of protest by South Africans

The emergent work of African and Coloured writers born in, though often living in exile from, the Republic of South Africa has been dominated by the political fact of apartheid. Most find the short story the best medium of protest, among the most successful being Bloke Modisane; Todd Matshikiza; Ezekiel Mphahlele; Richard Rive; a Cape Coloured writer, Alex La Guma; a journalist, Can Themba; and Arthur Maimane. In the late 1950s, *Drum* magazine provided an important forum for these writers.

Autobiographies are among the most moving and significant works by black South Africans. Mphahlele's *Down Second Avenue* (1959) has become a South African classic; and Alfred Hutchinson's *Road to Ghana* (1960) and Bloke Modisane's *Blame Me on History* (1963) are passionate documents in which the authors try to overcome by expressing it the humiliation of their experience as Negroes living in white South Africa.

The outstanding non-white novelist in South Africa in the 1960s was Alex La Guma. His superb eye for detail makes it possible for him to let the horror of a situation speak for itself, so that he never becomes sentimental or overemotional in describing it. His three short novels, *A Walk in the Night* (1962), *And a Threefold Cord* (1964), and *The Stone-Country* (1965), are tense works of protest. The subtlest treatment of the recurrent theme of race relations, however, was perhaps that by Lewis Nkosi in his play *The Rhythm of Violence* (1964).

The outstanding poets of the 1960s were all relatively young—Mazisi Kunene, William Kgositsile, K.A. Noortje, and Dennis Brutus. Brutus is a restrained and mature poet whose chief concern is to preserve human sensitivity in a world of brutality. Though personal suffering and the sufferings of his people loom large in his poetry, it never becomes self-pitying:

> Somehow we survive
> and tenderness, frustrated, does not wither.
> Investigating searchlights rake
> our naked unprotected contours;
> over our heads the monolithic decalogue
> of fascist prohibition glowers
> and teeters for a catastrophic fall;
> Boots club on the peeling door.
> But somehow we survive
> severance, deprivation, loss.
> Patrols uncoil along the asphalt dark
> hissing their menace to our lives.
> most cruel, all our land is scarred with terror,
> rendered unlovely and unlovable;
> sundered are we and all our passionate surrender
> but somehow tenderness survives.
>
> (From his first collection, *Sirens, Knuckles, Boots*, Mbari Publications, Ibadan, 1963.)

(U.Be./Ge.Mo.)

Bantu. Literacy came to most of Bantu-speaking Africa in the 19th century with the advent of missionaries and the establishment of mission schools. By the second half of the century, Africans were editing newspapers and participating in translating the Bible and John Bunyan's *Pilgrim's Progress* into the Bantu languages; and the first African minister of religion to be trained abroad, the Xhosa Tiyo Soga, completed his studies at Glasgow University in 1856. Generally, therefore, the period from the mid-19th century, when Western education was becoming established, to the early 20th, when Bantu-language books began to roll off the presses, was one of culture shock, affecting the majority of Africans through Christianity and a significant minority through education.

Influence of Christian missionaries on Bantu writing

Since missionary education had made them question the validity of their own culture, African writers at first turned their backs on their own rich heritage of story, song, poetry, myth, legend, and chronicle and adopted Western forms—the novel, drama, and rhyming and metrically structured poetry. In narrative, reform on Christian lines became an obsession: the "Pilgrim's Progress" theme, the "prodigal-son" theme, the "unquestioning-obedience" theme, the "reward-of-sin" theme, etc. abound. In third-person narrative the author often interrupts the story to sermonize, warn his characters of impending danger, or muse parenthetically over an incident. All this could also, of course, be partly due to the African heritage of didactic tales and of audience participation in their telling.

Christian influence is most prominent in characterization: in general, a character is either Christian and good or non-Christian and bad. Virtue is typified by an overemphatic version of the "turning-the-other-cheek" dictum, which degenerates into an inertia that tolerates provocation by simply not reacting, or by a robot-like reaction in which complete obedience is the ideal. As the "good" characters become less true to life, so the rogues are, in contrast, human, alive, fun loving. Because characters are either saints or satans, they lack that most intriguingly human trait—unpredictability. Change from satan to saint or saint to satan is often sudden and mercurial. Much concern is expressed at the threat to social cohesion arising from periodic absences from their tribal villages of the young men, who go away to find work in the mines or in industry. The "country" is associated with innocence, the city with corruption.

Problem parents are not spared, and, stylistically, this results in a transgenerational plot pattern. First, the background is given—the parents' mistakes in bringing up their children; then, the story proper shows its effects. The flashback technique seems unknown.

Drama has three main technical faults: "mobile scenes" allow an actor to move over a distance much wider than the stage can legitimately represent; characters offstage continue to be seen and heard, while those onstage are forgotten; and parts of the story are narrated, not shown in action.

In poetry, many have attempted to use rhyme and syllabic metrical structures unsuited to the Bantu languages, while many are apparently to a great extent ignorant of the phonology of their own languages and sometimes, therefore, "rhyme" letters, not sounds. When they use the techniques of heroic poetry—a genre forming part of their own cultural heritage—their poetry immediately improves.

Outstanding Bantu writers

These criticisms apply to many, but by no means all, writers in Bantu languages. Despite the need to adjust to new forms, many writers have achieved distinction. Foremost among them is Thomas Mofolo of Basutoland, who, for his three novels written in his native Sotho, is generally regarded as the first important African novelist of the 20th century. The third, *Chaka* (1925; *Chaka, a Historical Romance*, 1931), is a Sotho classic. Later writers include Xhosa A.C. Jordan (*Ingqumbo Yeminyanya*, 1940, "The Wrath of the Ancestors"), the Zulu poet B.W. Vilakazi, of Natal (*Inkondlo ka Zulu*, 1935 [revised 1965], and *Amal'Ezulu*, 1945; Eng. trans. together *Zulu Horizons*, 1962), the Xhosa poet S.E.K. Mqhayi (*Inzuzo* ["Gain"]), Kopano Mushashu, H. Citepo, and the Zulu J.L. Dube, whose *Insila kaShaka* (1930; *Jeqe, the Body-servant of King Shaka*, 1951), about the chief who brought the Zulus to power, was the first novel to be published by a Zulu in his native language. (D.P.K.)

Yoruba. The Yoruba (who live mainly in Nigeria and partly in Dahomey) have an unusually rich oral tradition of myths, legends, folktales, and songs. Their written literature probably began in 1844 with the translation of

the opening chapters of St. Luke's Gospel by Samuel Adjai Crowther, a former slave who, freed by British traders, trained as a missionary and finally became a bishop. As in the Bantu languages, early literature consists mainly of translations from the Bible, the hymnbook, and Bunyan's *Pilgrim's Progress.*

Novels of Fagunwa

Original literature began with *Igbo Olodumare* (1947; "The Forest of God"), the first of a series of fantastic novels by Chief D.O. Fagunwa that gained immense popularity among the Yoruba. By 1970 the total edition of Fagunwa's five novels ran into hundreds of thousands.

Fagunwa's novels are loosely constructed adventure tales containing many folklore elements: spirits, monsters, transformation flights, gods, magic, and witchcraft. But, though the plots are conventional, the style is lively, and Fagunwa's sense of humour is typically Yoruba. His language is visual: a sad man "hangs his face like a banana leaf"; a quarrel "sticks in the throat like a fishbone"; a liar "has blood in his belly but spits white saliva."

Fagunwa created a vogue for fantastic novels among the Yoruba and probably had some influence on Tutuola, whose works in English belong to the same genre. Other writers tried to break away from this tradition, and I.O. Delano's *Aiye d'aiye Òyinbo* (1955; "It's a White Man's World") was the first of a series of realistic novels.

Modern Yoruba literature finds its most vigorous expression in the theatre, with its origins in the morality and biblical plays performed in the breakaway African churches. The first professional Yoruba theatrical company, Ogunde's Theatre Party, was founded by a playwright, Hubert Ogunde, in the 1940s. Ogunde's plays blend elements of music hall and slapstick with astute political satire. The most famous, *Yoruba Ronu* (performed 1964; "Yorubas, Think"), was such a biting attack on the prime minister of Western Nigeria that Theatre Party was banned from the region—the first instance of literary censorship in Nigeria. Ogunde laid the foundations of Yoruba theatre's commercial success and prepared audiences all over Nigeria for his followers.

His colleague E.K. Ogunmola (more successful as actor, mime, and director than as playwright) refined Ogunde's techniques, replacing saxophones with Yoruba drums and writing much more tightly constructed plays, gentle social satires, in which he laughs at, without passing judgment on, folly and weakness. His most typical play is *Love of Money* (performed *c.* 1950 and widely played under its English title), but his greatest success was with an adaptation of Tutuola's *Palm-wine Drinkard* (performed 1963).

Historical plays of Duro Lapido

As literature, the most rewarding dramas are Duro Lapido's historical plays. Making full use of the richness of traditional Yoruba poetry, Lapido created a trilogy about the history of the kingdom of Oyo—*Ọba Kò So* ("The King Did Not Hang"), *Ọba moro* ("The King of Ghosts"), and *Ọba waja* ("The King is Dead"), which were published together in 1964 as *Three Yoruba Plays.* In its entirety this work has the power and ultimate serenity of ancient Greek tragedy. The most outstanding Yoruba playwright, Wole Soyinka, though he writes in English, is not uninfluenced by trends in Yoruba theatre.

Yoruba is perhaps the only African language to produce a magazine devoted to contemporary writing. *Olokun* (1967– ; published twice yearly at Ibadan University) features the work of such modern poets as S.A. Babalola and A. Faleti. (U.Be.)

Swahili. Swahili is the only Bantu language with a tradition of written literature, in Arabic script. The oldest manuscript dates from the 17th century. Apart from some diplomatic correspondence and chronicles, until the 19th century, when missionaries began to collect stories from oral tradition, all Swahili literature was poetry. It drew its inspiration from the Islāmic literatures in Arabic, Persian, and Urdu taken to the east coast of Africa by Oriental traders and scribes. Though some early narrative poems in Swahili are translations or paraphrases of the long odes (*qaṣīdahs*) of the classical period of Arabic literature, none has been traced to an Arabic epic.

Swahili epic poetry is written in four-line stanzas, each line having eight syllables. The most important epics preserved are the *Muhamadi* (6,280 stanzas), a versified life of Muḥammad, the longest written epic in any African language; the *Rasi 'l-Ghuli* (4,384 stanzas), describing the punitive expedition of ʿAlī, Muḥammad's son-in-law, against the evil king whose name gives the poem its title and the conquest for Islām of his country, the Yemen; the *Herekali*, or *Tambuka* (1,145 stanzas), the oldest Swahili epic (the earliest manuscript is dated 1141 [AD 1728]), which describes Muḥammad's legendary campaign in AD 630 against the Byzantine emperor Heraclius; and the *Huseni* (1,209 stanzas), about the life and death of Ḥusayn ibn ʿAlī. Its representation in Swahili literature is of particular interest for the study of Shīʿism (a branch of the religion of Islām) in East Africa.

Epic poetry in Swahili

Other epics (by 1972 more than 70 titles were known) treat Muḥammad's wars against the unbelievers or tell stories of the women in his life (Khadījah, ʿĀ'isha, Fāṭimah) and of the biblical prophets such as Adam, Ayubu (Job), and Yusufu (Joseph). In the early 19th century, secular poetry began to be written, and an epic cycle of *c.* 1900 describes the Swahili wars against Germany. The tradition of the national hero King Liongo, *Liongo Fumo*, a complex of songs that plays an important part in Swahili culture, is of great interest for the study of myth and ritual.

The first secular poet known by name is Muyaka ibn Haji al-Ghassaniy (1776–1840), whose style remains popular. His poems were edited as *Diwani ya Muyaka . . .* (1940). Two important 20th-century poets were Sheikh Amri Abedi and Mathias E. Mnyampala.

Most prose writings are based on traditional legends in a vivid Oriental style, first collected and recorded by Europeans. Modern writers of this type of Oriental saga are David Diva and Omar Shariff. The first writers to introduce the novella form into Swahili literature were James Mbotela in *Uhuru wa Watumwa* (1934; "Freedom for the Slaves") and Muhammad Saleh Abdullah Farsi in *Kurwa na Doto* (1960), a story set in his native Zanzibar. The first detective writer is Muhammed Said Abdulla, whose *Mzimu wa Watu wa Kale* appeared in 1960.

Swahili is the only Bantu language in which history was recorded before the arrival of the Europeans. The town chronicles of Lamu, Mombasa, and Pate have been published, but the Swahili version of that of Kilwa is lost. An important publication is Liajjemi's chronicle of the Wakilindi (1962).

The greatest 20th-century writer in Swahili, Shaaban Robert, stood between two civilizations (his father was a Christian; he returned to Islām) and two literary periods. He wrote didactic stories, Oriental in tradition and style, and also introduced the essay. His fluency and command of language reveal him as a true master of letters. (Ja.Kn.)

III. African music

Music composed and performed in traditional forms by indigenous Africans south of the Sahara is considered to be "African music." North of this area there prevails a distinct musical style heavily influenced by Islāmic music. This definition of African music excludes the performance by Africans of music whose origin is clearly foreign to their communities, such as the musical styles of America, Europe, and the Middle East, prevalent in certain regions of Africa. African music is the "classical" music of the continent. In that it is transmitted from ear to ear with uninterrupted continuity from past generations of Africans, it is also folk music. It is constantly evolving as performers add or subtract elements of the music at will while remaining within the accepted framework of local practice. This process gradually transforms the original or previous conception of a given piece of music into something different but always contemporary.

Although there are many tribes that have little tradition of carving or other plastic art, there is no community in the whole continent without its own music, which in most cases is considerably complex. African music is characteristically polyrhythmic (consisting of simultaneous, distinct rhythmic patterns) and polyphonic (consisting of simultaneous sounds or melodic lines). Like folk musi-

cians elsewhere, African musicians do not actively conceptualize the abstract principles of their music. Yet the unhesitating participation by all members of an African community in a musical performance implies the existence of common, though unverbalized, principles underlying their music making.

This discussion of African music will survey briefly the relation of music and society in Africa, the process of musical composition, and the characteristics of musical scales, songs, and melodies. There will follow more detailed discussions of rhythm, of part singing and harmony, of interlocking techniques and inherent note patterns. Finally, there will be an examination of the different instruments used in African music.

MUSIC AND SOCIETY

Social functions of music. As with largely preliterate communities elsewhere, most African music is socially controlled; *i.e.*, organized in relation to the activities and needs of everyday life. Use of music in a social or private event may be either traditional or spontaneous. Traditional musical activities are mostly combined with such fixed occasions as festivals, rites, and religious celebrations. Spontaneous music making occurs in the private and recreational spheres. In traditional combinations, there are rigid prescriptions regarding roles, responsibilities, general framework, and performance. In spontaneous music, the instruments, songs, and texts—with certain exceptions—enjoy great freedom of expression. Traditional musical activities include marriages, funerals, rites of cult groups such as warrior associations, communal labour, celebrations in connection with chiefship, and worship of the gods. In many cases, the musical activity takes much of the actual time of the event. Such places of performance as the village square, the street, the shrine, the courtyard, or the chief's palace are the musical centres of African communities.

Songs associated with life cycles

Throughout Africa, songs of several kinds are associated with the various ages of life—from infancy, through childhood, adolescence, and maturity, to old age. The repertoire of a single village community may include several hundred songs in traditional styles, all of which are appropriate to one or another of the many seasonal, religious, and social occasions. In addition, secular dance songs and songs of entertainment follow the familiar cycle of novelty, popularity, and decline within a relatively short period. The currency of a dance or popular song is usually longer in rural areas than in towns.

Work songs are the constant companion of much manual labour such as hoeing, threshing, and pounding grain in a mortar. These songs provide the rhythmic compulsion that keeps the work going at a steady pace and thus makes it a form of dance. In Malawi (in southeast Africa along Lake Nyasa), for example, three women will pound together in one mortar, alternating their blows either in rhythm with their song or in cross-rhythm with it (*i.e.*, with rhythms at variance with those of the song). In many situations of communal labour in Africa, there occurs the same responsorial (soloist/group) principle used in the sea chantey in Europe and America: the leader singing different phrases and the body of workers answering with the refrain. Hauling, lifting loads, digging, and hammering are typical rhythmic activities in which responsorial singing is commonly employed.

Cross-cultural influences. During its long history, African music has been subject to many changes. Most influential in recent times is its encounter with Western music, religious as well as academic and popular. Such influences include Afro-American music by way of jazz and Afro-Brazilian music. Dating further back was the Arabization of the African Middle Ages. Arabization in West African music, for example, led to the reduced use and sometimes complete disappearance of harmony. Melodies were also influenced: in Arab-influenced areas there is a preponderance in vocal and instrumental music of the portamento (sliding from one note to the next) and ornamental melodic style characteristic of Arabic music. The Sudanic area (stretching from the Atlantic coast eastward to Ethiopia, south of the Sahara and north of the Equator/), particularly the ancient Fulbe and Hausa states, reveals the greatest increment in Arabic and Oriental instruments. Among them are the *algaita* oboe; the long-tubed *kakaki* trumpet of the Hausa and Nupe peoples; and a number of bowed string instruments.

Impact of Western music

The complex structure of African music has unwittingly created social and psychological repercussions as the continent has developed in the 20th century. Western educators at first dismissed African music as "primitive," and substituted for it in the schools the simpler forms of Western music. Missionaries, fearing the possible association of African music with heathendom, discouraged it. Both professions, together with the phonograph industry, represented the weight and authority of successful material culture. They have persuaded a proportion of Africans—especially those in towns and institutions—to accept their foreign musical prejudices. The result has been the adoption of a popular facade of exotic (to Africa) and semi-exotic music. Such music lacks the normal structure and continuity of African traditional music and is constantly spurred into activity through the usual channels of commercial publicity. Thus in modern Africa, there are the many forms of Western-influenced urban social music that, originating in the streets, conquered the ballrooms, the radio stations, and the commercial disc market. In Ghana and Nigeria, it is named highlife. Its main features are intertribal appeal and the use of Western band instruments and of chords typical of the Western system of harmony. In addition, some African musicians and scholars have begun composing works aimed at the concert hall, church, and school, but which seek to recreate traditional musical styles.

The influence of Western music is important in that it has stimulated misgivings about the integrity of indigenous African culture. Thus many African leaders stress the need for imitation of foreign arts rather than originality. No such doubts appear among the active practitioners of African music who have not broken the thread of artistic continuity. It is to these musicians that future Africans will look for inspiration if they wish to achieve further maturation of their unique musical art.

THE STRUCTURES OF AFRICAN MUSIC

Diverse historical backgrounds, intense attachment to regional spoken languages, widely contrasting environments, and great distances between peoples have created a rich repertoire of many differing styles of African music. They have, however, many characteristics in common and there is evidence pointing to certain degrees of uniformity in the abstract principles underlying African music.

Speech tone and music

In structure, African melodies are closely associated with each of the many languages of the continent. In addition to deriving distinctive characteristics from the stress and rhythm of the spoken language, they are influenced by tone. In most African languages, the meaning of a given sound depends on the tone at which it is spoken. For sung words to be understood, melodies must rise and fall in agreement with the tones of the spoken words. Dependence of melody on speech tone is a common characteristic of musical styles throughout the continent. Other common traits include spontaneous creation; improvisation by performers within traditional frameworks of melody and rhythm; frequent use of polyphony and harmony; and complex rhythmic organization.

Composition. Many composers of African songs maintain that their melodic inspiration is founded upon the tone profile of spoken phrases in their own language. Good craftsmanship demands that sung melody should follow as closely as possible, with certain exceptions, the rise and fall of normal tone sequences. In a song of more than one verse, the initial melody must be modified to accommodate the tone sequences of the new words. In instrumental music, when the composer discovers a satisfactory sequence of notes, he plays it immediately upon his instrument to crystallize its form, "to get it into his hands." In this way, both the tone contour of sung words and the physical action of the hands upon the instrument

act as memory aids to the composer, equivalent to the function of musical notation in other cultures.

The composer of new songs or the person who adds new words or variations to an established style of music is recognized as an important member of society, for he helps to crystallize the general opinions and taste of the community. In preliterate African society, only those songs that are popularly considered to be valid remain in common use and can therefore be taken as genuine expressions of local sentiment. Songs of approval or disapproval voiced by all the village may have profound social and moral effect upon behaviour. Songs of praise or derision may be directed toward individuals or groups, and, if supported by public opinion, may have considerable political significance.

Status of composers and performers

Musicians are not usually considered as a class apart. Everyone normally takes part in the singing and dancing to the best of his ability or, if sufficiently skilled, in the playing of instruments. Gifted musicians may become semiprofessional through popular demand. In several communities, professional minstrels and small troupes of actors tour the towns and villages. They give shows with mime to the accompaniment of drums and other instruments and recite legends or tell stories. In many cases, songs are a characteristic part of the performance—a rudimentary form of light opera.

Musical scales. A musical scale is a particular series of notes within an octave (as, from c to c′ on a piano keyboard, encompassing eight white keys). To most African musicians, the series of notes in a scale is usually thought of in descending order from treble to bass, as with the early Greeks. Many Africans also have a "magnitude" concept of pitch according to which treble notes are "small" and bass notes "large." This contrasts with the Western "altitude" concept, with treble notes "high" and bass notes "low."

The choice and use of distinctive scales appear to be an identifiable characteristic of the music of different tribal or linguistic groups. Within a single community, the exact pitch of the tonic, or basic note of the scale, may vary; but the scale's pattern of intervals (distances between notes) remains relatively consistent whatever its starting point. The awareness of the "correct" ideal pitch of the intervals of the scale is notable among instrumentalists who play in ensembles, such as the xylophonists of the Chopi of southern Mozambique, the lute and lyre players of the Ganda of Uganda, and the *mbira* (thumb-piano) players of the Shona of Rhodesia and central Mozambique.

Basic scale types

Examination of the tuning of instruments of fixed pitch, such as the xylophone and *mbira*, shows that about 40 percent of African communities use pentatonic scales—*i.e.*, scales with five different notes in each octave, as, 1–2–3–4–5–(1′). About 40 percent use heptatonic, or seven-note scales, and the remaining 20 percent use hexatonic, or six-note scales. The interval of the perfect octave is common to most African scales. Other intervals usually considered important are the perfect fourth and the perfect fifth. These are sometimes modified, either up or down, by the application of some local tuning system, such as the equidistant system, which is found in many parts of Africa.

In the equidistant system, all of the intervals of a scale (whether pentatonic, hexatonic, or heptatonic) are the same size. The Western major scale, by contrast, uses intervals of two sizes, five whole tones and two half tones. Intervals as small as the Western half tone are rarely encountered in African scales. Nevertheless, there are ethnic groups like the Guro and Gere nations of the Ivory Coast whose harmonic singing gives the impression of being highly chromatic, due to the use of "augmented" and "diminished" melodic intervals (intervals wider or narrower than "normal").

More than one scale may be used in a single community for different purposes. Various classes of instruments are recognized as demanding their own distinctive scales, so that, for example, the unequal tuning of *mbiras* and xylophones within the same locality may prevent their being played together.

Song forms. Song types are often determined by their function: dirges for funerals, hunters' songs, heraldic songs for chiefs, praise songs, songs of insult, criticism, commentary, narration, proverbial songs, religious songs. They may be sung privately by individuals, with or without instrumental support, or they may be sung publicly by single performers, groups, choirs, with bands and orchestras, according to function and occasion. Group singing is particularly common and usually utilizes the basic formula of cantor and choir: soloist's lead and group's response.

Songs are usually built on short melodic units. A period, or unit, of two phrases (A,B), for example, may be repeated over and over. Such a form, called litanic from its repetition of a short melodic formula, can develop a number of extensions, for example A B A′ B′ or A B A C. The counterparts to such litanic forms are sequential forms, in which, for example, choir and cantor alternate sequences of unrepeated phrases, as A B C D. Litanic and sequential forms may also be employed in combination. The treatment of the text is in most African music strictly syllabic (*i.e.*, one syllable of text per one note of music). For song types with "open-end" forms (litanic, sequential) that can be added on to indefinitely, short melodic or rhythmic formulas are often used to signal the end of a piece.

Melody shape. The general pitch pattern of a melodic unit is falling; *i.e.*, it starts on a high tone and falls, either direct or in a series of steps, to a lower tone. The movement from note to note is often stepwise (without skipping any notes of the scale) in seconds (as, d–c) or in thirds (as, e–c). Intervals larger than a fourth rarely appear in sequence. Similarly a sequence of more than three ascending tones is unusual in most parts of Africa. Rising, falling, and level melodic direction is strongly guided by speech tone. Departures from this guiding relationship between words and melody are considered "artistic" and call for great ingenuity. (H.Ty./A.M.D.)

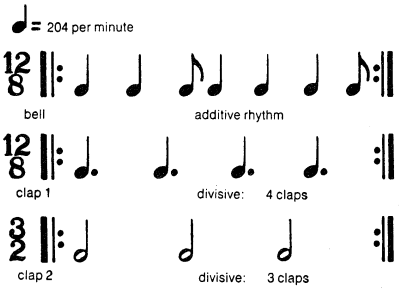
Rhythm in African music. African music is based on two very different forms of rhythm. Divisive rhythm divides time into equal pieces and results in a regular pulse (as in most Western music). Additive rhythm adds together little pieces of time of unequal length with a stress at the beginning of each piece, resulting in irregular beats. Further, African rhythmic technique falls into two distinct classes: the technique used in drumming on the one hand, and that used for songs, hand clapping, bells, rattles, and other instruments on the other. The latter may be called the "background rhythm," as it is against this that the drums usually play.

Background rhythm. The principle of background rhythm is well illustrated by songs accompanied by hand clapping. The essential principle is that the stresses of the song and the stresses of the claps must not keep in step. A rhythmic clash or tension is normally set up between the pulses of the claps and the stresses in the song melody. The simplest and most usual way of achieving this is to use a divisive rhythm for the claps and an additive rhythm for the song. Each clap will have either two basic time units per clap, or three. The irregular song stresses are bound to clash with either. The situation is more complicated if the claps use an additive rhythm, such as that often called by scholars the standard pattern, consisting of two-beat units (♩) and three-beat units (♩.).



The additive song stresses now cut across an additive clap pattern.

Still more complex is the frequent use by Africans of a multiple background rhythm. This may consist of several different clap patterns played simultaneously, together with a bell or two and a rattle, each with its own rhythm-pattern. These patterns will be either divisive or additive. Usually the claps will be divisive and may include clapping two against one, three against two, or four against three: the bell and rattle often have additive patterns.



Against all these, the song uses an additive pattern. The essential feature of such background rhythms is that they start their patterns on the same beat together, then cross over each other during the pattern, then start together again each time the patterns repeat.

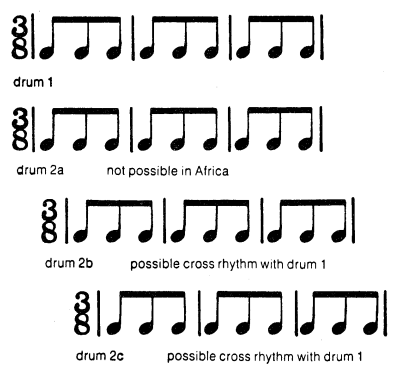
Influences on choice of rhythm

The additive rhythm of the song is not a fixed pattern in any one song. It arises from two factors: the shape of the tune and the natural emphases arising from it; and the normal speech accents of the song text. (Such accentuation is similar to the stresses in ecclesiastical plainsong.) There are also some songs, especially for children, in which the rhythmic stress is divisive. In such songs the melodic accent overrides the influence of speech accent.

All songs that take a clap must be organized so that in each section there are two divisive claps or a multiple of two (as in the above example, four claps by the first clapper). Or each section must contain three divisive claps or a multiple of three. Or there may be a combination of these, such as two plus three. The whole song, including all integral repeats, nearly always contains a total number of claps that is a multiple of two or three (such as 16, 24, or 36). When additive clap patterns are used in the background rhythm, they also normally add up to a total that is a multiple of two or three.

The songs dealt with so far are all songs that the African finds it natural to clap to. There is, however, another class of songs which, he says, cannot take a hand clap. Such songs are in completely free rhythm. They are necessarily additive in structure, for the rhythmic stresses arise once again both from the nature of the tune and the spoken accents of the text.

Drumming rhythm. The rhythmic technique of African drumming is fundamentally different from that of the background rhythms. The basis of the system is to set up a permanent cross-rhythm between each drum. Thus if two drums are playing in triple time (three-beat patterns), there are three possible ways of combining them. Only two combinations are normal in Africa (options *b* and *c* below):



Note that the main beat of the second drum must not coincide with the main beat of the first: in options *b* and *c* they are permanently crossed, a continuous rhythmic tension being thus established.

But drumming is not merely beating out rhythms. Each drum is capable of producing beats of different pitch and of different quality according to the manner of hitting the drumskin. Every drum pattern is therefore a rhythmic tune, and, furthermore, the tune of each individual drum is distinguished from all of the others by the basic pitch of the drumskin. In a drum set, each drum lies at a different pitch.

The example given above is only schematic: in practice, the first beats of each drum pattern will be crossed, but each drum will also be playing a different pattern. But it also shows yet another basic feature of drumming which is that whatever their patterns, and however crossed, the several drum patterns must be built out of the same basic time units. Thus if the basic time unit of the first drum is an eighth note played at the rate of 420 notes per minute, then the other drums must make their patterns out of this same unit of time. The only exception known is the rare case where one drum plays at four against three to the other.

Dance and instrumental rhythms

African dances exhibit the full flowering of African rhythmic procedures. Customs vary from tribe to tribe. An example is a case where a bell, a rattle, hand clap, song, and three drums—small, medium, and master drum—are used, as with the Ewe people in Ghana. The song, which has an additive rhythm, is first sung rubato by the leader (rhythm consciously diverging from strict beat). It is then continuously sung a tempo (in strict rhythm) by leader and chorus. The bell now enters, closely followed by the rattle, both usually playing additive rhythms. The people start clapping one or more divisive rhythms. The small drum starts playing a divisive rhythm whose main beat is permanently staggered with that of the bell. The master drum comes in with a long irregular pattern rhythmically crossed with everything else. Lastly, the middle drum enters with a short standard reply to the master's pattern (each of the master's patterns has its own reply on the middle drum). This pattern will be crossed with the other drums, and may be either additive or divisive. There is now a whole galaxy of crossed rhythms, and it is against this astonishing ensemble that the people start dancing, each dancer selecting one instrument to dance to, and thus dancing in cross-rhythm to all the others. Such a rhythmic thrill is entirely unknown in Western music.

Rhythmic variations

The master having established his first pattern by several repetitions now embarks on his main task. This is to play freely created improvisations using some fragment of the rhythmic pattern as a basic motif. His object is to work up the dancers to high enthusiasm and at the same time to accompany the song in suitable rhythmic terms. The variations are highly additive in rhythm. Having exhausted all possibilities of the pattern, he now starts on another one—to each dance there are a number of master-drum patterns. The middle drummer changes to the reply proper to the new pattern, the small drum keeps on with its original pattern, and the master starts variations based on his new pattern. African dance music is thus a combination of fixed patterns and free, spontaneous creation. One can never hear the same dance played exactly in the same way. The whole system is vitally creative and of ever-changing interest.

How can such a cross-rhythm ensemble end? There are two ways. In some areas, all that happens is that one drummer gets tired and stops playing: the cross-rhythm is broken and collapses, and the dance just peters out. In other areas (*e.g.*, among the Ewe in Ghana), the master drummer gives a signal on his drum, at which all the drums change to the "ending pattern," which, at its last few notes, brings all the drums into phase, with a final simultaneous beat by them all.

The full complexity of African rhythm is found preeminently in the forest areas—around the Gulf of Guinea and in the Congo basin. By and large, the further one moves from these areas the weaker becomes the cross-rhythmic setup, so that the music changes into something approaching the Western system, in which divisive rhythms tend to preponderate. Nevertheless, the above principles of background music and of drumming never disappear entirely. Dance music is the flower of African rhythmic virtuosity. All forms of rhythm in music, whether vocal or instrumental, are based on the essential rhythmic principles exhibited in the dance, although the degree of complexity varies considerably in different parts of the continent. (A.M.J.)

Multipart singing and harmony. Multipart singing and harmony in several regions of Africa have been observed by travellers from the earliest periods of Western contact with Africa. From the 18th century on, written transcriptions of such music have also been available. As early as 1719, for example, the missionary Peter Kolbe transcribed the music of the Hottentots (a Khoisan people) at the Cape of Good Hope. His account described a type of polyphony characteristic of the music of some Khoisan peoples even today. Numerous sources describing harmonic multipart music, sometimes with transcriptions, are found in the 19th and early 20th centuries.

With the beginning of the 20th century, phonograph recordings became available. Wax cylinder recordings made before World War I illustrate the harmonic techniques of several African peoples.

Basic principles of harmony

Two principles of harmonic organization can be distinguished in African multipart music: homophonic and polyphonic. In homophonic multipart singing, all melodic lines, though different in pitch, are rhythmically the same. They begin and end together. In polyphonic singing, the combining individual lines differ rhythmically. They have noncoinciding starting points and may be of different length. In the context of African music, both homophony and polyphony are usually considered linear, each combining voice having equal right. (In contrast, in Western part music, one part, usually the topmost, may be given more importance.)

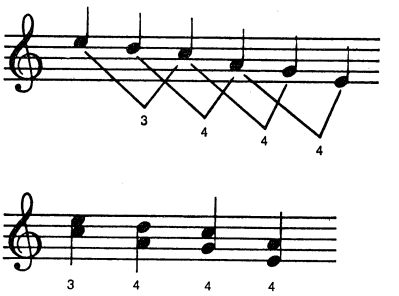
Geographical distribution of homophony

Homophony is the most important principle of traditional harmony in vocal music of Africa south of the Sahara. Found in several large geographical patches spread from West to Central, East and Southeast, and South Africa, it comprises many different stylistic types. Polyphony, on the other hand, is restricted in its geographical distribution. The vocal music of Khoisan communities (in the western part of southern Africa) is predominantly polyphonic, as are the vocal styles of the Pygmy tribes of both the Ituri (Zaire) and the Upper Sangha (Congo-Brazzaville, Central African Republic). In other parts of Africa, isolated islands of polyphony often occur among or between largely homophonic communities. Thus the otherwise homophonic Gogo, a people of central Tanzania, employ polyphony in certain songs, and Nyakyusa children of southwestern Tanzania use yodel and polyphony in some of their songs. A distinct style of polyphony is found in much of the music of the peoples of the lower Zambezi valley and parts of Mozambique, Rhodesia, and the Transvaal. Vocal polyphony also occurs in South Africa.

Homophonic part singing. One of the basic techniques in multipart vocal music in several African communities is that melodic lines thought by the singers to be identical are sung at different pitch levels. In many instances this leads to the rigid parallel singing of a melody at distinct pitch levels; for example, parallel fourths, one part beginning, say, on c, the other on f, or parallel thirds, the parts beginning, say, on c and e. Such parallel singing is homophonic, for all voice parts move together. There are two main streams of African harmony, some tribes using thirds, others using mainly fourths and fifths or octaves and unison singing. In many areas harmonic movement must be parallel because of the importance of speech tone.

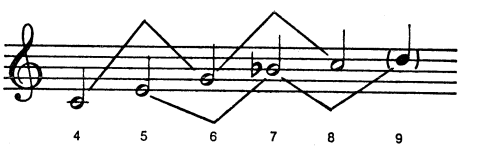
The actual harmonies that ultimately emerge from this parallel singing technique depend on the type of scale used by the singers. Thus parallel singing based on pentatonic scales necessarily differs from that based on heptatonic scales, because these scales have different intervals. Many communities in the harmony areas of Central and East Africa and some in West Africa form harmonies by a principle common to much harmonic part singing south of the Sahara: the harmonizing note for a given note is regularly found in the next note but one of the scale. In the case of two singers, the second one skips one step in the scale and duplicates the melodic line of the first singer at the new level. The precise harmonies that result from this "skipping process" vary from one African community to another according to the type of scale the musicians use; harmony and scale type are interrelated.

In many cases, the particular scale used allows the homophonic part singing to go beyond rigid parallel singing. An example occurs in some music of the Pangwa, Ndendeule, and other peoples in southern Tanzania and Malawi, specifically in one type of pentatonic scale. When the skipping process is used for harmonizing melodies, the result is a sequence of parallel fourths interrupted at one point by a third. Thus the example shows how notes in a pentatonic scale (e′–d′–c′–a–g–e) can be harmonized using the skipping process.

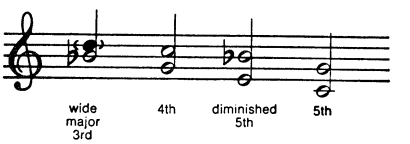


In another mode of the same pentatonic scale, d′–c′–a–g–e–d (from top to bottom), plain parallel fourths would be obtained.

A fascinating harmonic system results from application of the skipping process to the scale of the Gogo of Tanzania. Their scale is an extract of the natural harmonic series—the notes ("partials") produced, say, by a natural trumpet or by a string vibrating simultaneously in halves, thirds, fourths, etc. In this case, the sequence of notes is chosen from the higher partials, usually the fourth through the eighth, or more rarely, through the ninth. This scale gives rise to a typical chord progression that is the hallmark of most Gogo vocal music. The following shows the notes of the scale and their number in the series of partials.

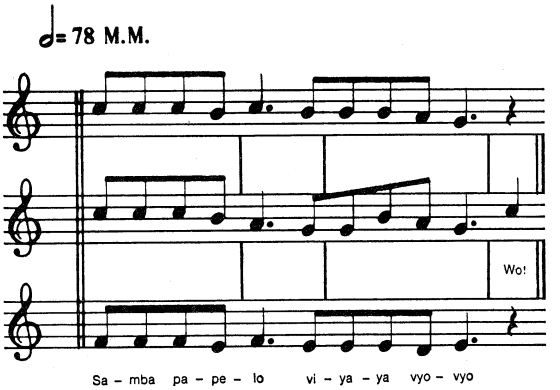


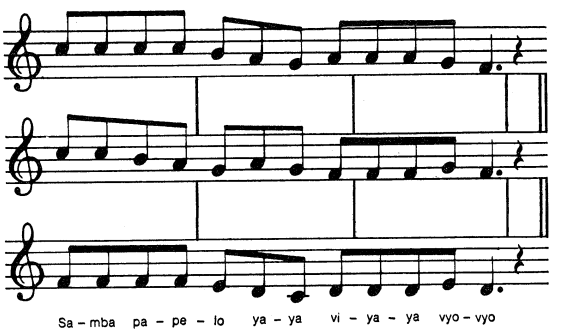
(The seventh partial is flatter in pitch than Western notation indicates.) The optional ninth partial is shown in black and in parentheses. Braces indicate notes forming chords when the skipping process is used. A typical chord progression resulting from this scale is

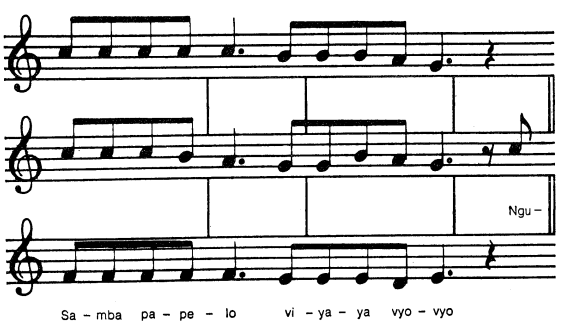


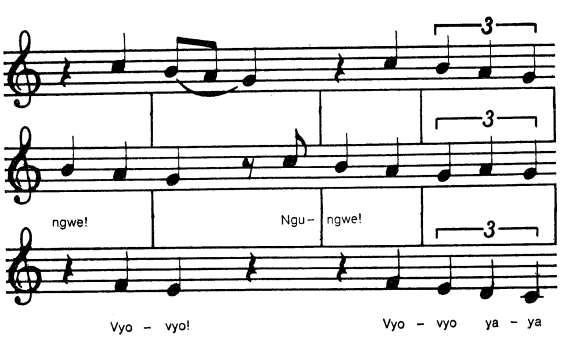
Chains of thirds

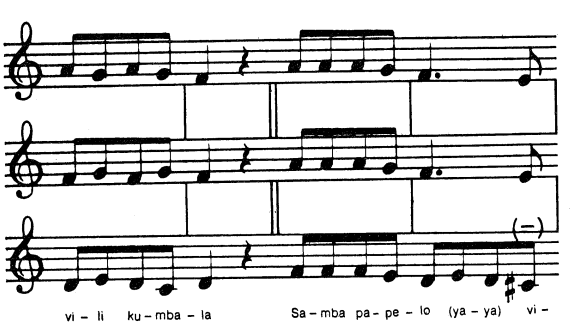
In heptatonic scales the application of the skipping process yields chains of thirds. According to the type of scale used these can be alternating major and minor, parallel major, or "tempered" neutral thirds. (Unlike the fourth, fifth, or octave, the third can vary somewhat in size without losing its identity. Western music uses major thirds, as, c–e, and minor thirds, as, c–e♭; a neutral third is anywhere in between in size.) One of the largest African heptatonic-scale areas using homophonic harmony encompasses most of Angola and northwestern Zambia and southern Zaire. Here the skipping process is simultaneously employed in two directions, both up and down the scale. The result is three-part harmony in triads (thirds plus fifths, as, c–e–g). Theoretically, all three lines move in parallel. Nevertheless, each singer can leave his own level at any time and move into that of his neighbour. The result of this liberty is that the harmonizing voices very frequently move in oblique and contrary direction relative to one another (rather than in parallel direction). The following extract of a song from Sakateke, Angola, provides an example of this type of homophonic harmony.

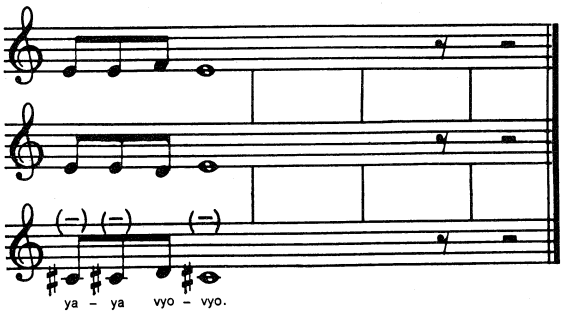






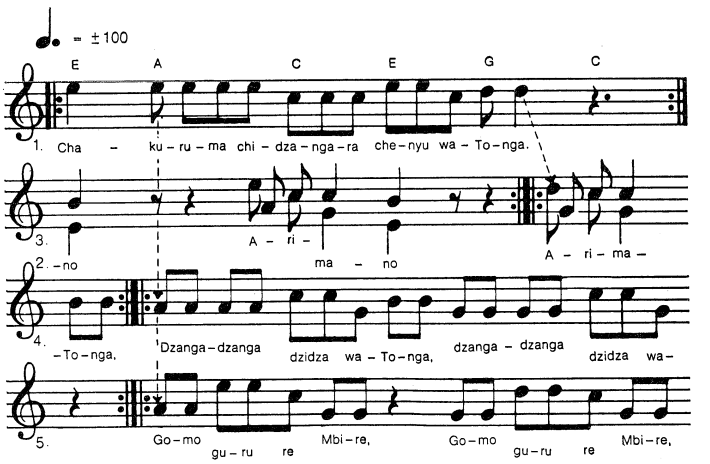






Polyphonic singing. In contrast to the type of homophonic part singing just discussed, in polyphonic part singing the individual voice parts are rhythmically independent. An analysis of a threshing song of the Karanga, one of the Shona peoples of Rhodesia and Mozambique, will illustrate what may happen in a polyphonic part singing area. This particular song, recorded in Rhodesia, is in five voice parts. Parallel movement is not allowed, so the skipping process cannot operate. Instead, the harmonic principle is that the parts must all follow a fixed chord sequence, here E A C / E G C. The letters refer to the basic notes of the chords. The chords themselves usually consist of two notes, a basic note, say c, and the fifth, as, c–g, or c–g′.

Using this chord sequence, vocal parts are constructed which contrast as much as possible melodically and rhythmically. The rhythmic cycle is 18 pulses long. As in drumming, cross-rhythms are set up: parts two and three answer part one at the end of its cycle; parts four and five begin shortly after the beginning of part one's cycle. The individual rhythm patterns of each part are also distinct. Part one uses divisive rhythm, dividing the 18 beats into six groups of three beats. The remaining voice parts use additive patterns built up of two-beat and three-beat units.



Instrumental music, such as for the *mbira* (thumb piano, see below *African musical instruments*), panpipes, and xylophones also tends in this part of Africa to be based on standard chord sequences. An example for the *mbira* is C E G / C E A / C F A / D F A. (A.Ty./G.K.)

Interlocking techniques. An important technique of combination in many kinds of African instrumental music is described as interlocking. This is one of the techniques of cross-rhythm and results in composite patterns made up by two or three instruments or parts alternating in playing notes. It is found in parts of East, West, and Central Africa.

A basic characteristic of interlocking is the absence of a common guide-pulse to be taken as reference point by all players. In a Western music ensemble, or a jazz band, all the players share a "beat," one common metric point of departure. They may even beat their feet to mark it. While there are many traditional African musics in which such a common reference pulse does exist (even where there is polymeter), in several others the musicians

in a group relate their parts to individual reference pulses, which can stand in various relations to each other.

One type of relation is for the pulse of one performer, or group of performers, to fall exactly in the middle of the other's. This type of interlocking occurs, for example, in the music of the *amadinda* and *embaire* xylophones of southern Uganda. A special type of notation is now used for these xylophones, consisting of numbers and periods. A number indicates that a player strikes a note; the value of the number refers to the note in the scale, as 5, the fifth note of the scale. An underlined number should be read an octave down; *i.e.*, 5 is an octave below 5. A period indicates that no note is struck. Numbers and periods occupy an equal space of time known as an elementary pulse, the smallest segment of time used in this type of music.

The following is an example of interlocking as played on the *amadinda*. The melodies are actually played in parallel octaves; that is, each melody is played at the notated pitch and also at the pitch an octave below:

Musician A plays: 4.3.4.1.3.3.4.2.3.4.2.1.4.3., etc.
(in parallel octaves) ↑
Musician B plays: ...5.3.3.5.5.3.5.2.3.5.1.1., etc.
(in parallel octaves)
↑ = place of entrance for musician B

In interlocking music of this type, one musician's positive action of striking a note always coincides with a negative action or "non-strike" of his fellow musician, who at that moment lifts his beater. The effect is such that both series of equal-spaced notes seem to interlock like the teeth of a cog wheel. Each of the two musicians, however, feels his own series of notes as "on beat."

Variations of the interlocking technique

In the very fast *mangwilo* xylophone music of the Shirima of northern Mozambique, the interlocking technique is exploited further. In some compositions by two virtuoso players, each musician interlocks with the right hand only. The left hands play different rhythm patterns superimposed over the interlocking pattern.

Triple interlocking is another type, used, for instance, in drum music in Zambia and also in southern Uganda in the music of the *akadinda* xylophone. Here a group of three musicians plays a short pattern of equally spaced notes in parallel octaves. Three musicians sitting opposite them interlock with another pattern that fits two equally spaced notes between each note of the first group's pattern. In numerical notation it looks like this:

Three musicians play: 2..4..2..2..1..1..2..5..
(in three octaves) ↑
Three opposite musicians play: .14.35.14.35.14.35.14.35
(in three octaves)

Interlocking techniques allow African instrumentalists to produce resultant patterns—overall patterns formed by all the players—that are unbelievably rapid. The resultant pattern of the above *akadinda* musical example is: 21443521423511413521453 5. This series of 24 notes, when played by expert musicians, is at a speed of approximately 600 notes per minute. But each musician, for himself, in terms of his own beat, plays one third that fast.

Inherent note patterns. Closely associated with interlocking techniques, but not necessarily depending on them, is the composition of inherent note patterns. These are rhythmic and melodic patterns that emerge when series of notes in distinct intervals are played at high speed.

The human ear does not perceive isolated particles of sound, but a "gestalt." When sequences of many notes are played rapidly, the ear cannot follow each note. As a result, the hearing tends to pick out and re-group the material forming several melodic–rhythmic patterns, seemingly independent from each other. Thus the total heard image of the music differs from the pattern of notes actually played. In a series of notes that are large intervals apart, for example, the ear picks out the notes of about the same pitch level and perceives them as a group. This psychological perception of a "gestalt"—an inherent note pattern—is an important element in listening to and composing some kinds of African instrumental music, particularly in Central and East Africa.

Function of inherent note patterns

Inherent note patterns are not accidental or coincidental; they are recognized and consciously employed by African musicians. In southern Uganda, there are even specific terms referring to them. The main function of inherent note patterns is to suggest words—text passages of the song that is outlined by instrumental accompaniment. Thus in the music of the *ennanga* harp of Uganda, the inherent note patterns suggest certain phrases of the vocal part. In a performance of the traditional harp song "Olutalo olw'e Nsinsi" ("The Battle of the Nsinsi") by the former court musician Evaristo Muyinda, one inherent pattern seemed to speak the words "Batulwanako ab'edda!" ("How They Forget Those Ancients!") long before they were actually sung. Muyinda often introduces a new phrase of text by first accentuating the corresponding inherent rhythm on the *ennanga*. Once the melody is firmly established as a "gestalt," it is sung. By slight accentuation or melodic variation during performance, the harpist may bring one or another of the already existing inherent patterns into prominence. This results in a musical development of the song's text.

(G.K.)

AFRICAN MUSICAL INSTRUMENTS

The existence of musical instruments in West, Central, and southern Africa has been noted since the 14th century. From the description of early travellers up to the 17th century, it is clear that African coastal communities had a rich and varied musical repertoire. The same can be said of most African peoples to this day and particularly of those who live in the interior in well-wooded regions where effective materials for the construction of instruments is close at hand. Present-day African musicians make and play several varieties of sophisticated instruments. But they also make good use of natural objects, spurning nothing that would suit their purpose. For this reason, the range of objects put to musical use in Africa may be far wider than in many other societies that rely exclusively upon conventional instruments of mass production.

Like all musical instruments, African instruments fall into four major categories, based on the way in which the instrument produces sound. In chordophones, or string instruments, a vibrating string produces sound. In aerophones, or wind instruments, a column or body of air is made to vibrate. In idiophones, the instrument's body itself vibrates, or parts of it; an example is a bell or xylophone. In membranophones, a stretched membrane vibrates; the prime examples are drums.

Chordophones: string instruments. Chordophones are musical instruments in which a string is set in motion by striking, plucking, or bowing and sometimes by a current of air. There are four main types: zithers, lutes, lyres, and harps. All are found in profusion throughout Africa.

The zither family. The strings of a zither are stretched between the two ends of the body. According to the shape of the body, zithers are classified as bow, stick, board, trough, or raft zithers. Most African zithers need an external resonator (besides the body itself) to increase the sound. Such resonators may be a small sheet of metal, a gourd or box, or often the open mouth of the player. The string or strings of a zither may be open (*i.e.*, one note to a string), or they may be stopped by the fingers to produce higher notes. Sometimes a string may be divided into two segments, each producing a different note. This is done either by pulling the string back toward the body at some point by a small cord, or by wedging a length of stick between the body and the string. Most stick and bow zithers have only one string. Trough, board, and raft zithers may have from five to eight (or more) strings on which chords can be played.

Variations of the zither

Several African communities use a friction bow, a bow zither in which a string or a thin band of palm leaf is made to vibrate by rubbing a stick along a series of notches cut into the instrument's body. In this case, the mouth is the usual resonator; the vibrating string is held across the open mouth. Harmonics are produced as on a Jew's harp, by altering the shape of the cavities of the throat and mouth as if by "silent" singing.

In the *lesiba*, a stick zither of the Sotho of Lesotho, the player sets the string in vibration by alternately blowing and drawing in his breath across a short length of quill attached to the end of the string. As long as the string vibrates there is an uninterrupted sequence of notes. The pitch is varied by stopping the string with the fingers at the far end. Harmony may be added by simultaneous humming or singing.

The lute family. Lutes comprise the family of string instruments that have a body and a neck along which the string or strings are extended beyond the body itself. A lute may also have a bridge, a piece of wood that the strings pass over and which transmits their vibrations to the instrument's body, which consequently serves as a resonator. Lutes may be plucked or bowed; the word "fiddle" is a general term for bowed lutes. Bows in East Africa are frequently made of sisal fibre, which is rubbed with rosin to increase the friction of the bow on the string. In Malawi, a simple sliver of bamboo, lubricated by spittle, is used. The necks of African lutes do not have fingerboards, as do guitars and violins. The strings are usually stopped by pressure from the backs of the fingers, and not, as on a violin, by pressing the string against the fingerboard. Occasionally the string is pinched between thumb and finger. Most African members of the lute family have only one string.

Several African stringed instruments are not easily classified, being considered lutes by some, zithers or harps by others. One African instrument consists of a body and several necks in the form of thin bows, one for each string. The strings are attached directly to the edge of the body and the tip of each bow. This instrument is called a *pluriarc*, but it can also be classified as a multiple bow harp. Another is the West African *kora*, usually known as a harp lute. Twenty or more strings are stretched in two banks from a neck, over notches in the two sides of a bridge, which stands on the calabash resonator body, and attached to the edge of the body.

The *pluriarc*

The harp family. In harps, the strings are attached to and run vertically away from the soundboard or the membrane that serves as a resonator (in a lute, in contrast, the strings are parallel to the soundboard). All African harps are plucked. The instrument may be held horizontally or vertically. Five to eight strings are usual, tuned by tuning pegs. The African harp is not found with a pillar between the body and the tip of the neck, as on the European harp. For this reason, and because of the curved shape of the neck, it is sometimes called a bow harp. Harps occur mostly in the tropical regions of Central and East Africa; their distribution is limited. In the *ennanga*, a bow harp of the Ganda people of Uganda, a small ring of lizard skin encircles the bow near each tuning peg. This is adjusted to touch the vibrating string lightly so that it emits a buzzing rather than a clear tone. Such a buzzing tone is cultivated in many other African instruments.

The lyre family. The lyre has a body with two extended arms connected at their upper ends by a yoke or crossbar to which the strings are attached. Most African lyres are tuned by friction pads on the yoke. Like tuning pegs, they regulate the string's tension, hence its pitch. Some lyres, such as the *litungu* of the Wanga of Kenya, use a bridge; others such as the *endongo* of the Ganda use none. Like the African harps, lyres are plucked by the fingers of both hands.

Aerophones: wind instruments. In wind instruments, a column or body of air is made to vibrate by a stream of air, such as the player's breath.

Pipes. African pipes are made of reeds, bamboo, or, in some regions, the tip of an antelope horn. The performer blows across the end of the pipe, producing a single note. To produce double notes some players open a hole in the tip of the horn or reed and open and close it with a finger or thumb. Two-tone messages can be sent from one hunter to another without alarming their quarry. Ensembles of end-blown pipes are frequently played together by a number of performers; each is responsible for inserting his one note into the music at the correct moments.

When a number of simple end-blown pipes of different pitch are bound together in a raft they are called panpipes. Panpipes may be played by one performer or in ensembles, as with the single pipes. Among the Nyungwe, of the Zambezi River valley in Mozambique, over 30 panpipe players may take part in the *nyanga* dance, sounding their pipes with alternating sung notes and dancing together in a circle.

Uses of panpipes

Flutes. Flutes with a number of finger holes are found in most regions of Africa. Some flutes are open at each end. These are held vertically and are blown from one end. Such flutes may have a V-shaped notch cut into the upper or mouth end. Transverse flutes (held horizontally like a European orchestral flute) may be closed at both ends or open at the far end only. Flutes are usually played solo because of the great difficulty in making multiple-holed flutes sound in tune with each other; the *ndere* flute ensembles of the Ganda people are an exception.

Reed pipes. Reed pipes use a single or double vibrating reed to set the pipe's air column in motion, as in the clarinet or oboe. They are rare in Africa except in regions where they were introduced by Arabs.

Trumpets and horns. Wherever there are cattle or wild antelope, horns are used as single note noisemakers or, with longer horns, as trumpets. Ensembles of open-ended instruments made of horns, gourds, or a mixture of both materials, are used in royal music in many chief's courts in East Africa. Such horns produce two notes and are employed in the same way as in the pipe ensembles. The *makondere* horn bands of Uganda are well-known examples.

Singing horns, or kazoos. Singing horns are usually made from hollow gourd stems. They have a mouth hole cut in the side and are sometimes fitted with a membrane over the smaller end. As they are sung into, the membrane emits a buzzing sound. Such a buzzing membrane is called a mirliton or kazoo. Singing horns, such as the *malipenga* of the Cewa people of Malawi, are used to give the singing voice a semi-instrumental quality.

Idiophones: hard vibrating instruments. These instruments include objects that are struck, shaken, or otherwise set in motion. They produce either melodic sounds, as with the xylophone, or non-melodic noise, as with rattles, bells, seedpods, clappers, and tubes, or any object that can be made to sound effectively. Three varieties of idiophone are particularly important: the xylophone, the *mbira*, or thumb piano, and the slit drum.

Xylophones. Xylophones are instruments in which several tuned, resonant wooden slats are struck to produce different melodic notes. In Africa, they are found exclusively in the better wooded regions. They vary from simple logs of wood laid across a pair of banana-stalk supports, as among the Ganda, to well-made instruments in which the slats are attached to a frame. Examples of frame xylophones, with a resonator for each note, are those of the Chopi of Mozambique and the Malinke of Guinea. Resonators vary from small tree-fruit shells to broader and longer shapes of cultivated gourds, graduated in size. Xylophones are played with rubber-tipped beaters or with short lengths of stick. They are tuned to the local pentatonic, hexatonic, or heptatonic scale. In some areas, the pentatonic and heptatonic scales used resemble similar scales found in Southeast Asian and Indonesian xylophones, suggesting cultural contact through Indonesian colonists who settled in Madagascar in medieval times. In East Africa and Zaire a single, large, unresonated, free-note (*i.e.*, nonframe) xylophone is played by two, three, or four performers at the same time. Among the Chopi of southern Mozambique, ensembles consist of as many as 40 instruments ranging in size from treble to double bass.

Mbiras. The *mbira* family of instruments is unique to Africa. *Mbiras* consist of an array of metal or fibre reeds attached firmly at one end to a soundboard or resonator. The instrument is usually held between the palms of the hands, and the tips of the reeds are plucked with the thumbs, or thumbs and first fingers. In northeastern Zaire and the Transvaal, the instrument is slung from

the shoulder by a cord and played with the fingertips. The *mbira* is also known as *likembe*, *marimba*, *kalimba*, *chisanji*, thumb piano, and by other regional names. The common term *sansa* is not correct; it is not found in Africa.

Different varieties are set apart by the nature of the array of reeds, their number and tuning, and by the type of resonating body to which the reeds are attached. Rectangular boards or boxes are common, but bell-shaped resonators, rafts of papyrus stalks, small fan-shaped bodies, or even metal bodies are found. Many instruments have, in addition to their soundboard, an external resonator made of a gourd or other material. In a few places larger bass instruments occur; *e.g.*, in Nigeria, Cameroon, and Uganda.

The tuning varies throughout the continent; pentatonic, hexatonic, and heptatonic scales being employed according to local tradition and taste.

Mbira music is highly melodic. The tone preferred by most African musicians is not clear but, rather, edged artificially by devices such as shells, metal rings, or mirlitons placed so that they buzz when the reeds vibrate. *Mbiras* are mostly used to accompany solo singing but are also played in ensembles of three or more.

Slit drums. Hollowed logs, or slit drums, are found in forest regions. They are made by carving a log into shape externally and hollowing it out internally with long chisels through a slit cut in the top of the log. They vary in size from ten or more feet long and four or five feet in diameter, to instruments small enough to be held in the hand and are often carved in the shape of an antelope. By carefully varying the thickness of the two flanks on either side of the drum, two main pitches or tones can be produced. Other tones can be produced by striking other parts of the body. As with membrane drums, wax may be added to the flank of a slit drum to lower its tone. Slit drums are used for signalling by sending two-tone messages and for production of loud, many-tone rhythms. They are among the most versatile of all African percussion instruments, and are often used in combination with membrane drums.

Membranophones: drums. In membranophones, a stretched membrane, usually a cattle, antelope, or waterlizard hide, vibrates to produce sound. These include the whole family of drums. In addition, mirlitons are often considered membranophones, because a vibrating membrane produces their characteristic buzzing sound.

Drums vary in pitch and tone colour according to the weight and shape of the body (as, pot-shaped, hourglass-shaped, cylindrical), the size of the membrane, and the manner of playing. They are played by the hands, sticks, or both. In size they vary from a few inches to six or eight feet. In addition to body size, weight, and shape, drums are classified according to whether they are single or double headed (*i.e.*, with a membrane at one or both ends); whether the body is open or closed; and according to how the membrane is attached. These characteristics affect the quality of the drum's sound. Wax or other material may also be placed in the centre of the membrane to lower its pitch or to produce a falling note rather than a clear-pitched note. A mirliton may be attached to the side of a closed drum; the pulsations of air inside the drum cause the mirliton to vibrate, giving the drum a buzzing tone.

Friction drums are made to vibrate by the friction of the player's wetted or resined hand rubbing on a thin stick, which is pegged to the centre of the drum membrane.

There are three usual ways of attaching the membrane. It may be pegged to the body, usually the case with heavier hides. It may be pinned with multiple small pegs or nails, usually with lighter membranes. Or it may be laced, as with double-headed drums. In West Africa, the lacings on the hourglass-shaped "talking drum" are pressed toward the drum body, tightening the tension of the membranes and thus raising the drum's pitch.

Musical role of drums

Drums are more closely associated by foreigners with African music than any other instrument. This is partly because when drums are played together their volume of sound may draw attention away from the sound of other instruments and voices. Drums in Africa are used mostly for entertainment and to accompany dances. In some instances, specially constructed drums are revered as symbols of authority and national honour and are played only upon ceremonial occasions.

Drums may be played singly, or more usually in small ensembles of two, three, or four. Batteries of drums are played by many communities and notably in several regions of East Africa, Rwanda, Burundi, and Zaire. In Uganda, for example, an ensemble of as many as 15 drums, tuned to a pentatonic scale of more than two octaves, is played by six drummers. Ensembles of up to 20 drums or more are played by the Makonde people of East Africa, and sets of as many as nine or ten wax-tuned drums are played by a single performer in Malawi.

(H.Ty.)

IV. Performing arts

GENERAL CHARACTERISTICS

The dance and drama of Africa are both exciting and original; the dance in particular can be fiercely beautiful or tenderly expressive. But perhaps the most important characteristic of each is its vigour, its life-enhancing quality and purpose. This is entirely appropriate, since each was developed by peoples who, whatever their local differences, were primarily concerned with life, its protection and augmentation, often in a hostile environment. The dance especially can be regarded as an African people's life in action, inextricably linked with its religious and cultural beliefs and illuminating the meaning of both as ties that bind a society together and at the same time give meaning to its structure. Unfortunately, this significance is being lost as the indigenous cultures of Africa give way to foreign influences and as their tightly knit communities break up with the rapid growth of urban mixed communities. African theatre, however, has responded to the stimulus of Western dramatic literature and has added to its richness and variety. In its turn, the West has received African dance forms and styles and readily acknowledges what seems to be a native genius for physical expressiveness and communication.

General observations about the nature of ritual dances performed by primitive peoples apply, by and large, to those that developed early in Africa. The occasions for ritual celebration, at which dancing was an essential part, included the hunt, harvest, birth, puberty, marriage, sickness and recovery, and death (Figure 6). Some dances were "pre-theatrical" and others contained the seeds of drama.

The terms theatrical and dramatic have acquired their fullest significance and technical meanings within the context of Western culture. A theatrical act demands, above all else, a distinction between performer and onlookers, the former displaying some special skill for the delight of the latter; thus an acrobat in the circus engages in a theatrical act. A dramatic act further demands that the performer should imitate an existence other than his own; thus an actor impersonating an acrobat in a play about a circus engages in an act that is both theatrical and dramatic. Many African dances, forming only part of a complete ritual, were pre-theatrical in the sense that they were performed by a section of a community that had developed special skills (of athleticism, grace, endurance), while the rest of the community watched and thereby formed an embryonic audience (Figure 7). The skills, of course, were not primarily acquired for purposes of display nor to give pleasure to the onlookers but to enhance the potency of the ritual; as one scholar points out, however, the nature of dances that require unusual skill in performance encourages an aesthetic desire to perform well, in addition to any other impulse. Some dances were semi-dramatic inasmuch as they employed mimesis, or imitation, the essential element of drama. Another scholar describes, for example, a rainmaking ritual in East Africa, in which this element of imitation is readily apparent: after water had been sprinkled on the ground from a new gourd bowl, the rain dance followed, "imitating with soft pattering feet the first slow thunder-

Pre-theatrical aspect of African dances

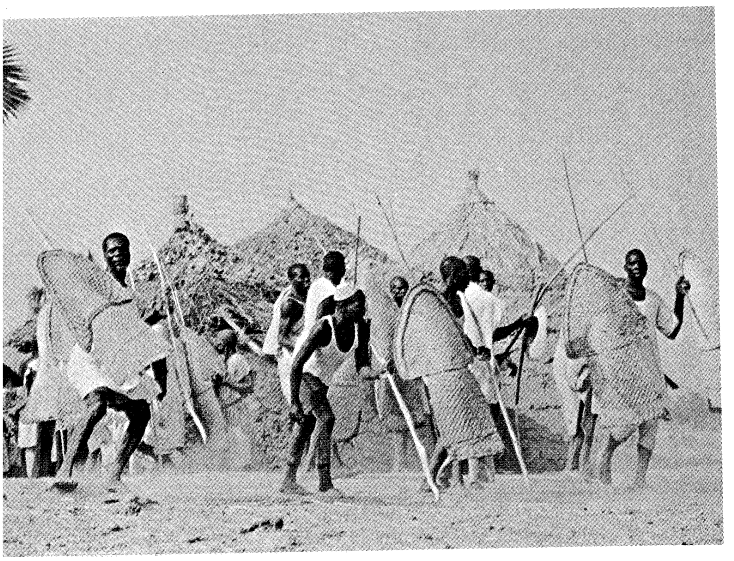




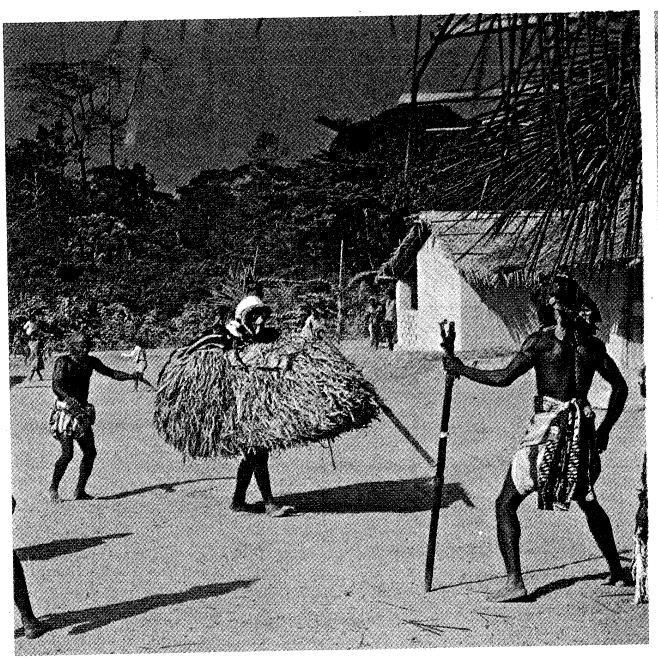
Figure 6: *Dances for ritual celebration.*
(Left) Mousgoum tribesmen staging a mock war dance during a funeral ceremony, Logone River, Cameroon. (Right) Unmarried Kao-Nyaro girls dancing before men of their village prior to a ceremonial tribal contest, The Sudan.
(Left) Marc & Evelyne Bernheim—Woodfin Camp, (right) George Rodger—Magnum

drops, then more swiftly, to the swirling of the women's skirts, the rapid falling of steady rain." In the hope of ensuring a successful hunt—for on its outcome could depend a community's food supply—dances were evolved in which the animals to be hunted were impersonated by dancers, often wearing skins and specially constructed masks to heighten the illusion; they were "captured" and "killed" by the hunters. The logic of this stemmed from a hope that, in the words of a modern scholar, "the desired goal, through being dramatized, becomes more readily accessible, almost as if once a pictorial hint is given to nature it might be hard to erase it."

Historians of the drama are strangely silent about the theatre in Africa before the 20th century, and it might easily be supposed that none worthy of the name existed independent of the dance. This was by no means the case. In parts of Africa a theatre developed that was exactly suited to the people's social and emotional needs. The dance, meanwhile, did not depend entirely upon religious inspiration; the natural urge to express joy and emotion was also satisfied.

AFRICAN DANCE

The kinds of dance developed by African peoples for performance on important ceremonial occasions range from the ecstatic, convulsive frenzy of the medicine, or shamanistic, dance to the subdued discipline (at any rate among certain peoples) of the funeral dance. The basic patterns followed in any and all of these dances might be either abstract or mimetic.



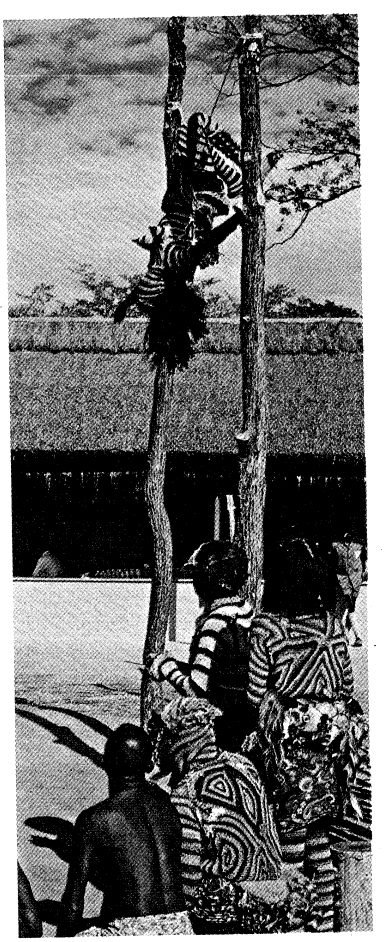


Figure 7: *Theatrical aspects of African dance.*
(Left) Fetish dancers of the Guere Wobe tribe performing during a funeral dance, Petit Duekoue, Ivory Coast. The dancers serve as "intermediaries" between the masked figure and the men of the village. (Right) Tribesmen of the Matable village performing an acrobatic ceremonial dance, Victoria Falls, Rhodesia.
(Left) Marc & Evelyne Bernheim—Woodfin Camp, (right) David Martindale—FPG

Abstract and mimetic patterns of the dance. *Abstract patterns.* The ceremonial mourning dance performed at the funeral of a member of the Akbia tribe (in the southern Sudan) provides an example of a dance of abstract pattern. At dawn, 30 women of the tribe, wearing garments made from natural forest creeper and plastered with mud, move in a single file around a mango tree. Their dance step is a simple, jerking, forward shuffle. The accompanying musicians repeat an unresolved phrase of five descending notes. Misery is expressed not by any concrete symbol or image but by the monotony of rhythm and movement: the women's concentration expresses mourning. In their long dance, which lasts for seven days and nights, they are also thought to be accompanying the deceased along the first, difficult part of his transitional journey into the world of spirits. The dance is thus full of meaning; it is a distillation of tribal belief, and it expresses a definite mood.

Dances performed for sheer joy

At this point, however, it should be stressed that dances were not always motivated by religious belief but were sometimes performed for sheer joy; such dances were abstract. As one dance historian explains it:

> Many . . . dances derive simply from social-expressive urges, the collective desire to move, to exercise rhythmically, to entertain oneself. Thus, after an evening meal, a Mambuti Pygmy of Central Africa may casually strike a drum, simultaneously hopping in place from one foot to the other, and eventually entice the entire camp to join him in an all-night affair. . . . Such dances are not meant to portray or depict anything, whether concrete or other-worldly. They are quite abstract in idea and intent, meant primarily for kinesthetic enjoyment.

Mimetic patterns. Those dances in which there is implicit some kind of "wish"—perhaps for fertility, for success in the hunt, for rain, for a good harvest—commonly contain mimetic patterns that match their generally symbolic theme, as in the rain dance described above. A dance performed during the six-day Incwala festival (Swaziland, southern Africa) may stand as a further example. During the festival, a number of things are celebrated—the new year, the chief's authority, a renewal of tribal energy—and one dance is performed for the purpose of "renewing the chief's strength." His warriors dance first, while they throw a new green gourd high into the air, from one to the other. At a certain point the chief joins in the dance, demonstrating his skill and strength by throwing the gourd ever higher and higher. The gourd is regarded as the repository of energy that will pass into the chief; therefore, the gourd must not be dropped, because, should this happen, the vital energy will be drawn into the ground instead of being transferred into the chief's person.

Variety of dance styles. Cultural temperament determines to some extent the dancing style that a given people develops. The outgoing peoples of Central Africa, for example, according to one scholar

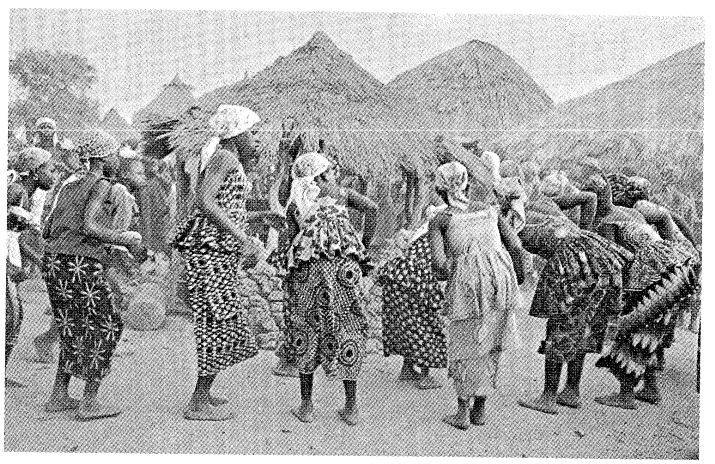
> exhibit the explosive, enlarged bodily movements so markedly missing from the Far Eastern dance world. The flinging up and down, the centrifugal leaping and kicking that bespeak an "outgoing" extrovert cultural temper appears definitely to belong to the dance-forms of Central Africa.

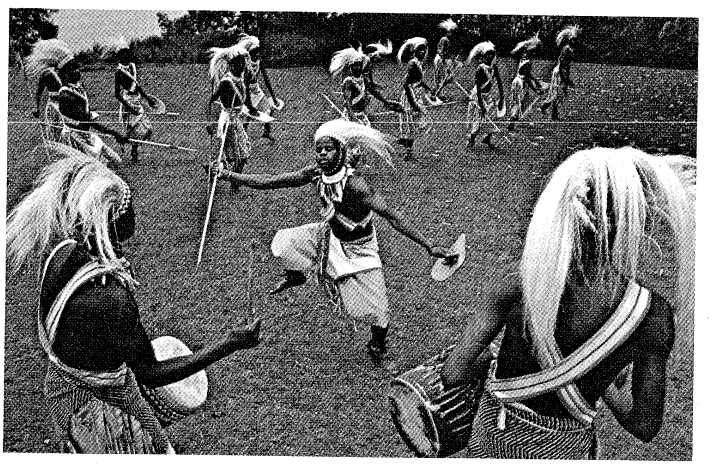
The same characteristic is found in extroverted peoples of other parts of the continent, such as the Zulus of the south. The reverse also holds good: the same scholar describes the Pygmy people of the Central African region, "particularly the branch or branches following nomadic life-patterns in the heart of the forest . . . shy, subdued, retiring . . . song, dance, and drumming during feasts . . . often very soft." (The acrobatic dancing of other Pygmy groups, which are far from "shy, subdued, retiring," impressed observers as long ago as the 7th millennium BC, when some of their number were brought to Egypt for the purpose of dancing in religious processions.)

The range of African dancing styles is very wide indeed (Figure 8); the same kind of dance will be danced very differently from one tribe to another, and even neighbouring communities will evolve an idiosyncratic style. Equally numerous are the kinds of bodily movement that give a dance its overall character; in southern Africa, for example, there are stamping dances (of the Nguni), shaking dances (Xhosa), striding dances (Sotho), step dances (Tswa, Zulu, Sotho), and acrobatic dances (Tswa, Ndau).

Kinds of bodily movement

The instruments used to provide musical accompaniment are everywhere chiefly those that mark rhythm





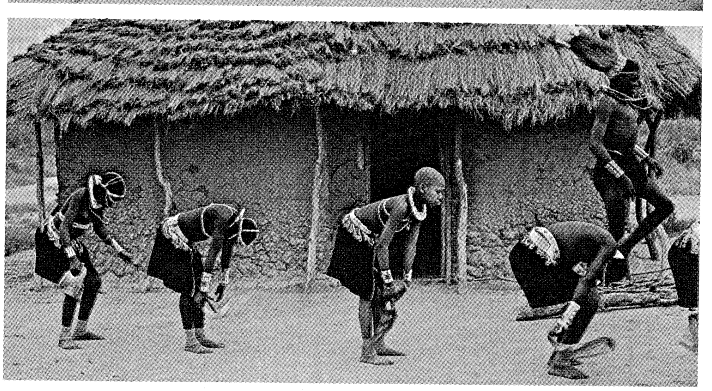


Figure 8: *Styles of African dance.*
(Left) Women of the Tiongofoloha village performing the slow, rhythmic movements of a funeral dance, Ivory Coast. (Right) The fast, agile movements of the Intore (Watusi) people during the ceremonial lion dance, Lake Kivu, Zaire. The headdress is symbolic of a lion's mane. (Bottom) Wakamba boys performing a type of "leapfrog" dance, Makindu, Kenya.
(Left) Marc & Evelyne Bernheim—Woodfin Camp, (right) George Holton—Photo Researchers, Inc., (bottom) Bill Horman—FPG

(above all drums and gourd rattles), though pipes are also used. Rhythm is rather frequently reinforced by clapping, while the human voice often cries out cadences that intensify whatever particular emotion is being expressed in the dance itself.

African dance in the 20th century. As the African peoples lose their cultural identity to superimposed cultures, the web that gives their indigenous dance coherence gradually loosens. Their themes have been handed down from generation to generation, and a break in continuity means that many are forgotten; others continue to be performed, but, as the culture that gave meaning to their original statement is eroded, the dances lose an inner life and significance that nothing can fully restore. Today, although traditional dances are performed, they are as often informed by nostalgia as by dynamic religious belief.

AFRICAN DRAMA

Over and above the dramatic patterns of the dance, exchanges of dialogue in heightened speech eventually began to appear in religious ritual ceremonies. In West Africa, for instance, where the communities of the Limba people supported a moon cult, dialogue appeared in a solemn ceremony held at the time of full moon. Before the moon dance was performed, the sacrificing priest begged the "moon god" (impersonated by the cult's chief priest, masked and robed) to send prosperity in the coming month. The "god" was then given drink. The dance that followed was performed by two groups of 14 dancers, representing the two phases of the moon. Toward the end of the month another ceremony was held, at which the chief priest appeared as the "dying god" who told his followers not to mourn because he would rise again to new life after three days and three nights.

Despite this theatrical instinct for dance dramas, there was little need in Africa for the development of an independent "religious" theatre comparable to Western tragedy. In an interesting study of religious processes among the Ndembu people of Zambia, in Central Africa, a modern scholar analyzes the dramatic structure of the society itself, which is such that it might be said to supply the cathartic effect of formal tragic action. The study suggests that in any crisis involving either an individual belonging to the community or factions within it, when feelings of anxiety, fear, and aggression are aroused, the people have recourse to "ritual remedy" (which involves the recounting of traditional causes of trouble and their associated feelings), thereby restoring amity and cooperativeness and regaining confidence in their capacity to cope with the present problem.

Theatre comparable to Western comedy

Development of popular theatre. There developed nonetheless something very like the Western concept of comedy as a vehicle for social criticism and recreation. Among its early occasions were the purification rites held annually to "cleanse" the people of any and every evil. One ethnologist has likened some of these purification rites to the Roman Saturnalia, festivals characterized by excess. Another has mentioned a freedom of speech at some purification rites that parallels the medieval "Feast of Fools" in the West, at which the established order was overturned and the "king" waited on his "servant." Even kings, chiefs, and gods were mocked:

> O King, you are a fool.
> We are taking the victory out of your hands.
> O King, you are impotent.
> We are taking the victory out of your hands.

The psychology behind this topsy-turvy situation is that, by mockery, the community "soul" would be "cool and quieted."

The instinct for mockery survived. A vigorous community theatre evolved and still survives in many rural parts of Africa. Performances celebrate seasonal festivals; everyone is involved, joining in well-known narrative expositions (which often recount the deeds of past kings and heroes) by forming processions, by chanting, singing, miming, and by responding in chorus. The themes and presentation are handed down orally from generation to generation and are known to all members of the community. There is no impatience for a "plot twist"; every moment is savoured as it arrives.

In the hands of secret associations of artisans—somewhat similar to trade guilds of craftsmen in medieval Europe—an almost professional African theatre developed. The entertainment devised was akin in material and versatile showmanship to that given by a commedia dell'arte troupe of popular players in Renaissance Europe. A modern scholar has described a performance given by the semi-religious Ekong association of Ikot Obong village in the Calabar province of Nigeria, a community of the Anang people. This particular association rehearses weekly for six years, and the entertainment devised during rehearsal is presented publicly in the village clearing during the dry season of the seventh year, afterward touring to surrounding villages and communities.

> During the six years of practice, a complicated seven hour routine was originated and mastered to perfection by the members. This routine included: music played by an eleven man percussion orchestra, featuring characteristic [*Ekong*] rhythms; patterns danced by a chorus of eighteen men; songs of ridicule sung by the orchestra and dance chorus and masked soloists; plays enacted by puppets and an entourage of fourteen men, some wearing masks; dances on stilts by two men; and the mystifying movements of a life-size wooden carving of a female fertility spirit enwrapped by the coils of a python. I was told that other [*Ekong*] groups present additional specialities, such as acrobatics, sleight-of-hand magic, and tight rope walking, by skilled entertainers.

The entertainment is carefully fashioned to be as relevant as possible to the neighbourhood. Detailed information about the private lives of certain individuals from a chosen community is secretly obtained, and these people are singled out for special ridicule. There is also more generalized satire, especially in the skits, and, of the five watched by the scholar, "two ridiculed nativist churches, and the other three corrupt diviners, domineering wives, and adulterous wives who run off with their lovers." Audience participation is expected, and the onlookers will come into the playing area and show their approval of any particularly excellent actor.

Generalized satire

Development of modern theatre. Even this "professionally" sponsored theatre declined with the intrusion of Western cultures. For example, whereas before the coming of the British in 1901, the Afaha subtribe of the Anang people had 28 Ekong organizations, 50 years later they had only two. As one modern scholar puts it:

> the concentration of population in urban areas, with the influence of the cinema and other vehicles of non-African culture, is creating a demand for ready-made leisure entertainment in place of the communally created functional drama. . . . So side by side with traditional drama is being formed an African theatre which exploits the ideas and techniques of the European stage.

African writers have adapted Western plays into African languages and have written original plays, and dramatists such as Wole Soyinka and Athol Fugard have made their mark in Western countries—as did Pearl Primus and her dancers in 1952, the Ballets Africains of Keita Fodeba in 1956 and 1957, the Negro jazz opera *King Kong* in 1961, and a Zulu adaptation of Shakespeare's *Macbeth* in 1971. Those associated with African theatre are aware that they have much to teach their colleagues in the West, who, beginning in the 1960s, became preoccupied with the problem of escaping the restrictions imposed by its "literary" drama and the respectable but often dull theatre it had produced. In the words of the Nigerian playwright Wole Soyinka,

> African drama is sophisticated in idiom. Our forms of theatre are quite different from literary drama. We use spontaneous dialogue, folk music, simple stories, and relevant dances to express what we mean. Our theatre uses stylised forms as its basic accepted disciplines. I am trying to integrate these forms into the drama of the English language.

(Si.M.)

V. Visual arts

The visual arts—sculpture, painting, and architecture—consist of more or less enduring forms in contrast to music, dance, and oral literature, which are ephemeral.

Once made, the products of the enduring arts continue to exist independently of their creators and their cultural context. Sculpture, thus, can be taken away and studied outside Africa. It is for this reason that "African art" has come, in the West, to mean almost exclusively sculpture. But, while it is true that sculpture has been Africa's greatest contribution to man's cultural heritage, the other visual arts—painting, architecture, and the decorative arts—have also contributed to it.

TYPES OF AFRICAN VISUAL ART

Decorative arts. The study of African art might well begin with the human body itself, for the body has served as a support for painting (especially the bodies of dancers in ceremonies accompanying passage rites marking birth, initiation into adult status, marriage, and death) and for sculpture in the forms of hairdressings and scarifications, both of which are often elaborated into designs of great beauty and are represented, in turn, on sculptures of human figures. Indeed in Africa, the scarifier and the hairdresser are regarded as artists, and, among the Guro, the hairdresser enjoys considerable prestige.

Pottery is made in agricultural communities wherever there is suitable clay and sufficient wood for the open fire in which the pots are fired. Decoration by incision is most widely practiced; carved roulettes are also used. Painting is not common south of the Sahara, although entire pots are often washed all over with a pigment.

Basketry is made by either the simple interweaving of elements or by the coiling of multiple strands, around which is wrapped a flat element stitched through the next inner turn. In general, coiled basketry is characteristic of the savanna areas, and woven baskets are common in the forests. Mats and containers are made by both techniques.

The weaving of raffia fibres was practiced in antiquity and reached its artistic peak in the pile cloths of the Kuba of Zaire. This art is widespread in agricultural communities, usually employing cotton, although some of the finest weaving, such as the *kente* cloths of the Ashanti of Ghana, also employs imported silk and similar threads. Women usually weave wide cloths on a vertical loom, but in West Africa men weave on a horizontal loom, producing long strips only two to four inches wide. In West Africa, one or more techniques of resist dyeing are used on imported cloth. The dye is resisted by applying a pattern of starch either with a stamp, as in the *adinkra* cloth of the Ashanti, or through a metal stencil or painted freehand, both of which are used in Yoruba *adire*. In Benin, Dahomey, and Ashanti, appliqué decoration is also found.

Leatherwork is especially important in the savannas, where pastoralism is practiced. It has reached a high standard among the Hausa and the northern Yoruba, both of whom make pouffes for export. In East Africa, many peoples not renowned for sculpture are skillful leatherworkers (*e.g.*, the Masai) and often paint the surfaces of their leather shields.

Jewelry is made in gold and silver by lost-wax casting and by smithing. Imported glass beads have been used for centuries in most of Africa, as necklaces and bangles, as decorations on leather garments, and applied to satchels and calabashes.

Architecture. A variety of materials is used in African architecture. Buildings made entirely of stone are relatively rare, but, in some cases, clay buildings have been carefully maintained and preserved in their original form for several centuries. People of simple technology (*e.g.*, the Bushmen) provide shelters from twigs; Pygmies in the forests build simple beehive huts thatched with leaves. Round houses characterize the dwellers in The Sudan and the grasslands of East Africa (where the Masai, however, construct rectangular buildings), while in the Congo Basin both round and rectangular houses are found. In the forests, houses of a basically rectangular form are usual, elaborated in accord with local custom, the plastic qualities of clay being exploited not only to decorate the surface but sometimes with the effect of turning the building itself into a kind of sculpture. Western materials, such as cement blocks and corrugated metal or asbestos sheeting, are encouraging a swing from round to rectangular forms, which are more easily built from these materials. In the larger cities, skyscrapers, conference halls, and other buildings have been built similar to those pioneered in the West but sometimes affording scope to African artists in requiring carved doors or mural decoration in paint or mosaic.

Painting. The earliest surviving African art consists of painting and engraving found in rock shelters of the Sahara from at least as early as the beginning of the 5th millennium BC and continuing into the 20th century there and sporadically elsewhere in Africa as a minor form of artistic expression wherever suitable smooth rock surfaces are available. Other forms of painting that have received less attention include the above-mentioned body painting and the painting of house walls, utilitarian objects (Figure 9), and sculptures. In the case of masks, the painting may, as among the Yaka, enhance the sculptural forms; but it may well completely ignore them, as among the Kalabari Ijo (Ijaw). Apparently, for the Kalabari Ijo, the act of painting is of greater importance than that of carving because, although the sculpture provides a home for the spirit, it is the act of painting that brings the spirit to reside in the mask.

Rock painting and engraving

Sculpture. *Mediums.* Although wood is the principal medium of African sculpture, many other mediums are employed: copper alloys, iron, ivory, terra-cotta, unfired clay, and, infrequently, stone (Figure 10). Unfired clay is and probably always was the most widely used medium in the whole continent, but, partly because it is so fragile and therefore difficult to collect, it has been largely ignored in the literature. Small figurines of fired clay were excavated in a mound at Daima near Lake Chad in levels dating from the 5th century BC or earlier, while others were found at Zimbabwe in deposits of the later part of the 1st millennium AD. These imply an even earlier stage of unfired clay modelling. About the time of these lower levels at Daima (which represent a Late Stone Age pastoral economy), there was flourishing farther to the west

By courtesy of the Musee de l'Homme, Paris



Figure 9: Dogon painted stone circumcision stool, Mali. In the Musée de l'Homme, Paris. Length 35 cm.



Figure 10: *Mediums of African sculpture.*
(Left) Wood doors carved by the Yoruba sculptor Olawe of Ise, *c.* 1910. In the British Museum. Height 1.83 m. (Top right) Luba ivory neck rest, Luba cultural area, Congo. In the Charles Ratton Collection, Paris. Height 17.8 cm. (Bottom right) Fon iron image of Gun, the god of iron and war, Dahomey. In the Musée de l'Homme, Paris. Height 1.65 m.
By courtesy of (left) the trustees of the British Museum, (bottom right) the Musee de l'Homme, Paris; photograph, (top right) Holle Bildarchiv, Baden-Baden, West Germany

the fully Iron Age Nok culture, producing large, hollow sculptures in well-fired pottery, some of the stylistic features of which imply yet earlier prototypes in wood.

Copper-alloy castings afford evidence of very great sculptural achievements from as early as the 9th century AD, when the smiths of Igbo Ukwu were casting lead-bronze, which is highly ductile, and smithing copper, which is not. Some three or four centuries later the smiths of Ife, seemingly unaware that unalloyed copper was not suitable for casting (or perhaps wishing to demonstrate their virtuosity), used it to produce masterpieces such as the seated figure in a shrine at Tada (see below *Nupe*) and the so-called Obalufon mask in the Ife Museum. In fact, however, zinc brasses were used more than unalloyed copper. The largest corpus of this work is from Benin, where zinc brasses were used almost exclusively, the zinc content increasing with time. These copper-alloy castings, together with terra-cotta sculptures the traceable history of which goes back even further, are the main evidence for the early history of sculpture in sub-Saharan Africa.

That iron should be used as a medium of sculpture may appear unusual; but in many parts of Africa wood carvings are made by blacksmiths. Among the Bambara, the blacksmiths work in both mediums, in totally different styles; the Yoruba blacksmiths, however, work solely with metal, making skillfully sculptured staffs in wrought iron. It has been suggested that the abstraction found in many sculptural styles, especially in the western Sudan, may be due to their having been made by blacksmiths, the forms of forged iron objects being carried over into wood sculpture; but this theory carries little conviction.

Stone sculpture

Stone sculpture occurs in several separate centres, employing both hard and soft rock, but there is usually not much evidence of a development through time in a single place. Ivory is a highly prized medium in many parts of Africa. Its fine texture makes it suitable for delicate sculpture, while its rarity leads to its employment in many societies for items of great prestige.

Tools. African wood sculptures are carved with similar tools throughout the continent. An ax may be used to fell the tree; but an adz, with its cutting edge at right angles to the shaft, is used for the substantive work of carving. The skill achieved with this tool is astonishing to the Western observer. Thin shavings can be removed with speed and accuracy, creating a surface (especially when the form is convex) that shows slight facets that catch the light and add to the visual interest. More intricate work is done with knives. A pointed iron rod heated in the fire may be employed to bore holes in a mask for attachment to the costume and to permit the wearer to see. The surface of the sculpture is sometimes polished with the side of a knife or sanded down with rough leaves. Details are commonly picked out by charring with a red-hot knife (as among the Ibibio), or the carving is immersed in mud to darken its surface before oiling (as among the Dan group).

It has been suggested that native-made iron was of poor



Figure 11: Wood figures in the pole style, southwestern Ethiopia. In the Ethnographical Collection, Zürich. Height (left) 47 cm, (right) 41.9 cm. The woven cloth is also of Ethiopian origin.
Holle Bildarchiv, Baden-Baden, West Germany

quality, barely adequate for the task of carving any but the softest woods and that it was this technical inadequacy that led to cubistic qualities in African sculpture. Native wrought iron, however, is capable of accepting a very fine cutting edge, so it would appear that the qualities of African sculpture do not result from technical inadequacy but rather from the artist's intention.

Style. Attempts have been made to define a "primary" style of sculpture in Africa—a style which, although in contemporary use, is thought to have survived from the earliest period of African wood sculpture. Usually this "primary" style is equated with pole sculpture (Figure 11), in which a minimum amount of wood is removed from a piece of tree trunk to convey an impression of a human figure, with facial features sketched in and arms represented in low relief against the trunk. This rather abstract style is contrasted with "round" sculpture, which, because it is more naturalistic in its representations, is thought to have originated later. But, like other polarities that have been suggested, such as "concave" and "convex," both extremes of any pair of opposites may be shown to occur simultaneously in many groups. Artistic style is perhaps better seen as a continuum between the two extremes of abstraction and naturalism, neither of which has a valid claim to priority in time; for both styles occur even within the corpus of Nok sculpture, the earliest surviving in sub-Saharan Africa.

Sculpture and religion. Certain striking characteristics of African sculpture cannot be explained without knowledge of African religious beliefs. For example, the human navel is often exaggerated, especially in figures of ancestors such as those of the Lulua (Bena Lulua). This appears to be because it marks the link between one generation and the next. All African religions emphasize the continuity of life between the dead (ancestors), the living, and those not yet born. It is probably for this reason, rather than because of a preoccupation with ideas of fertility as such, that ancestors are commonly represented with emphasized sexual organs as well as navels.

Belief in the continuity of life

Another general notion is that the representation of an object shares in the life-force of the object itself, *e.g.*, the representation of an animal brings something of the characteristic qualities of that animal to bear on the situation. This perhaps explains why African sculptures have been referred to as idols, even though they are not strictly objects of worship. The mystical participation of the object in its image—which also lies at the base of some European witchcraft practices—enables the African to control the forces represented. He may harness the life-forces of many animals to produce a concentration of life-force more powerful than the one he fears (as, for example, the Dogon in driving away to their final abode the spirits of the ancestors); or he may, by providing an image of an ancestor (as, for example, in the ancestor "stools" of the Kalabari Ijo), furnish his spirit with a place to live now that it is no longer confined to a human body. For the living must know where the spirit is so that they can avoid offending it and can keep it happy by offering it food.

SUDANIC CULTURES

Types of art. *Painting and engraving.* The earliest known African art consists of more than 30,000 engravings and paintings on rocks in the Sahara. At the time most of these works were executed (from about the beginning of the 5th millennium BC into the 2nd) the area was open savanna, comparable to the Sudanic belt of the present day, supporting animals no longer found in the desert but represented in the art. Representation of the changing fauna makes it possible to divide the art into a succession of periods, the divisions being confirmed by changes in style and in the economy and artifacts possessed by the artists. In later periods the art spread into the western Sudan, across to the Horn of Africa and, possibly, into southern Africa (see below *Southern African cultures; Southern Bantu; Bushmen*).

The earliest engravings (in southern Oran, the Tassili, and Fezzan) reflect a hunting economy and represent such wild animals as the extinct buffalo *Homoioceras antiquus* (formerly called *Bubalus*, hence the name Bubalus period assigned to these earliest engravings), the elephant, rhinoceros, hippopotamus, giraffe, ostrich, and large antelope. The human figures are armed with clubs, throwing-sticks, axes, and bows. Frequently very large (ranging from a few inches to more than 20 feet), the drawings are fairly naturalistic representations (Figure 12).

Radioactive-carbon dating indicates that there was human occupation (if not drawing) in the Tassili from the mid-6th millennium BC. Paintings in the Tassili seem to begin a little later than the engravings, for *Homoioceras* is not represented, though elephant, ostrich, and giraffe are. One scholar has distinguished upward of 30 styles of paintings, the earliest assigned to the Archaic period, which seems to correspond to the later part of the Bubalus period of the engravings. Because these paintings often represent men and women with globular heads or wearing masks reminiscent, in a general way, of those used in black Africa at the present time, it has been inferred that the artists were black.

Cattle period

There follows, both in painting and in engraving, the Cattle period. *Homoioceras* is no longer represented, though other wild animals are, and the appearance of domestic cattle indicates that pastoralism had by then become the basis for human life. The bow is the principal weapon. Bones of domestic cattle and of *Homoioceras* were found together in a deposit dated by radioactive carbon to the mid-4th millennium BC, thus dating this transitional phase of the art. The style of engraving is less naturalistic, the poses are stiffer; in contrast, the paintings are more naturalistic, with compositions and a sense of space, if not strictly of perspective. The Cattle period ended with the introduction of the horse about 1200 BC.

The Horse period is subdivided into three periods. The first is the Chariot sub-period, in which the elephant was the only pachyderm still depicted, cattle continued to be represented, and mouflons, or wild sheep, and domesticated dogs appeared. The earliest chariots were carefully rendered with a single shaft and a horse on each side; later chariots consist only of a shaft with two wheels, and human figures are reduced to two isosceles triangles set apex to apex. Spears and shields are introduced and, later, daggers. The Horseman sub-period reflects a change from horse driving to horse riding, though chari-





Figure 12: *Rock painting and engraving in the Sahara.* (Top) Painting of herdsmen and cattle, Sefar, Tassili, Algeria, Cattle period. 3.05 m × 1.05 m. (Left) Engraving of an elephant, Bardai, Chad, Babulus or Cattle period. Height 20.3 cm.
(Top) J.D. Lajoux, (bottom) Emil Schulthess—Black Star

ots continue. Next, the camel was introduced, possibly as early as 700 BC and certainly by Roman times, producing the Horse and Camel sub-period. Cattle had become very rare, and the drawings were reduced to between seven and 16 inches (17 to 40 centimetres) in length, whereas in the earlier sub-periods they were between 10 and 20 inches (25 to 50 centimetres) in length, and sometimes as long as three feet.

Because continuing desiccation led to restricted distribution of the horse (represented mainly in Mauritania), the Camel period reflects only present-day fauna: camel, antelope, oryx, gazelle, mouflon, ostrich, humped cattle, and goat. At first the spear was the only weapon depicted, but later the sword and firearms were added, weapons that are still in use. The style is highly schematic, and sizes of the drawings range between six and eight inches (15 to 20 centimetres). This period has continued up to the present time, for their owners still paint and engrave representations of camels in the Sahara.

Sculpture of the Nok culture

Sculpture. The earliest known sculpture of large size in the Sudan is that produced in terra-cotta by the Nok culture, which flourished extensively in northern Nigeria from the 5th century BC into the early centuries AD (Figure 13). These people were the first-known manufacturers of iron in western Africa, furnaces at Taruga having been dated between the 5th and early 3rd centuries BC; they continued, however, to use stone tools. Of well-fired clay, their sculptures represent animals naturalistically; human figures, however, are depicted with heads that are usually tubular, but sometimes conical or spherical, and with simple tubular trunks and limbs. Although generally found in fragments, it would appear that the human figures often must have been four to five feet (120 to 150 centimetres) in height. The sculptural form sometimes presents features that appear to have been carried over from wood carving, probably implying the existence of earlier sculptures in that less durable material. The heads seem to have been large in proportion to the rest of the body, a feature which, like the forms of eye and nose and many other details, has parallels in recent sculpture in many parts of West Africa (see below *West African cultures*). Thus, the art of Nok indicates the antiquity of many basic canons of West African sculpture, but the precise relationship between ancient and modern forms is obscure; for the modern Sudanic forms tend to be somewhat skeletal, in contrast to the fuller, more Nok-like forms of the Guinea Coast.

Frank Willett



Figure 13: Nok terra-cotta head, Nigeria. In the Jos Museum, Nigeria. Height 21 cm.

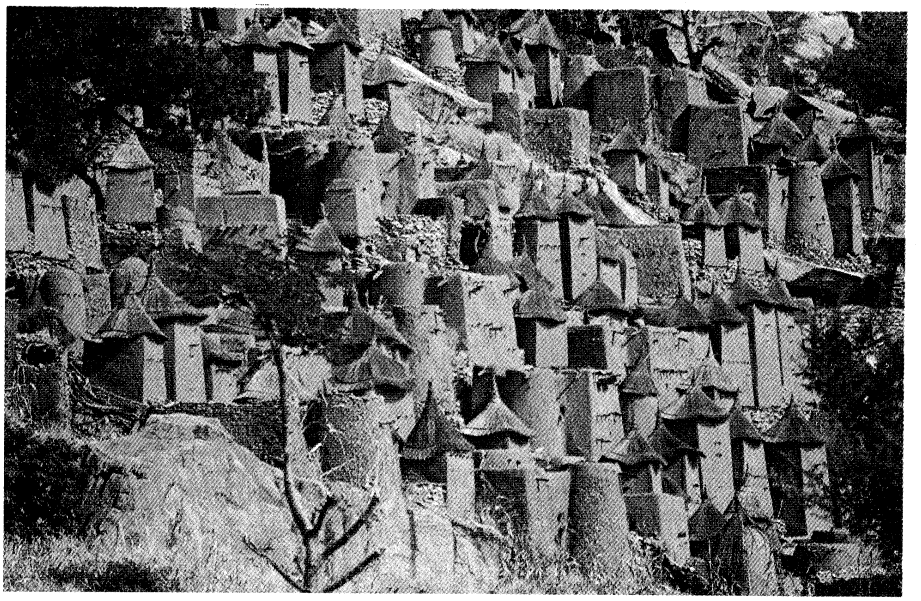


Figure 14: Dogon village, Mali.
Victor Englebert—EB Inc.

Not far from the Nok area, at Daima near Lake Chad, small, simple clay animal figures were by the 5th century BC being made by a population of Late Stone Age herdsmen. A little later they began to extend the legs of the animals, and some time after AD 1000 they began to make animals covered with little spikes. The last are similar to examples found on sites of the Sao culture in the Chari Valley, where more elaborate human figure sculptures, thought to represent ancestors and probably spirits, have also been found. Radioactive-carbon dates for these range from the 5th century BC to the 18th century AD.

Many other sculptures in terra-cotta of uncertain age have been found farther west; for example, around Jenne and Mopti. They may have some association with the medieval empires of Ghana and Mali, which, for all their extensive trade contacts across the Sahara, did not significantly change the basic structure of society in the western Sudan. Cults of the earth and of agriculture continued unimpeded, so that many of the sculptures used in these cults today probably represent ancient traditions.

The oldest wood sculptures to survive (three have been dated by radioactive carbon to the 15th to 17th century AD) are from caves in the Bandiagara escarpment, south of the bend of the Niger River in the area now occupied by the Dogon (see below), who attribute them to an earlier population, the Tellem. These figures, usually of simplified and elongated form, often with hands raised, seem to be the prototype of the ancestor figures that the Dogon carve on the doors and locks of their houses and granaries; but investigations have confirmed that the Tellem are ethnically a different people from the Dogon, so the art style appears to have been handed on from one people to another.

The hut as a basic architectural unit

Architecture. The traditional architecture of The Sudan consists of a basic unit, the cylindrical mud hut with a conical thatched roof, which may stand completely alone, as among the Tiv, or be grouped into compounds with separate huts connected by a wall, as among the Dagomba. The walls of the individual huts, commonly decorated by painting or low-relief decoration, are often built upon a ring of stones, which may be all that survives of a deserted settlement. This form has been modified in a variety of ways. The Jaba (Ham) have made the plan oval and built transverse walls inside with large cavities that serve as granaries. The Masa (Musgu) continue the walls upward and inward, forming a single curve of wall and roof.

With Islām, which began to spread among the Negro peoples of The Sudan in the 11th century AD, came the mosque. There were already 12 in the city of Ghana in 1067–68. In the early 14th century Mansa Mūsā, ruler of Mali, brought as-Saheli from Egypt to design mosques at Gao and Timbuktu, probably modifying the established tradition. He is credited with the introduction of the fired brick. Accompanying Islām came a square house unit with a flat mud roof supported on beams, the weight of which usually required a different technique of building —sometimes using pear-shaped, dried-mud bricks, as in Kano, and everywhere requiring buttressing to counter the outward thrust. These houses are usually incorporated into family compounds by means of an encircling wall. Many communities that are not Muslim, such as the LoWiili, also build square houses, and it is a moot point whether or not this shape represents Muslim influence. The Dogon build square, flat-roofed, multi-story houses and tall, circular granaries, as if they had drawn on both traditions (Figure 14). The Nupe, however, though effectively Muslim, live in compounds composed of circular huts, which they sometimes pave with circular pottery tiles broken with a mallet to produce a flat, hard-wearing surface. At Daima, at the end of the 1st millennium AD, circular huts were paved with potsherds set on edge, a technique otherwise mainly found among the Yoruba (where it seems to be as old) and at Benin.

Decorative arts. Leatherwork is highly developed—so-called Morocco leather comes from northern Nigerian sheep and goats. Horse trappings, saddlebags, pouffes, and satchels are made, while leather decorations are applied to basketry hats worn by the pastoralists of the area. Baskets are usually made by the coiling technique and are often lidded. Pottery, made among settled communities, may be decorated by incision and rouletting (*e.g.*, the Hausa). Calabashes are widely used and may be incised or decorated with pokerwork. Silver jewelry is cast and then chased.

Dogon. Dogon sculpture is intimately linked with the cult of the ancestors. Figures are made to house the spirits of the deceased on the family shrine, and masks are used to drive the spirits away at the end of the mourning period. One type of mask, called *sirige*, has a very tall, flat projection above the face (a feature found also in the masks of the Mossi and Bobo), which is said to represent a multi-story house. The Great Mask, never worn and made anew every 60 years, represents the primordial ancestor who met death while he was in the form of a serpent. Paintings on rocks and on stone slabs used for circumcision seem to depict masks. Iron staffs topped with human figures are also made, and some personal ornaments are cast in bronze.

Bozo. Sometimes regarded as the earliest inhabitants of the region, the Bozo fishermen of the Niger Bend sculpt large bull masks, which apparently are carried on poles.

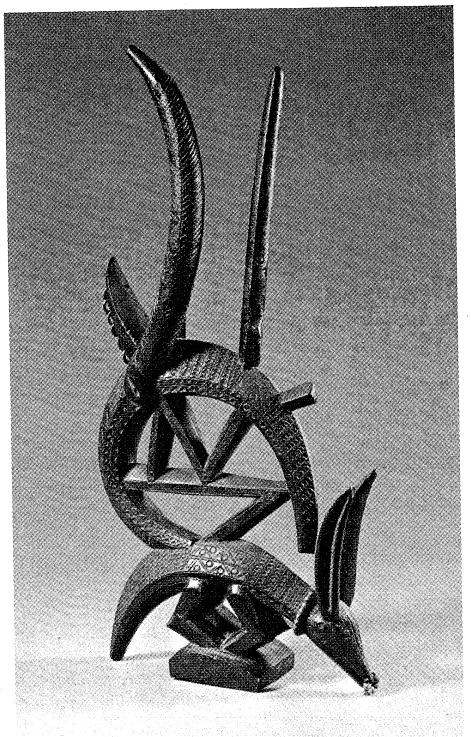


Figure 15: Bambara dance headdress in the form of an antelope, representing the spirit of Tyiwara, who introduced agriculture, wood, Mali. These headdresses, attached to a wickerwork cap, are worn by farmers who, at the time of planting and harvest, dance in imitation of leaping antelope. In the Nationalmuseet, Copenhagen. Height 50 cm.

By courtesy of the Ethnographical Department, National Museum, Copenhagen

Bambara masks

Bambara, Malinke, and Marka. The Bambara have six societies to which males can belong, and each has its own type of mask. The Ntomo is for young boys before circumcision. Their masks have a line of vertical projections placed transversely over the human face, representing man as God first created him. The Komo is the custodian of tradition and is concerned with all aspects of community life—agriculture, judicial processes, and passage rites. Its masks are of elongated animal form decorated with actual horns of antelope, quills of porcupine, bird skulls, and other objects. Masks of the Kono, which enforces civic morality, are also elongated and encrusted with sacrificial material. The Tyiwara uses a headdress representing, in the form of an antelope, the mythical being who taught men how to farm (Figure 15). The Kore, concerned with the sky and with the bringing of rain to make the crops grow, employs masks representing the hyena, lion, monkey, antelope, and horse. In addition are created masks of the Nama, which protects against sorcerers. Originally forming successive grades of initiation, the latter five societies now seem to function independently. Originally intended to promote the fertility of the crops by mystical means, the Tyiwara ceremony is now largely concerned with honouring the most energetic farmer and with encouraging others to greater efforts. Ancestor figures of the Bambara clearly draw on the same artistic tradition as do many of those of the Dogon; so also do their sculptures in wrought iron. Rectangular intersection of flat planes is a stylistic feature common to Bambara and Dogon sculpture. The masks and figures of the Malinke are closely similar to those of the Bambara. While the masks of the Marka derive their form from Bambara models, they make more extensive use of brass sheet decoration; they appear in pairs, apparently representing the courtship of man and woman.

Mossi. The Mossi have been largely Islāmized; but members of the Wango society, at funerals of important persons, still employ a mask with an oval face, surmounted by a tall frontal plank with pierced decoration, resembling the *sirige* mask of the Dogon, from which it is probably derived. A simplified form also is made in sheet iron. The masks are said to be made by a group of blacksmiths of Kurumba origin.

Kurumba. The Kurumba are best known for their antelope dance headdresses, attached to a basketwork cap. They may owe something to the antelope headdress of the Bambara but are very different in function and style. Moreover, they are elaborately painted in geometric designs before each appearance at a funeral, at which headpieces in the form of a crocodile and one of more abstract form are also used. Recent examples have been carved so as to cover the wearer's face.

Bobo-Ule and Bobo-Fing. The Bobo-Ule hold ceremonies invoking the deity Do to send rain at the time of planting and again at harvest. For these ceremonies, they use a variety of masks. One has a flat, circular face and projecting mouth and beak, with a tall vertical panel above (resembling Dogon *sirige* and Mossi masks) decorated with geometric designs in black, white, and red paint (see Figure 4). These masks are said to represent owls. A similar mask, with horizontal panels, represents a butterfly (Figure 16). There are also masks in the form of birds, scorpions, antelope, and cattle that do not have these projecting panels. All of these masks are also used for funeral ceremonies. The Bobo-Fing carve masks in hardwood, representing human faces and ram, antelope, and bush-cow heads, with strongly developed sensitivity to sculptural form.

Lobi-speaking peoples. The various Lobi-speaking groups do not seem to use masks at all but produce ancestor figures, commonly made of mud as well as wood, with heads apparently used as puppets. They also make a characteristic three-legged stool, with a head and two legs at one end and a third leg sloping outward at the opposite end. These are used by older men, who pound the ground with them in a septennial *dyoro* festival that honours the Volta River, on the banks of which these people live.

Sculpture of the Lo society

Senufo. The Senufo produce a rich variety of sculptures, mainly associated with the Lo society (known more widely as the Poro), to which all adult men belong and which maintains the continuity of religious and historical traditions, especially through the cult of the ancestors. During initiations, headpieces are worn that have a flat, vertical, round or rectangular board on top, decorated by paint or by pierced work. Many wood carvings of male figures depict these headpieces, sometimes on rhythm pounders used by young initiates, who beat the earth to call upon the ancestors to take part in the ceremony and purify the earth. Several types of mask are used. The *kpelie*, a human face with projections all around, is said to remind initiates of human imperfection. Animal-head masks usually combine characteristics of several creatures—hyena, warthog, and antelope. A type of animal mask called *waniugo* has a cup for a magical substance on top; these masks blow sparks from their muzzles in a nighttime ritual protecting the village from sorcerers. Among the Naffara group of the Senufo, masks of similar form, but with an interior cavity too small for a human head, are carried on the top corner of a rectangular,

By courtesy of the Museum of Primitive Art, New York; photograph, Lisa Little

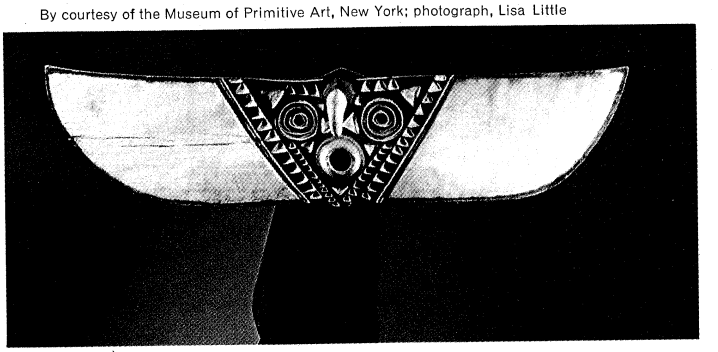


Figure 16: Bobo-Ule mask representing a butterfly, painted wood, Upper Volta. The mask is used to invoke the deity Do in ceremonies asking for rain and for fertility of the fields. In the Museum of Primitive Art, New York. Length 1.30 m.

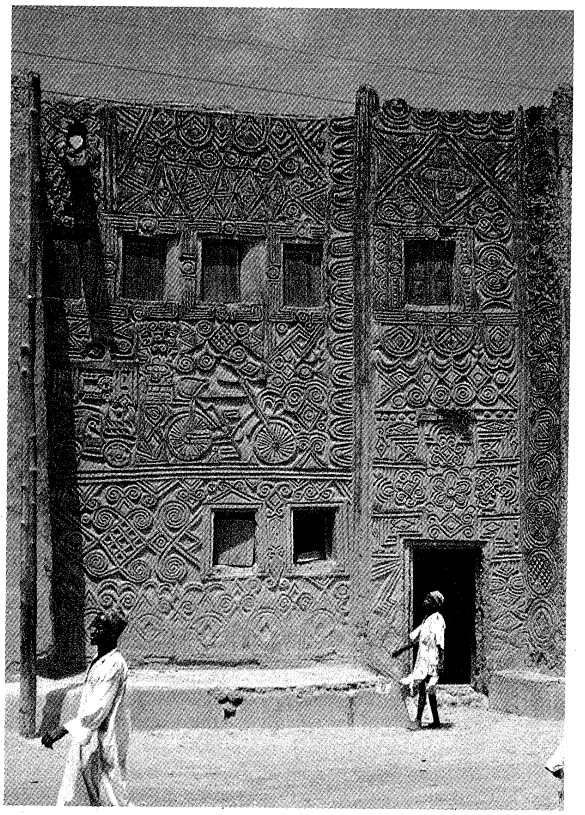


Figure 17: Hausa building with molded low-relief decoration, Zaria, Nigeria.
Frank Willett

tentlike costume called *kagba*. This mask is the symbol of the Lo, which only initiates may see. In the Korhogo region, *deguele* masks appear in pairs at funerals. They are of plain helmet shapes topped with figures whose bodies are carved to resemble a pile of rings. Figures of the hornbill are used in initiation, and groups of birds on a pole are trophies for the best farmer. Figures of women and of horsemen are used in divination. Shrine doors and drums are carved in relief, and small figures and ritual rings are cast in bronze. For many years the Senufo have been producing large quantities of carvings for the European market.

Nafana. The Nafana and their neighbours use flat masks known as *bedu*, which appear in pairs, the male about five and the female about seven feet (150 and 210 centimetres, respectively) tall. They appear every night for a month at the winter solstice, blessing every house in the village against disease and misfortune. They are also used for funerals.

Fulani. The Fulani are basically nomadic pastoralists who range across from Senegal to the Cameroon grasslands. Many of them have settled on the Upper Niger, where they have formed a kind of caste system for their various artists—goldsmiths and leatherworkers, blacksmiths and potters—all of whom have been drawn from the indigenous population. Indeed, the only truly Fulani expression of the visual arts (apart from the decoration of calabashes) consists of body decoration, which reaches its height in the annual *gerewol*, a beauty contest between men, whose faces are painted and who wear metal bracelets, bead necklaces, and head ornaments. The women regularly wear elaborate hairstyles (often featuring golden rings around separate locks of hair), together with a profusion of jewelry. The Fulani express their quest for beauty in all aspects of their lives, even in the way they pack their few possessions onto their transport cattle, carefully arranging their decorated calabashes and their woven covers to show to advantage. Fulani designs are geometric and are thought to antedate their conversion to Islām.

Body decoration

Hausa. In Nigeria the Sudanic belt has long been dominated by the Muslim Hausa who, since the 19th century, have been ruled by Fulani emirs. No representational wood sculpture is produced by the Hausa, though for centuries their buildings have been decorated inside with molded and painted low-relief decorations (Figure 17), which have more recently been applied to the exteriors. Both decorative and of a high technical standard are their crafts: leatherwork for saddles, bags, hilts, and sheaths; gold and silver jewelry; ironwork; pottery; weaving and embroidery (Figure 18). Of special interest are containers made from split green cowhide dried over a mold in what is called the *tandu* technique. These lidded bottles are extremely light and almost indestructible.

Other groups in northern Nigeria. The dominance of the Hausa and Fulani has driven many groups to take refuge in the rocky hills of northern Nigeria. Most of these peoples have been little studied from an artistic point of view and have become known only through a very few examples of their work. Some of the peoples who became Muslim are thought to have lost their traditions of sculpture because Islām discourages the depiction of living creatures.

Nupe. The Nupe have been Muslim for some centuries and are best known for their glass beads and their sheet metalwork decorated in engraved and repoussé (hammered in relief) designs that are part of the Islāmic heritage. Their stools, in a variety of forms, show similar Islāmic decoration. Yet they have also produced many doors carved in low relief in a blend of decorative designs, which are consonant with Islāmic teaching, and

By courtesy of (right) the Tropenmuseum, Amsterdam; photograph, (left) EB Inc.

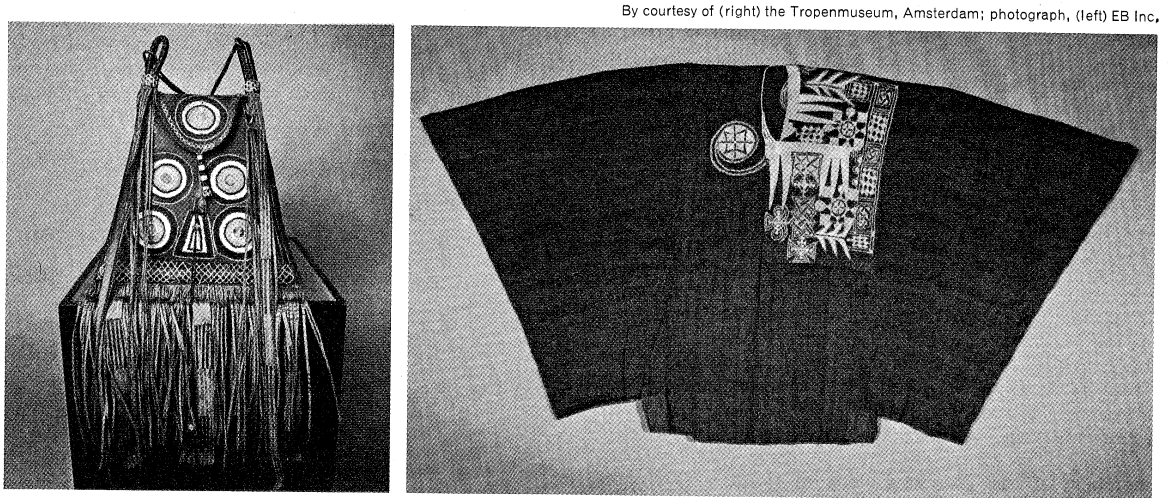


Figure 18: *Hausa crafts, Nigeria.*
(Left) Leather bag decorated with appliqué and stitching. In the Herbert Baker Collection, Chicago. (Right) Man's embroidered cotton robe. In the Tropenmuseum, Amsterdam.

stylized representations of God's creatures, which are not. Masking traditions continue vigorously, however. Carved and painted masks are made for the *elo*, a purely secular performance intended only to entertain (nowadays on the Prophet's birthday), and perhaps introduced, like the *gugu*, which appears with it, from the Yoruba. (The name *gugu* appears to be derived from the Yoruba *egungun;* see below *West African cultures: Yoruba*). The *elo* mask has a human face with a motif (sometimes a human figure) rising above it, flanked with stylized horns. The *gugu* wears a cloth mask decorated with cowrie shells, but sometimes Yoruba *egungun* masks are used. The *ndako gboya* appears to be indigenous and entirely pagan in character; a spirit that affords protection from witches, it is controlled by a small secret society that cleanses communities by invitation. The mask is not sculptured but consists of a tall tube of white cotton supported inside on a bamboo pole about 12 feet (370 centimetres) long.

That Nupe art should have been influenced by the Yoruba is not surprising. Not only do Yoruba live among the Nupe, but there are bronzes in the Nupe villages of Tada and Jebba—one of them apparently an Ife work, and another in a more recent Yoruba style, with its hands clasped in a gesture of their Ogboni society. Others of this group, which include the largest castings ever made in black Africa, share features with Benin sculpture and have other elements that are widely distributed in time and space on the Lower Niger. Nupe tradition says these sculptures were taken from Idah, the Igala capital, in the early 16th century. Many of them were probably already ancient, but unfortunately nothing is known of ancient Igala bronze casting (see below *Igala*).

Tada and Jebba bronzes

Afo. The most important of the groups known from relatively few pieces is the Afo, who live just north of the confluence of the Niger and Benue rivers. Best known for mother-and-child figures, which apparently represent the tutelary spirit of the village, their carvings have been found among many of their neighbours. They also employ enormous bush-cow masks covered with detachable small figures; highly abstracted dance headdresses; and masks with white faces thought to have originated with their neighbours to the south—the Idoma and Ibo (see below *West African cultures: Ibo*).

Mama. The Mama are known mainly for animal headdresses that brilliantly abstract the essential characteristics of the bush cow, the waterbuck, and the reedbuck for the *mangam* ceremonies, which encourage the crops to grow. Carvings of human figures appear unfinished, with facial features sometimes totally omitted.

Ham and Koro. The Ham and Koro around Nok have abstract dance headpieces decorated with red abrus seeds, which are used in agricultural ceremonies, and cups, in the form of human figures, used in an annual sacrifice and in second burial ceremonies subsequent to interment. The Koro apparently carve for both groups. The Ham also have a masquerade *dodo*, which wears a netted costume covering the entire body, without a sculptured mask.

Montol and Goemai. The Montol and Goemai have netted fibre masks, with stylized features, worn with palm-leaf costumes. Carved masks called *gugwom*, representing wooden crocodiles, are used at the millet harvest and at the burial of chiefs. Both groups use carvings of human figures in healing rites. The figures are obviously traded in the area, for their style is found also among the Kanam, Yergum, and Angas; but their source is uncertain. In the southern part of the Goemai area is found another mask, that is the Goemai version of the *aku onu* mask of the Jukun (see below).

Jukun. The Jukun, situated athwart the Benue Valley, which was a route of contact between the northeast and the confluence with the Niger, was a strong kingdom by the late 14th century AD and by the 17th century controlled Bauchi and Gombe, far to the north. Its power was ultimately broken by the Fulani in the early 19th century. Despite its historical importance, little has been published on its arts. It is known, however, that bronzes were cast. Surviving pieces include small animals with wide-open mouths, intended as containers; two large bronze neck rests; and a cylindrical cast box with lid. The best known Jukun sculpture is the *aku onu* mask, perhaps the most abstracted form of the human head known on the whole continent. The mask represents a beneficent spirit who speaks for the ancestors both when good fortune is celebrated and at funerals.

Tiv. The Tiv place paired wooden figures—male and female—to the right of the door of the house of the senior wife as part of a religious rite for warding off catastrophes, famine, and disease. In collections, the Tiv are represented more commonly by their very detailed brass castings—mostly snuff takers—for the modelling of which a rubbery latex is used instead of the more usual wax.

Idoma. Idoma sculptors commonly carve without commissions, so that the customer not only chooses what he wants but also determines the use to which it is put. This may be a recent change, however, for formerly all of the dance societies had characteristic masks. That of the *ogrinya*, for men who had killed an elephant, a lion, or a man, was a headpiece often with a white, heart-shaped face. The Akatakpulakpula, a society of elders, had an elephant mask six feet (180 centimetres) long. Both societies perform at harvests and at funerals of members. White-faced figures usually represent the spirit Anjenu and are used in curing the sick.

Igala. The Igala aristocracy claim descent from the Jukun, but no sculptural evidence for this has been published. Evidence of contact with Benin, with which the Igala were frequently at war, does exist, however: the *ata* of Igala wears a fine 15th–16th-century Benin brass mask as his sign of office. The castings at Tada and Jebba (see above *Nupe*) are supposed to have been brought from Igala. Eastern Igala masks, with flat or concave whitened faces, show influence from the Idoma, Ibo, and perhaps Afo. Western Igala masks are usually black and varied in form. All seem to be primarily concerned with the ancestor cult, but some have judicial functions too. Figures are used to protect the family's well-being. Formerly, doors were carved in relief.

Igbira. The Igbira, on the opposite bank of the Niger, have many dance societies, few of which employ carved masks. The societies that do use masks traditionally employ those that have prominent foreheads, noses, and mouths; their surfaces are left unsmoothed. Currently, they also use indiscriminately masks imported from many other traditions—such as Yoruba, Edo, Ibibio, and Bassa Nge—probably because to them the dance is far more important than the sculpture.

Bassa Nge. The Bassa Nge (literally We Are Not the Bassa) are a Nupe group that settled north of the Niger–Benue confluence about 1869. They produce masks that are a blend of human and animal features, with the surfaces decoratively carved all over.

Vere, Mumuye, Chamba, Bagirmi, and Kotoko. The Adamawa Mountains and their foothills along the Nigeria–Cameroon border are the home of several sculptural traditions. The Vere possess a complex bronze-casting tradition, as well as figure carving in wood. The Mumuye figures are carved so that the arms and sometimes even the abdomen enclose space within the sculpture. They seem to be ancestor, guardian, and healing figures. The Chamba make masks in the form of a bush cow with a very wide mouth, but it is not yet certain whether they make their own figures or obtain them from the Mumuye. To the north and east, wooden dolls are found among the Bagirmi, wax ones among the Kotoko. Across the eastern Sudan little sculpture is found among the non-Bantu-speakers, except for human figures of such simple form that their style is difficult to identify, until one reaches the Mangbetu (see below *Central African cultures: Mangbetu*).

Mambila. Mambila carvings are done in a boldly cubistic style and are painted red, black, and white. The Mambila hold masked dances at sowing, harvest, and funerals. The principal dancer wears a mask that encloses the head; he is accompanied by dancers wearing cap-masks and others wearing netted costumes only. The

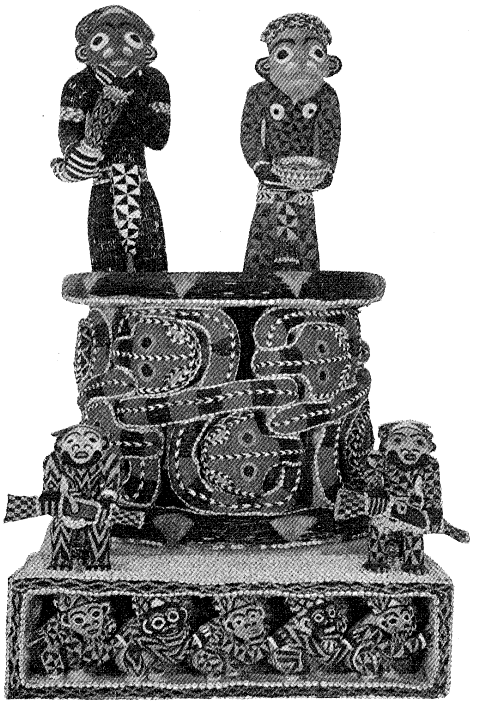


Figure 19: Bamum beaded throne of a king, Cameroon grasslands. In the Museum für Völkerkunde, Berlin. Height 83.2 cm.
Holle Bildarchiv, Baden-Baden, West Germany

Mambila also carve ancestor and protective figures, paint on panels of split bamboo, and weave the standing grass into human figures.

Groups in the Cameroon grasslands area. The Cameroon grasslands area can be divided into three stylistic areas. The Bamileke area is composed of a number of separate chiefdoms, the best known being the Bangwa and the Bacham. Human figures are composed of a highly expressive blend of rounded and angular forms. The Bamum, a large kingdom whose dynasty is from Tikar, has developed roundness of sculptural form almost to its extreme, producing figures with big, inflated cheeks. Among the Tikar, Bekom, and Babanki, the forms are rounded but not exaggerated. Throughout the grasslands there have been exchanges of art objects and diffusion of the brass-casting technique, confusing the more detailed stylistic picture. In general, however, all of these societies are hierarchical, with sculpture mainly intended to reflect the power and importance of the king: palace doors and columns; thrones (Figure 19) and drums; masks, some of which are worn by the king and others worn at his enthronement and funeral. The Babanki held the sculptor in such respect that he could accede to the throne. Houses in this area are constructed of raffia poles and are basically square in form, with a steep pyramidal roof of much larger size set on top, overhanging to provide a veranda on which male and female ancestor figures stand flanking the door (Figure 20).

Construction of houses

WEST AFRICAN CULTURES

Types of art. *Architecture.* The architecture of the West African rain forests is characterized by rectangular mud-walled huts with roofs thatched with broad leaves. In Ashanti, four such houses joined together by short walls are set around a square courtyard, the wall surfaces being decorated in low relief. Among the Yoruba, the separate units are constructed as a single building, with a continuous veranda open onto the courtyard. When constructed on a small scale, the courtyard, called an impluvium, becomes a room with a roof sloping into the centre; the rain from the roof is caught in pots or a tank, the excess conducted away in drains. Houses of this type were also built in Benin. Nowadays, they are being replaced by multi-story houses, usually rectangular in plan and built of mud bricks or cement blocks and corrugated iron. In Yorubaland and westward to Sierra Leone the fronts of such houses are often decoratively treated in relief and paint in a style known as Afro-Brazilian, since it was introduced by liberated slaves returning from Brazil. Islām began to penetrate the area only in the 19th century. The mosques consequently tend to resemble Christian churches, with which they were, and are, in competition. In preference to traditional materials, both are usually built of cement blocks or stone and have corrugated iron or asbestos roofs.

Decorative arts. Pottery is highly developed; vessels intended for ritual use are often decorated with high-relief sculpture. Calabash carving employs both incision and excision with great skill. The most richly woven cotton cloths in the whole continent are made in West Africa; *e.g.*, in Ashanti (Ghana), Akwete (Ibo, Nigeria), and Okene (Igbira, Nigeria). Gold and silver jewelry is widely made, while the Ashanti and Baule have developed a technique of laying gold foil over carved wood.

Sculpture. The sculpture of the area tends to be well rounded and often very smoothly finished and blackened or else painted in several colours. Many traditions achieved a high degree of naturalistic representation, which reached its peak in medieval Ife. In areas that have not developed strong centralized kingdoms, masks are an important means of social control as well as a means of reinforcing the sense of group identity through initiation and the cult of the ancestors. In the discussion below, the many ancient centres of sculpture in West Africa are taken up geographically, from west to east.

Bijagós. In the Bijagós Archipelago live the warlike Bijagós, who, in rather simplified forms, carve human figures seated on stools, bowls supported by human figures with human and animal forms on the lids, and staffs with figures. They also carve naturalistic masks of wild bulls, which are carried on the prow of the royal war canoe and also used in dances. Their buildings are round and often painted.

Nalu, Landuma, and Baga. On the coast opposite are the Nalu, Landuma, and Baga. All three share the male Simo society, which promotes agriculture; the society is their version of the Poro (see below *Dan-Ngere*). Some types of sculpture have been attributed to more than one group. The Nalu are best known for their highly coloured *banda* masks, four feet (120 centimetres) or more in length, which combine human, crocodile, antelope, serpent, and bird features. The Landuma employ headpieces representing bush cows, a motif repeated on their combs.

Photo Hoa-Qui

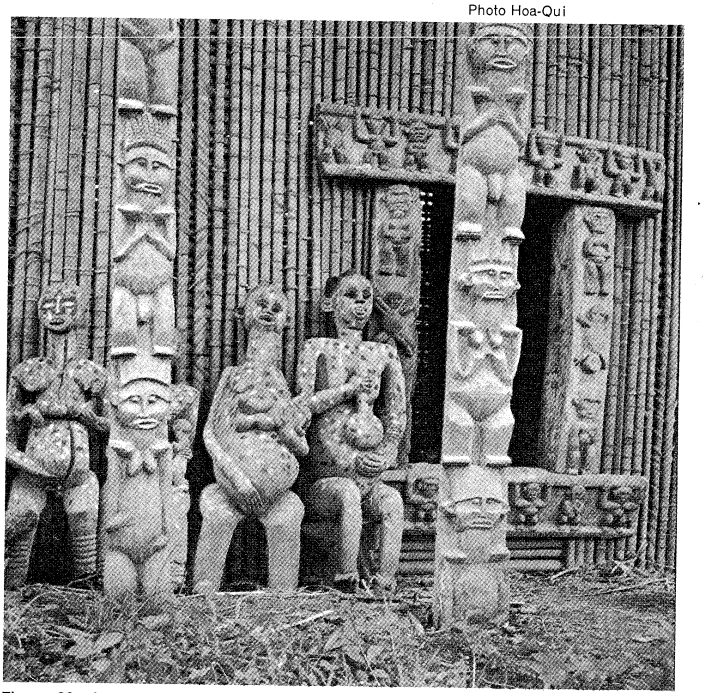


Figure 20: Ancestor figures, carved doorframe, and veranda posts on a Bafoussam chieftain's house, Bamileke area, Cameroon grasslands.



Figure 21: Baga *nimba* mask used in rituals of the Simo society during the harvest and threshing of the rice crop, wood and fibre, Guinea. In the British Museum. Height 1.22 m.
By courtesy of the trustees of the British Museum

For the rites of initiation into adult life, the Basundi society employs sinuous representations of the python, which is associated with the rainbow; the carvings are between eight and 14 feet (240 and 430 centimetres) tall. The Baga, recent migrants from the Sudan, carve *anok*, or *elek*, bird heads with human features stylized in a manner resembling western Sudanic work. These are used for protection against sorcery, in agrarian rituals, and at funerals of important men. Better known is their massive *nimba* mask (Figure 21), with its great head cantilevered out from a columnar neck; the head and neck are supported on the upper part of a female torso, carved so as to rest on the shoulders of the wearer, who sees out through a hole between the breasts, his body hidden in raffia fibre. This mask appears at the harvest and threshing of the rice crop. Tall drums supported by a human figure are also carved.

Temne. The art of the Temne has been little studied; one very elegant female figure has been repeatedly illustrated, and a few striking masks that resemble the hippopotamus are known. These masks are reportedly used by members of the Rabenle, or Maneke, society on the death of a chief.

Masks of the Sande society

Mende and Sherbro. The Mende are best known for smooth, black, cloche-shaped *bundu* masks (see Figure 1), representing the face of the Sande society, which is responsible for initiating girls into womanhood. This is the only women's society on the continent known to use masks. Because the mask is "found" beside a stream and is supposed not to be an artifact at all, the carver in this case is anonymous. Members of the corresponding male society, Poro, wear animal-skin masks. The women also have the Yasse, a divination and healing society that employs slender human figures called *minsere*. *Gongoli* masks are also used, apparently entirely for entertainment and consequently of no fixed form. In preparing their rice farms, the Mende often uncover figures carved in soapstone, which they set up in shelters to protect the crop. The figures are similar in style and are thought to be similar in date to Afro-Portuguese ivories carved in the 16th century for Portuguese traders in the adjacent Sherbro area. (Others were carved in Benin.) The Sherbro share *bundu*, *gongoli*, and Poro with the Mende.

Dan-Ngere. The Dan-Ngere complex of styles is one of the most studied and least understood of the major sculpture-producing groups in Africa. It is named after two extremes of stylistic variation: the smooth, restrained style of the Dan, De, and Diomande; and the cubistic, grotesque style of the Ngere (or Guere), Wobe, Kran, and Bete, a less extreme form of which is found among the Kru and Grebo. Accurate attribution of a work is impossible without field data, for it has been demonstrated that a single carver will produce masks in both of the extreme modes of the range of style. Nor is it possible to give a clear overall picture of the role of the mask in these societies: one scholar reports that they were used to obtain the intercession of the ancestors; another maintains that they are materializations of bush spirits who wish to take part in human life; and yet another explains that the mask originates in rites differentiating the sexes and is therefore a symbol of virility. The masks certainly have many functions—initiation, keeping the peace, supervising agricultural activities, protection from fires, and entertainment. In many parts of the area the masks have been taken over by the Poro, the men's secret society, and act as the society's agents in maintaining social control. Miniature, easily portable masks, representing and sharing in the power of the larger masks, protect an initiate when he is away from home. The carvers also produce the large anthropomorphic rice ladles used by the mother of the heir apparent at the harvest feast; chief's staffs; and female figures that seem to be prestige items, as are small figures cast in brass among the Dan and Kpelle.

The Poro society

Guro. The Guro speak a Mande language related to that of the Dan; but their arts, especially their human face masks, show strong influence from the Akan-speaking Baule. The Guro make polychrome antelope masks generally thought to be for the Zamle society. They appear at public festivals, nowadays *le quatorze juillet* ("the 14th of July," celebrating the fall of the Bastille in 1789), together with other animal masks gaudily coloured and often having rubber ears. Heddle pulleys for weaving, and a number of rather robust figures are also carved.

Ashanti and Baule. The Akan-speaking peoples share a common cultural and artistic heritage best seen in the kingdom of Ashanti, which was founded in the early 17th

Frank Willett

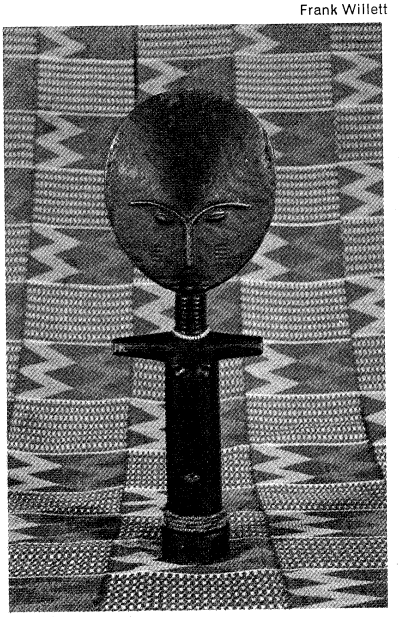


Figure 22: *Akua-mma*, a wood doll that represents the Ashanti ideal of beauty, Ghana. The dolls are carried by young girls and pregnant women to encourage their babies to be beautiful. In the C.M. Hutchins Collection. Height 26 cm.

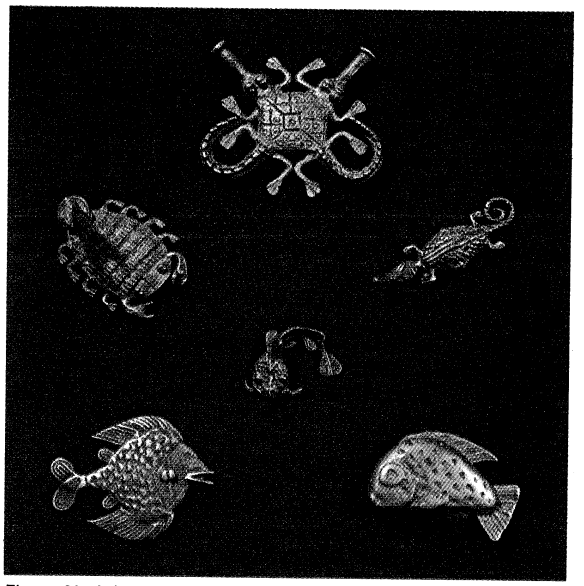


Figure 23: Ashanti brass weights for weighing gold dust in Ghana. The top weight, which is in the form of two crocodiles with a common body, is a reminder of the proverb "We have different mouths, but the same stomach," meaning that each individual wishes to have credit for his contribution to the common good. In the Herskovits Collection. Maximum length 7.6 cm.
Frank Willett

The golden stool of the Ashanti

century when, according to legend, the golden stool descended from heaven into the lap of the first king, Osei Tutu. The stool is believed to house the spirit of the Ashanti people in the same way that an individual's stool houses his spirit after death. The Ashanti also carve *akua-mma* (dolls with disk-shaped heads embodying their concept of beauty and carried by pregnant women to encourage their babies to be beautiful; Figure 22); figures of queen mothers; and staffs for royal spokesmen, which, like the handles of state swords, are covered in gold foil. The success of the Ashanti kingdom depended on the trade in gold not only with Europeans at the coast but with the Muslim north. Gold dust was the currency, weighed against small brass weights (Figure 23) that were often geometric and symbolic or were representations recalling well-known proverbs. Cast-brass ritual vessels, *kuduo*, used in funeral ceremonies, bear indications of Islāmic inspiration. The Ashanti also cast fine gold jewelry, as do the Baule, who separated from them in the mid-18th century.

Baule gold weights are similar to those of the Ashanti, but the Baule also have types of sculpture that none of the other Akan peoples possess: masks (which, like their low-relief doors, seem to indicate Senufo influence) and standing human figures of uncertain use but apparently sometimes used as ancestor figures. The figures and human masks, the latter reported to be portraits used in commemorating the dead, are elegant—well polished, with elaborate hairdressings and scarification. More roughly finished are the *gbekre* figures, representing minor divinities as human figures with animal heads. Masks are made also to represent the spirits of the bush: antelope, bush-cow, elephant, monkey, and leopard. Boxes for the mouse oracle (in which sticks are disturbed by a live mouse, to give the augury) are unique to the Baule, whose carvers also produce heddle pulleys, combs, hairpins, and gong mallets.

Anyi, Fanti, and Brong. The Anyi separated from Ashanti in the second half of the 19th century. In their area, terra-cotta heads and figures have been found, which are thought to be commemorative and to date from the 17th century. Their queen-mother figures, carved in wood, have great elegance. The Fanti make dolls with tall, rectangular heads. The Brong have a special reputation as skilled brass casters.

Fon. The Fon kingdom, with its capital at Abomey, was founded in the early 17th century. The palace is decorated with painted relief panels modelled in clay, symbolically representing the different kings and the events of their reigns. The kings are represented also by iron staffs and messengers' staffs with openwork iron symbols on a wooden haft, as well as by a small number of large wooden statues combining human and animal attributes. The thrones of the kings are similar in form to Ashanti stools, but much taller, and are preserved as cult objects. Small figures cast in brass, often in groups, are prestige items employed also to decorate the royal tombs. Brightly coloured appliqué cloth is used in the worship of the gods; on state umbrellas and chiefs' caps; and for the banners of Dokpwe, associations whose members are obliged to help each other (Figure 24). Popular sculpture is mainly in the form of *bochio*, wooden figures serving as guardians of the household; and calabash carving is widespread. The greatest achievements of Fon art, however, are the large sculptures for Gun, the god of iron and war, made from sheets of copper or iron.

Representations of Fon kings

Yoruba. Fon art has been strongly influenced by that of the Yoruba, who believe that the world was created at

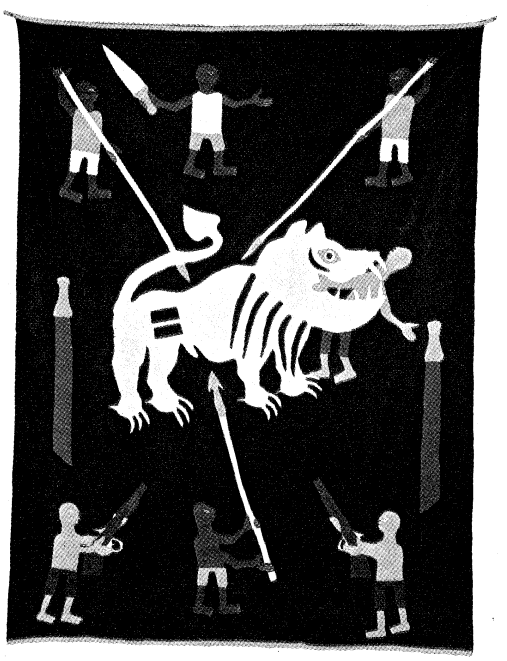




Figure 24: *Fon decorative arts, Dahomey.* (Left) Appliqué banner representing a lion hunt. In the Museum voor Land- en Volkenkunde, Rotterdam. (Right) Lidded bowl made from a calabash. The design is produced by cutting through the brown outer surface into the lighter layer below. In the Herskovits Collection. Diameter 22.2 cm.
By courtesy of (left) the Museum voor Land-en Volkenkunde, Rotterdam; photograph, (right) Frank Willett



Figure 25: Brass figure of an *oni* ("king") of Ife, Nigeria. In the Museum of Ife Antiquities, Nigeria. Height 46.7 cm.
Frank Willett

Ife. Excavations have shown that naturalistic sculpture in brass and terra-cotta was being produced by the 12th century AD. On grounds of its unusually large scale and of details of style and subject matter, the art is thought to have been derived from that of the Nok sculptors (see above *Sudanic cultures*), though the nature of the connection is still obscure. The Ife sculptures represent royal figures and their attendants, and life-size portrait heads in brass were used as part of funerary effigies. In the Middle Ages, Ife appears to have had widespread political importance, and the naturalism required for royal commemoration seems to have influenced the basic development of Yoruba sculptural style. Throughout Yorubaland, human figures (Figure 25) represent a fundamentally naturalistic canon, except for bulging eyes; flat, protruding, and usually parallel lips; and stylized ears. The evolution of these characteristics can be observed in a number of terra-cotta sculptures at Ife.

Variants of the basic Yoruba style

Within the basic canon of Yoruba sculpture, many local styles can be distinguished, down to the hand of the individual artist. Individual cults, too, have their own characteristic variants of the basic style. The Ogboni, a men's cult of the earth that holds the ultimate political power, even over the king, employs large drums covered with low-relief sculpture and pairs of staffs, usually of iron, topped with human figures in the least naturalistic of Yoruba styles. Osonyin, the universal medicine cult, employs a staff of wrought iron topped with a bird. Universal, too, is the Egungun cult, which ensures that the dead are properly buried and employs a variety of wooden masks and headpieces for the purpose, though many Egungun wear only a costume with a net over the face. Ifa, the divination cult, and Eshu (or Elegba), the principle of uncertainty, are also universal and work together. The diviner employs flat trays with low relief, carved margins, which include the face of Eshu; tappers of ivory or hardwood; small ivory figures of Eshu; beaded bags; carved bowls to hold the 16 palm nuts used in the primary act of divination; and wrought-iron staffs. A representation of Eshu is present in most shrines (usually in the form of an unworked stone) and must be propitiated before offerings can be made to other divinities. In sculpture he is represented with a long cap hanging down his back and often smoking a pipe or blowing a whistle. His devotees employ paired staffs strung with cowrie shells and carried round the neck on a leather strap. Staffs for Shongo, the thunder god, bear the symbol of a double ax. On his altars are placed carved mortars, for the pounding of food in a mortar sounds like thunder; on the wall behind hangs his leather bag, with a motif based on the extensive gesture of a Shongo dancer. As Shongo was king of Oyo, his cult is mainly restricted to areas that were once under Oyo domination. The Gelede society is found only in the southwest of Yorubaland. The dances are performed by men to please the women, who are considered to be potential witches. The masks appear in pairs and have a variety of superstructures intended to please observers. Restricted to Ekiti is the Epa cult, which is connected with both the ancestors and agriculture. The mask proper, roughly globular, has highly stylized features that vary little; but the superstructure, which may be four feet (120 centimetres) or more in height, is often of very great complexity—for example, a king on horseback, surrounded by two tiers of attendant warriors and musicians. The most widely distributed cult is of twins, Ibeji, whose birth among the Yoruba is unusually frequent. Their effigies (Figure 26), made on the instructions of the oracle, are among the most numerous of all classes of African sculpture. They are valuable as standardized treatments of the human body that occur in practically every Yoruba village, thus providing extensive data for the study of stylistic variation. Carved doors and house posts are found in shrines and palaces and in the houses of important men. Fulfilling purely secular functions are bowls for *kola* nuts, offered in welcoming a guest; *ayo* boards for the game, known also as *wari*, played with seeds or pebbles in two rows of cuplike depressions; and stools, spoons, combs, and heddle pulleys.

Mask of the *epa* cult

The sites of Ilorin, Esie, and Owo. To the north, in the Ilorin area, the artistic impact of Islām can be seen in house posts and other secular sculptures carved in geometric forms. Nearby is Esie, where about 800 sculptures in soapstone were found by the local Yoruba population some centuries ago. Their origin is obscure: a few features can be paralleled at Nok, others at Ife, others at Jebba and at Igbo Ukwu; they are by no means certainly Yoruba and might be ancient Nupe works. The city of Owo, to the southeast of Yorubaland near the frontier with the Edo-speaking peoples, developed an art style—indeed, a whole culture—that is a blend of Yoruba and Benin traditions. Ivory carving is especially important, and wooden heads of rams and of humans with rams' horns are used on ancestral altars. Second burial effigies, life-size and naturalistically carved in wood, have been made during the 20th century but were developed

Frank Willett



Figure 26: Pair of Yoruba twin figures (*ibeji*), wood, Nigeria. Height 27.6 cm. The starch resist dyed cloth is also Yoruba. In the Frank Willett Collection.

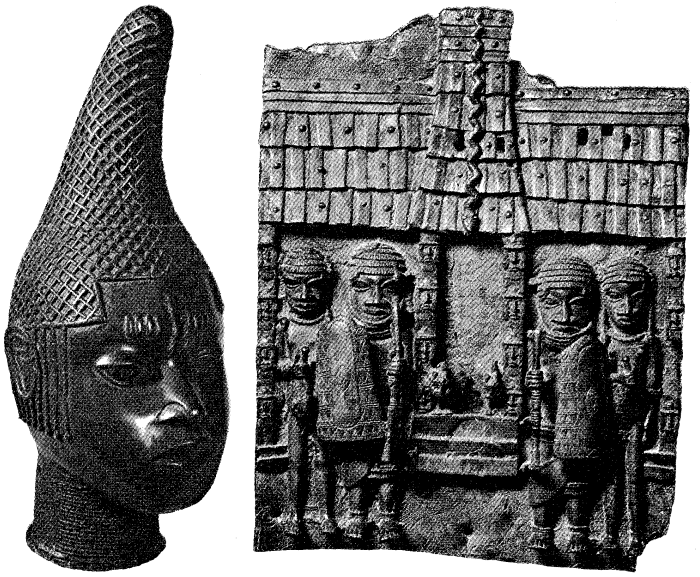


Figure 27: *Benin bronze castings, Nigeria.*
(Left) Commemorative head of a queen mother. In the British Museum. Height 40 cm. (Right) Plaque representing a gateway of the royal palace. In the Museum für Völkerkunde, Berlin. Height 48.3 cm.
By courtesy of (left) the trustees of the British Museum, (right) the Museum fur Volkerkunde, Berlin

from wickerwork forms such as are still used in Benin and in Ibo towns formerly under Benin influence. Excavations in 1971 revealed a large number of terra-cotta sculptures, clearly related to those of Ife but with some Benin features. The site was dated by radioactive carbon to around the 15th century AD.

Edo peoples. *Benin.* According to traditions in Benin, the dynasty was founded from Ife, from whence, in the late 14th century, knowledge of brass casting was introduced for the manufacture of commemorative heads for royal altars. These heads have been grouped in stylistic sequence from moderate naturalism, thought to be inspired by Ife models (one Ife figure having been found in Benin), through increasing stylization. The sequence has been largely confirmed by oral traditions that serve to date features represented in the art, such as the queen mother's curved conical crown, which can hardly antedate the introduction of the title in the early 16th century, and the wings that were added to the king's crown in the early 19th century. The stylistic sequence has been extended to include figures in the round, groups on a common base, and plaques (Figure 27). The rectangular shape of the plaques, their narrative content, and in some cases their attempt at perspective have been attributed to the influence of illustrations in books carried by the Portuguese, who were in contact with Benin from the late 15th century. The technique of brass casting, however, had been introduced at least a century earlier. Bronze bars had been imported, probably from the interior, as early as the 13th century, but these were made into bracelets in Benin only by smithing and chasing techniques, not by casting. Sculpture in brass and ivory (see Figure 2) was a royal prerogative. Lesser chiefs had to be satisfied with wooden heads covered with brass sheet for their ancestor altars; wood was also used for staffs on these and the royal altars and for *kola* nut boxes. Terra-cotta heads were made for shrines in the brass casters' quarter; and life-size groups of royal figures in mud are still made for the cult of Olokun, divinity of the sea and of wealth.

Commemorative heads

Edo peoples outside Benin. Outside Benin, the Edo peoples live in villages that have many localized cults of nearby topographical features and founder heroes. The *ekpo* masquerade, occurring to the south and east of Benin, is performed by the warrior age group in ceremonies to purify the village ritually and to maintain health. At Ughoton, to the southwest of Benin, a different type of mask is used, in the cult of the water spirit Igbile. Both the cult and the art style seem to have derived from the Ijo.

A number of bronze castings found in Benin were made elsewhere. In the absence of more precise information about their origin, they tentatively have been classified with others as the Lower Niger Bronze Industries. They include pieces from Tada and Jebba (see above *Sudanic cultures; Nupe*), a group from the Forcados River, and others from various parts of the delta of the Niger River.

Lower Niger Bronze Industries

Ijo. The delta is occupied by Ijo fishermen, whose masks for the cults of the water spirits are in the form of aquatic animals, especially hippopotamus and crocodile. The western Ijo use *ejiri* figures, in which the head of the household is represented upon a highly abstract quadruped said to represent the guardian spirit of the family. Similar objects are made by the Edo-speaking Urhobo, to the north of the Ijo, where they are used in a cult of aggressiveness by the warriors. Among the eastern Ijo, known as the Kalabari, shrines for the water spirits have figures that are often of large size, though frequently kept hidden from view. They also have masks, similar to those of the western Ijo, worn by men who belong to the Ekine society, which places great emphasis on dancing skill in performances that reflect situations of everyday life. There are also shrines with sculptures for the village heroes and stools with a human face on the front, which ancestral spirits occupy as they occupy figures in The Sudan. In New Calabar, rectangular screens are made to commemorate ancestors. They are fashioned by carpentry (presumably learned from European trading hulks) into a low-relief frontal group in which the ancestor is flanked by supporting figures—much like the king in Benin plaques, by which the screens appear to have been directly inspired about two centuries ago. All Ijo sculpture is starkly cubistic; although carved nowadays by a machete (rather than the almost universal adz), it is not clear whether the style permitted its use or whether the tool determined the style.

Ibo. On both sides of the Niger, but mainly to the east, live the Ibo. Traditionally, they have lived in small village settlements, without chiefs. Only on the northern and western edges of the area, under influence from Igala and Benin, are large indigenous towns and hereditary rulers found. Yet in the traditional Ibo society there was strong social pressure toward individual distinction—men could move upward through successive grades by demonstrating their achievements and their generosity. Among the Western Ibo, a mark of such distinction is the *itchi* scarification across the forehead that is reflected in many of their sculptures; the *ikenga*, shrines for the cult of a man's personal powers, are an expression of the same drive. The decentralized Ibo social structure has produced a great variety of sculptural styles.



Figure 28: Leaded bronze ceremonial object, thought to have been the head of a staff, decorated with coloured beads of glass and stone, 9th century. From Igbo Ukwu, Nigeria. In the Nigerian Museum, Lagos. Height 16.8 cm.
Frank Willett

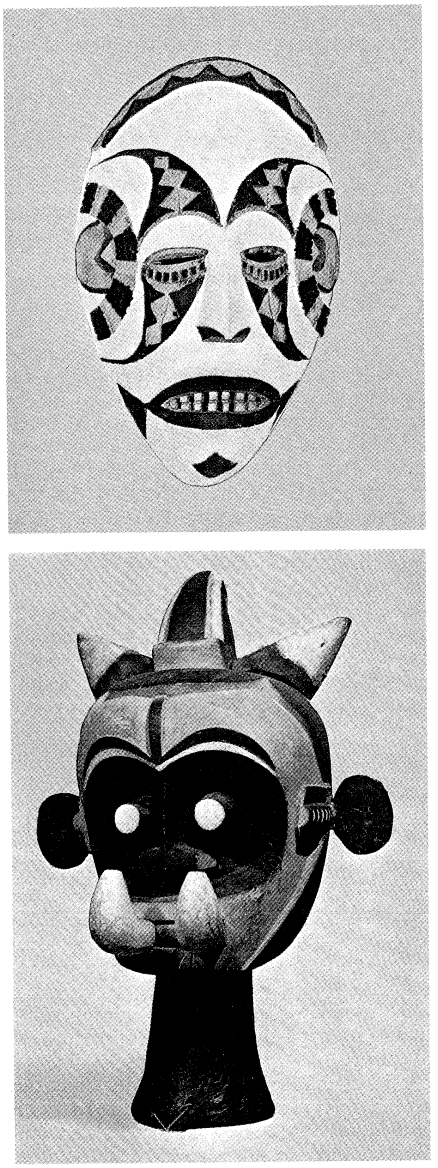


Figure 29: *Southern Ibo Ekpe society masks, Nigeria.*
(Top) Maiden spirit mask symbolizing beauty and peacefulness, painted wood. In the Nigerian Museum, Lagos. Height 21.3 cm. (Bottom) Elephant spirit mask symbolizing ugliness and aggression, painted wood. In the Nigerian Museum, Lagos. Height 48.3 cm.
Frank Willett

Earliest Ibo sculpture

The earliest sculpture known from Iboland is from Igbo Ukwu, where the grave of a man of distinction and a ritual store, dating from the 9th century AD, contained both chased copper objects and very elaborate castings of leaded bronze (Figure 28). The earliest artistic castings from black Africa, these pieces consist of ritual vessels and other ceremonial objects with intricate surface decoration, often small animals and insects represented in the round. Their stylistic affinities are still obscure.

A very great variety of masks is found among the Ibo. The masks are employed in a variety of dramas: social satires, sacred rituals (for ancestors and invocation of the gods), initiation, second burials, and many public festivals, which now include Christmas and Independence Day. Some masks appear at only one festival, but the majority appear at many or all. Best known are those of the Northern Ibo Mmo society, which represents the spirits of deceased maidens and their mothers with masks symbolizing beauty. Among the Southern Ibo, the Ekpe society, introduced from the Cross River area, uses contrasting masks to represent the maiden spirit and the elephant spirit (Figure 29); beauty and ugliness; peacefulness and aggression. A similar contrast is found in their Okorosia masks which correspond to the Mmo of the Northern Ibo. The Eastern Ibo are best known for masquerades associated with the Iko okochi harvest festival, in which the forms of masks are determined by tradition, though the content of the play varies from year to year. Stock characters include Mbeke, the European; Mkpi, the he-goat; and Mba, which appear in pairs, one representing a boy dressed as a girl mimicking the behaviour of a girl, the other representing the girl being satirized. These same masks appear also at the triennial Okonkpo festival marking the end of initiation. Most impressive are the *ijele* masks of the Northern Ibo, which are 12 feet (370 centimetres) high. Platforms six feet (180 centimetres) in diameter, supporting tiers of figures representing everyday scenes, the masks are made to honour the dead to ensure the continuity and well-being of the community. They are costly and time-consuming to make; in this and in their subject matter, they correspond to the *mbari* houses of the Oratta Ibo. These very elaborate mud houses filled with brightly painted clay sculptures are intended to placate the earth deity Ala when a disaster has struck. They take years to build and are then abandoned to decay, for it is the act of building them that is effective, not the final creation.

Wooden figures are carved for ancestors of both sexes, varying from less than one to more than five feet (30 to 150 centimetres) in height. Those representing founders of the village are kept in a central shrine and sometimes become patrons of the market. Slit gongs, a kind of wooden drum, up to eight feet (240 centimetres) in length, with human or animal figures at each end, were formerly kept in a place of honour, where they served as war gongs and, since they have two notes, as "talking drums." Carvers also make doors decorated with geometric patterns for important men, as well as stools, mirror frames, trays for offering *kola* nuts to guests, dolls, and a variety of small figures used in divination.

Southern Ibo *ogbom* headpieces

At Osisa are found elaborate Riverain Western Ibo terra-cotta groups representing a man and his family and made for the yam cult, but there seems to be no tradition of terra-cotta sculpture in other Ibo groups. Riverain Western Ibo *ogbukele* masks for water-spirit cults derive from those of the Ijo, while Southern Ibo *ogbom* headpieces—tall female figures, usually seated on a stool—are derived from the Ibibio. They were used to honour Ala as giver of children. The headdress was not considered to be in any way supernatural; and, though only men could carry it, the whole community could join in the dance (which was abandoned in all but three villages as early as 1900).

Ibibio. The oldest sculptures of the Ibibio are several hundred ancestor figures, *ekpu*, of the Oron, some of which are thought to date from the late 18th century. They are bearded figures three to four feet (90 to 120 centimetres) high, so individual as to suggest portraiture, despite their abstract style. Most of them were allegedly burned by Biafran troops in the Nigerian civil war (1967–70). The Ekpo society uses black masks, often of naturalistic appearance and with movable jaws, to maintain social order and propitiate the ancestors; the masks often represent disease and deformity. One dance, *ekpo urua*, uses white masks expressing beauty. Similar masks and many others are used in *ekong*, elaborate displays put on by the young men for entertainment and for the prestige of their village; the displays also include satirical puppet plays of social comment. Slit gongs and dolls are also carved, and a few brass castings have been found. The Efik, or Riverain Ibibio, use skin-covered headpieces commonly considered characteristic of the Ekoi (see below).

Ogoni. Ogoni, or Kana, sculpture has been much influenced by that of the Ibibio, to whom they may be related. Their masks often have a movable jaw and a human skull surmounting the face. They are used for funerals, annual festivals, and meetings of the society. Larger masks, representing horned animals in athletic performances, are used by the *karippo* at the beginning of the farming season. The Ogoni also use puppets.

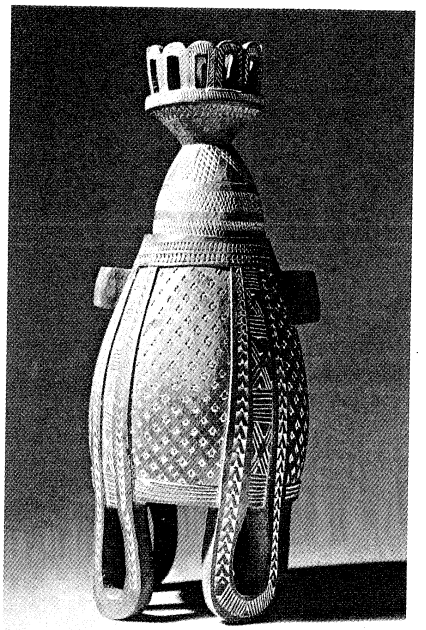


Figure 30: Issa wood milk pot, Somali. The vessel is first blackened and then carved in low relief. In the Musée de l'Homme, Paris. Height 50 cm.
By courtesy of the Musee de l'Homme, Paris

Ekoi masks and headpieces

Ekoi-speaking peoples. The Ekoi-speaking peoples (Anyang, Boki, Ejagham, Keaka, and Yako) are best known for their large, skin-covered masks, which have two or even three faces, and for their smaller headpieces, which represent a head or an entire figure. The headpieces and masks have metal teeth, inlaid eyes, and frequently pegs to represent hair which, alternatively, may be carved in elaborate coils. They are used by the Ekkpe society to ensure the fertility of crops and people alike and are copied widely by neighbours of the Ekoi. In the northern Ekoi area, around Ikom, are found circles of large stones, *akwanshi*, from one to six feet (30 to 180 centimetres) high, carved in low relief to represent human figures. They probably date between the 16th and early 20th centuries.

Bafo and Duala. In the forested coastal plain of Cameroon live the Bafo, whose masks have open mouths with long teeth. Represented with arms curving up to the chin, their human figures are carved both in wood and in ivory (as decorative ends of ivory trumpets). Farther south, the Duala have a very distinctive style of decorative curvilinear sculpture. The general form of Duala stools, which were objects of prestige reserved for elders, is similar to the stools of the Ashanti, though the support is very different in style. Canoe prow ornaments were carved in elaborate openwork tracery, vividly painted. Paddles, too, were brightly painted, as were the elongated masks representing horned animals, either an antelope or a bull. (F.Wi.)

EAST AFRICAN CULTURES

Southeastern (or Nilotic) Sudan. The southeastern, or Nilotic, Sudan is ethnically distinct from the Arabized northern Sudan. Agriculture is practiced, and cattle raising—a gauge of wealth and a Hamitic culture trait—is widespread.

Pole sculpture

There is little scope for differentiating local styles of surviving wood carvings, for all belong to a type known as pole sculpture, in which there is no elaborate treatment of limbs and features. Some of the larger ones, three feet (one metre) or more in height, are attributed to the Bongo and related peoples and appear to be ancestor figures that were placed on the graves of important people or around the villages. The Bongo also made smaller figures, used in murder trials to identify criminals by divination or psychological means or as ancestor figures within the huts. Other peoples, especially the Bari, also made ancestor or fetish figures that were kept in the huts. Wood carvings have been attributed to the Dinka and Jur (Djour), possibly incorrectly. The Ingassana have wooden figures called *tshalk*, used in family rituals.

The Shilluk made life-size, roughly carved wooden statues of their god Nyikang; clay figurines of bulls as cattle-fertility symbols; clay pipe bowls and figurines in hyena form; and masks, typically fashioned of a piece of gourd with applied facial features made of cattle dung and fishbone teeth.

Some peoples decorate their huts with wall paintings and reliefs; the Burun, for example, paint animal murals reminiscent of rock paintings. The Nuba make fine pottery of clay or cow dung, sometimes embellished with finely painted geometric patterns. Artistic taste appears in weapons, such as throwing knives, and in domestic utensils, elaborate coiffures, and personal ornaments.

Horn of Africa. The Ethiopian, Galla, and Somali peoples, the chief inhabitants of the Horn of Africa, are Hamitic in culture; that is, cattle breeders, often seminomadic. They are usually Muslims or pagans, and there may be elements of black stock.

Wooden tomb posts

The Gato, Konso, and Ometo peoples of southern Ethiopia carve wooden tomb posts about six feet, six inches (200 centimetres) high, of which one-third is buried. Made for deceased nobles or warriors, the posts are surmounted by a carved head and shoulders; sometimes the crested headdress bears a phallic symbol. A vigorous and distinctive style gives these figures great dignity. The Arusi, also of southern Ethiopia, make tombstones of like height, ornamented with engravings filled in with red or black, sometimes showing the deceased in rough relief. Similarly shaped gravestones occur in Somalia, bearing a phallic symbol, sometimes plain, sometimes adorned with some sort of decoration. This type of grave monument seems to be Islāmic in origin. Curiously, the graves of the Islāmized Arusi are richer in representational ornament than those of the non-Muslim Arusi or Somali, despite Islām's prohibition against the representational arts.

Masks of wood, leather, or gourd occur among some peoples of Somalia, perhaps owing to an admixture of black blood and culture.

A nomadic or semi-nomadic way of life can affect the applied arts; for example, pottery is less used than wood (Figure 30) or coiled basketry for making containers and dishes for food and liquids. In this region, baskets can be made waterproof and are often patterned in many colours and ornamented with cowrie shells and leather bands (Figure 31). Weaving and leatherwork can reach a high standard; handsome jewelry is made from silver and amber; wooden spoons and combs may be elaborately carved. The Somali have circular shields that are made out of antelope hide and have embossed designs; and their swords, daggers, and spears combine artistic feeling with usefulness.

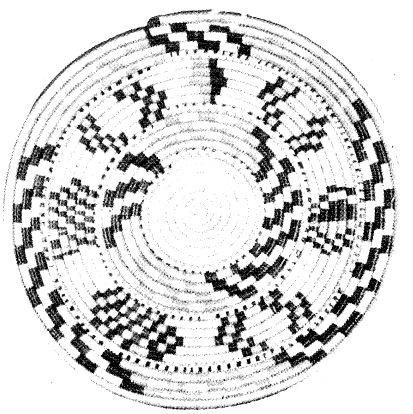


Figure 31: Ethiopian basket. In the Kirsten Vogeley Collection. Diameter 30 cm.
Maude Wahlman

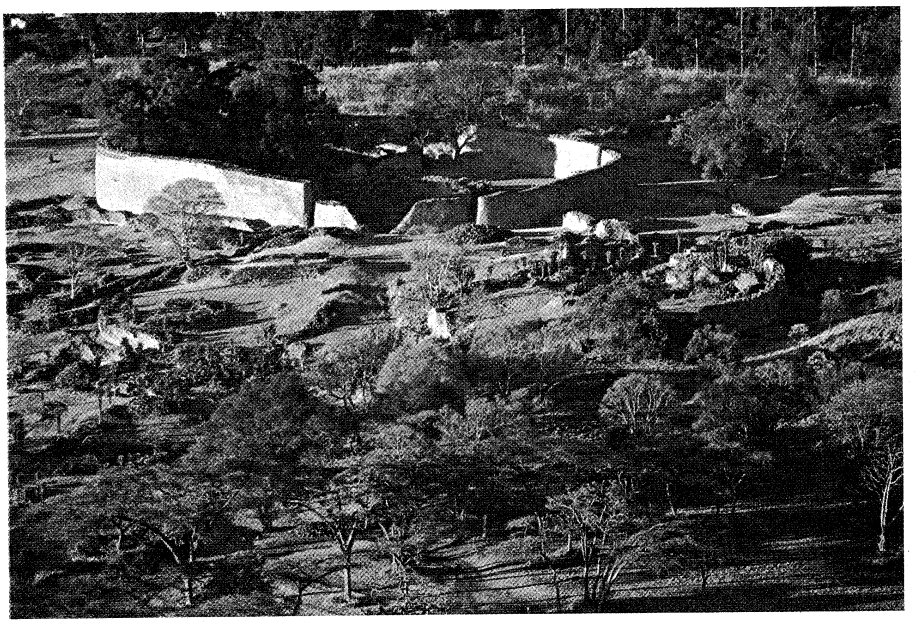


Figure 32: Aerial view of the Zimbabwe ruins, Rhodesia.
ZEFA

Region of Lakes Victoria and Tanganyika. The peoples living around Lakes Victoria and Tanganyika combine Bantu black and Hamitic elements, especially to the north, where a "Hamitic" aristocracy governed a Bantu peasantry.

A pottery head and torso from Luzira in Uganda (now in the British Museum), while probably not more than a few centuries old, is the oldest work of art known from this region and is thought to be a cult object. Surviving from the treasure of Chief Rumanika of the Karagwe (on the west shore of Lake Victoria), seen by the explorer Sir Henry Morton Stanley in 1876, are wrought-iron figures of two bulls and an eland, a bird of copper foil, and other metal objects that may have formed part of the chief's regalia (Linden-Museum für Völkerkunde, Stuttgart, West Germany). Compared to brass and gold, iron is little used as an art medium in Africa, and these iron figures are the only such known from East Africa.

The Kerewe of Ukerewe Island in Lake Victoria carved large wooden figures, about three feet (one metre) high, which appear to have been effigies of deceased chiefs. Other Kerewe carvings are smaller and may belong to a different tradition. These large figures, together with some from the neighbouring peoples, show affinity with Congolese sculpture, particularly in the treatment of the head and body scarification. Three wooden figures collected from the Bende are said to have been originally from Urua (Luba country, in southeast Zaire), which suggests that a trade in carvings could account for the diffusion of the Luba sculptural style. Face masks occur among the Haya (Ziba) and are said to have been worn by court jesters.

Ganda and Tutsi architecture

The peoples around Lake Victoria, notably the Ganda and Tutsi, have brought the architecture of the domed reed hut to a high peak of excellence in the building of royal tombs and palaces. Mats and screens used on hut walls are twill plaited or sewn in patterns of black against a pale, straw-coloured ground. Fine baskets, with a variety of motifs in the same colours, come from the same area, as well as fine black pottery burnished to a high lustre. The domestic equipment is made in great variety and is of high artistic merit. Also made are shields, painted or straw covered and patterned in black or natural colour. Bark-cloth robes are printed or painted in black on a rust-red ground colour.

Eastern Bantu. The area of the Eastern Bantu covers Kenya and part of Tanzania, including the Swahili coast. Arab civilization, through colonization over the past 1,000 years and the activities of traders, has left its mark. Arab ruins occur along the coast on the sites of former settlements, notably at Kilwa, Gede (Gedi), Pemba, and Zanzibar. Swahili art reached a high standard, whether in architecture, wood carving (especially of doors), silver- and metalworking, or finely plaited polychrome mats. Farther inland, direct Arab cultural contact is less obvious. The Islāmic interdiction on the graven image, however, may explain the almost complete absence of old traditional sculptures, an exception being the throne of the chieftainess of Buruku (Nyamwezi). A grave figure from the Zaramo, with articulated limbs, shows Arab influence. Only a few ancestor figures survive from the Shambala; these, like the Nyamwezi throne, combine figure carving with the common three-legged-stool form of the region. The Giryama produced grave planks lavishly chip carved and surmounted by stylized heads, analogous to those from Ethiopia and Somalia and belonging to the same Arab funerary tradition. Notable among the remaining peoples who produce sculpture are the Kamba, who have spontaneously developed in the 20th century a style of wood carving, embellished with coiled-wire jewelry ornament, sold in airport-type gift shops; formerly their art was applied to engraving gourds and inlaying stools with coiled-wire patterns.

Clay figures were made throughout the region, usually for a special purpose, such as for a harvest dance, for hunting down criminals, or for initiation ceremonies. Mostly they are plainly made and symbolic rather than representational in form. Murals occur on the mud walls of huts, sometimes decorative, sometimes ritual and magical. Pottery is normally simple in form and decoration; gourds ornamented with engravings or covered with beadwork are widespread. Stools may be elaborately made, as by the Kamba; shields painted with distinctive polychrome designs occur especially among the Kikuyu and Masai. Traders' beads and coiled brass or iron wire are the raw materials for elaborate personal ornaments in a variety of designs and colour combinations.

(Ma.A.C.)

Madagascar. Madagascar has been separated from Africa since Tertiary times. Its population was not from Africa but from Indonesia and Oceania; its art forms show affinities with these areas rather than with the continent of Africa. (F.Wi.)

SOUTHERN AFRICAN CULTURES

Southeastern Lunda-Luba zone. The southeastern Lunda-Luba zone describes the area covered by Zambia, Malawi, and Mozambique and including Rhodesia and southern Tanzania, where the carving of masks and figures reaches a level higher than that of the rest of



Figure 33: Painted hut and enclosure walls of a Ndebele village near Pretoria, Transvaal, South Africa.
Photo Almasy

eastern and southern Bantu Africa, and where, alone, helmet masks, common in West and Central Africa, occur. The art of this region may be regarded as an outlier of the Central African style of art; many of the peoples concerned have in fact migrated from the Congo Basin and belong to the Lunda or Luba groups.

Zimbabwe ruins

Zimbabwe ruins (Figure 32), perhaps the best known antiquity of black Africa, are associated with the former kingdom of Monomotapa. Archaeological work indicates that the site has been occupied since about the 2nd century AD. The apogee of building skill took place during what is called the fourth phase, the Rozwi occupation of 1450–1833, which also seems to have been the most prosperous period. The grouping of some of the Zimbabwe ruins resembles that of the reed-and-thatch-built palace of the paramount chief of Barotseland, suggesting that standard Bantu ceremonial architecture here underwent a local metamorphosis into dry-stone walling. Similar stone-built ruins occur throughout Rhodesia; the most famous are Khami, Inyanga, and Naletale. Birds on standards, ritual vessels, and small human figures, all carved in soapstone, are associated with Zimbabwe and Inyanga.

The Makonde, living on either side of the Tanzania–Mozambique border, are the most prolific and talented artists in the area. Figures of dancing women are carved with a latent animation and a feeling for movement rare in African art. Masks are more numerous than figures, and may be face masks, worn only over the face, or helmet masks, worn over the entire head. Face masks come mainly from the northern Makonde and are often used in stilt dancing; similar masks come from other peoples in southern Tanzania. Both types of mask, as well as the carved figures, usually feature lip plugs, formerly worn by both sexes but now only by women, and scarification of the face and body. The helmet masks made by the Makonde and the Yao are closely similar. In both, facial scarification is commonly rendered in applied beeswax; human hair is often used on the scalp; and beards are made of animal hair or vegetable fibre. Masks of female torsos also occur; these and the helmet masks were once worn only at dances. The Makonde also produce so-called airport art, often in ebony or ivory, in a great variety of forms and of above average quality.

In Zambia, the Mbunda, Luvale (Balovale), and Chokwe make masks; those of the former are made of wood; those of the latter two are made of painted coarse bark-cloth on a wicker frame. Each type is worn with a netted string costume or a fibre skirt. The masks may be worn at *makishi* dances (held at the new moon), in initiation ceremonies, or for public entertainments. One mask represents a maiden and is used in an acrobatic high pole dance similar in tradition to the Makonde stilt dance.

Small figure carvings are made in Zambia, mostly in the west. Decoratively carved food dishes, stools, headrests, thumb pianos, and snuff bottles come from Barotseland, Mashonaland, and Malawi. Distinctive baskets and clay pipe bowls carved in animal form come from western and southern Zambia; excellent potters, the Shona also use fine brass, copper, and aluminum wire intricately wound and plaited to embellish ax handles, hilts and sheaths, and snuff bottles in a manner that shows affinity with the Congo.

Southern Bantu. The Zulu and related peoples of the southern Bantu made wooden figures that are mostly undistinguished and may have been executed under European influence. Attractive small clay models of cattle, made by children, occur here, as through much of eastern and southern Africa. Much artistic feeling is revealed in such decorative arts as basketry, pottery, wooden vessels, stools and headrests, ceremonial weapons, spoons, pipes, and personal ornaments consisting of beadwork in great variety. Architecture shows the basic African hut form, circular or rectangular, with domed or conical roof: the Ndebele of the Transvaal not only paint the walls of their huts, which is customary, but also decorate their enclosure walls in a variety of coloured geometric patterns (Figure 33). (Ma.A.C.)

Rock paintings and engravings

Bushmen. In southern Africa, rock paintings and engravings occur in abundance (Figure 34). Most surviving paintings were probably made during the last two millennia, but the tradition seems to be much older. The distribution of rock art from the Sahara (where it began about the beginning of the 5th millennium BC) down the highlands of eastern and southern Africa may reflect a pattern of diffusion; but, as yet, no southern African site with either paintings or engravings has been satisfactorily dated to before the time of Christ. Bushmen were still making such paintings in the 19th century, but some of the surviving examples may be the work of Hottentots and Bantu-speakers. The way of life represented is mainly that of Late Stone Age hunters (which is also more or less the way of life of the Bushmen), living off a fauna that has not changed in the area since the middle of the Late Pleistocene Epoch. The earliest paintings depict a peaceful existence; later ones, however, from the 17th century onward, reveal the pressures created by incoming Bantu and Europeans—reflected in scenes of cattle raiding and subsequent fighting between groups (which are recognizable by their stature, dress, and weapons).

The paintings are in three main styles: monochrome, bichrome, and polychrome. The last style has a very



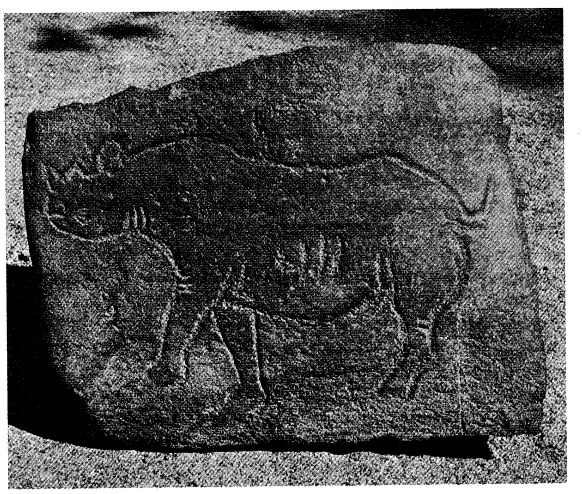


Figure 34: *Bushmen rock painting and engraving in South Africa.*
(Top) Bichrome painting of hunters and animals, Rytal Mountains, near Harrismith, Orange Free State. (Bottom) Engraving of a rhinoceros, southwestern Transvaal.
By courtesy of (bottom) A.R. Willcox; photograph, (top) Jean Vertut, France

restricted distribution in the southeast of southern Africa, an area to which the eastern Bushmen were confined by incoming Bantu in the early 17th century; yet works in this style are the finest achievements of the art, showing foreshortening and carefully composed groups. Ostrich-egg shells (Figure 35), engraved with linear patterns, are the only recent art form produced by the Bushmen, who discontinued rock painting and rock engraving during the 19th century. (F.Wi./Ma.A.C.)

CENTRAL AFRICAN CULTURES

Congolese cultures. Congolese visual arts are stylistically diverse and filled with what can be described in Western terms as spiritual and aesthetic vitality. In Congolese traditional life, as in that of other areas of Africa, neither the creation of art nor the art objects themselves are considered in such critical terms, for art is totally integrated into daily life. The carving of masks and figurines to be used in initiation rituals or other ceremonies, for example, is itself a part of the ritual and the ceremony.

Throughout the Congolese region, the artist's position is prestigious—Luba and Lunda artists, for example, wear an emblematic adz on their shoulders to signify the dignity of their position—even those peoples in which a man is selected to serve as an artist's apprentice because of a physical deformity that exempts him from becoming a warrior or hunter. One reason for the popular respect accorded the artist is the belief that he is in close touch with the forces of magic. To carve fetishes through which contact can be made between the living and the soul of the deceased, the artist and fetish maker must utilize magical powers; and it is the artist who may facilitate this spiritual contact by using the fetish in ceremonial rituals. His role is thus extremely important, for the spirit embodied in the fetish is capable of controlling the forces of evil, of causing sickness and adversity to enemies, and of protecting the individual and the community.

The artist's magical powers

Congolese art. The region referred to as the Congo consists of the modern countries of Zaire (formerly the Democratic Republic of the Congo) and the Republic of Congo (Brazzaville), which lies to the west of Zaire and is separated from it by the Congo River. The area falls into two major geographic divisions: the northern half is an equatorial rain forest inhabited by peoples who hunt, farm, and fish; the southern half is a savanna. It is in the villages of this southern region that the most highly developed political, social, and artistic culture has evolved. The Congo can be divided into areas distinguished on the basis of ethnographic as well as geographic charac-

By courtesy of the Ethnographical Collection, Zurich

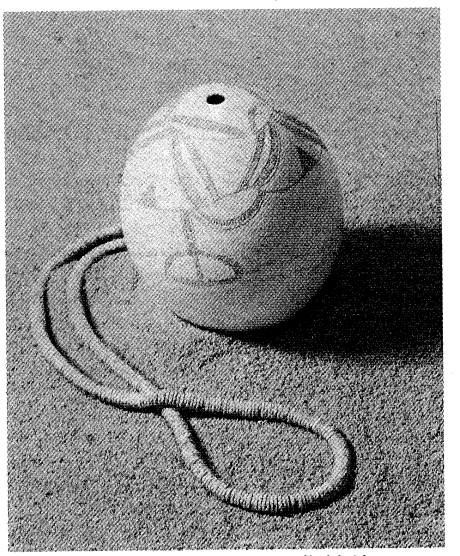


Figure 35: Bushmen ostrich-eggshell drinking vessel with incised designs and a necklace of the same material. From Lüderitz, South West Africa. In the Ethnographical Collection, Zürich. Height of vessel 14.9 cm.



Figure 36: Luba mask, painted wood, Luba cultural area, Congo. In the Herbert Baker Collection, Chicago. Height 61 cm.
EB Inc.

teristics. Most Congo peoples of the savanna have a common Bantu linguistic heritage. There is, in addition, a certain amount of artistic homogeneity. Stylistic differences within the two major regions of the southern savanna and the northern rain forest can best be seen by subdividing the areas according to the kingdoms that have determined the social, political, and artistic lives of the people. The savanna falls into the Lower Congo, Kuba, and Luba cultural areas; and the rain forest into the northern, northeast, and northwest areas.

In general, the style of Congolese art can be characterized as a combination of symbolism and realism, wherein naturalistic forms—predominantly human and animal—are rendered not in precise imitation of nature but in an exaggerated manner. It is this "non-naturalistic reality" that distinguishes Congolese art from West African art.

The most significant art form is sculpture, most commonly wood carvings: masks (used in dance, war, and ceremonies; Figure 36), ancestor figures, fetishes, bowls, boxes, goblets, staffs, pots and lids, pipes, combs, tools, weapons, and musical instruments. Similar objects are also carved in ivory, and in some cases copper, brass, and iron are used. In rare instances stone figures have been found that seem to have been influenced by the techniques used in wood carving.

Painting is not utilized greatly as a separate medium, but carved pieces frequently are painted. Masks and other pieces are covered with polychrome (Figure 37), the colours applied in wide patches and often in planes and angles upon smooth surfaces. In the huts in which rituals take place, wooden figures are hung on brightly painted walls. Painting is also used in textile design on woven fabrics. In the proportionally small number of Congolese paintings that do exist, the painter, like the sculptor, created figures whose large heads and short limbs are more imaginative than realistic. These paintings on bark, to be hung on walls, utilize linear perspective.

The Congolese use aquatic reeds to weave decorated mats, used for sleeping and for wrapping the dead, and baskets and boxes, which are used to contain foodstuffs as well as ritual objects. Basketry patterns and sometimes container forms have been imitated by wood carvers; textile weavers, too, use decorative motifs derived from basketry.

Pottery making, considered a woman's job, has depended on four forming techniques: molding; coiling, as in basketry; modelling on a board; and, more recently, throwing on the potter's wheel. Pottery forms are influenced by those of basketry and wood carving as well as by vegetal forms such as the calabash; decoration consists of traditional geometric incised or painted patterns. The pots are used for cooking and for carrying and storing food or as ceremonial objects.

Even architecture has a place in ceremonial tradition. Special huts are used in circumcision and initiation rites, and special buildings house ancestor figures. The exteriors of these ritual buildings are ornamented with carved wooden figures and posts, the interiors with woven mats and painted bark panels on the walls. Since the 1960s, spurred by a rising nationalist spirit and the growth of cities such as Brazzaville, Lubumbashi (Élisabethville), and Kinshasa (Léopoldville), there has been a resurgence of architectural energy in Africa. Most of the new architecture is by European firms that visualize the potential of these Congolese countries in terms of urban development and architectural landmarks.

The role of architecture in ceremonial tradition

Lower Congo (Kongo or Bakongo) cultural area. In the Lower Congo (or Kongo) area three substyles can be identified: the areas known as the coastal region, the Kwango River area, and the Teke (Bateke) region.

Seated mother-and-child figures, found throughout the Lower Congo region, have been interpreted by some scholars as reflections of Christian influence introduced by Portuguese missionaries who came to the area in the late 15th century and left two centuries later, having failed to shake the Congolese belief in the spiritual powers of the souls of ancestors. It is highly improbable that the Congolese mother-and-child figures reflect the influence of the missionaries' teachings; they almost certainly have been inspired by native religious beliefs in ancestor veneration and the continuity of the life-force. Stylistically, the figures are a combination of realism and abstraction. Occupying the space of the cylinder of wood from which they were carved, the bodies of mother and child are organically united, conveying tenderness, restraint, and vitality. It is in the studied poses of the figures and the traditional arrangement of body parts as well as in the typically enlarged and exaggerated heads that the realistic forms take on an attitude of abstraction and conventionalized stylization.

Use of the human figure

The human figure is used by the peoples of the Lower Congo in the decoration of almost every work—from ceremonial objects and domestic utensils to pieces of furniture and architectural ornament (such as the pillars of huts). In a narrative style of decoration, Lower Congo figures are shown in a greater variety of attitudes than are the figures characteristic of other areas of the Congo; and, because of their imaginative rendering of movement and posture, this Lower Congolese style is less marked by the frontality and symmetry typical of other Congolese art. Although the majority of carved figures are made of wood, many important pieces in metal and ivory have been found. Among them are numerous metal figures clearly influenced by the Portuguese missionaries—statuettes of Christian saints, for example. In addition to the figures, crucifixes were also produced, in brass or

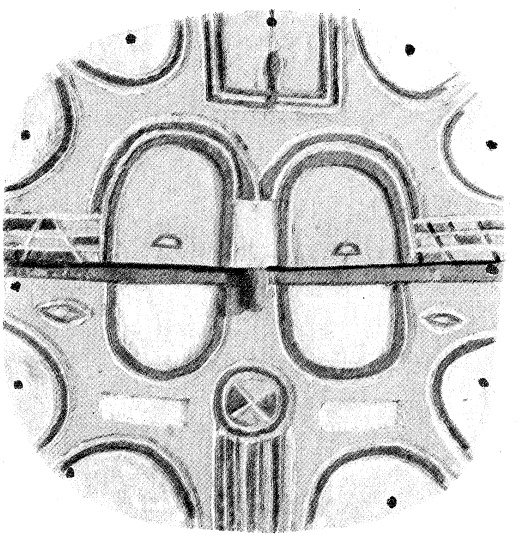


Figure 37: Teke (Bateke) mask, painted wood, Teke tribal region, Lower Congo cultural area. In the Musée de l'Homme, Paris. Diameter 34.9 cm.
By courtesy of the Musee de l'Homme, Paris



Figure 38: Kongo nail fetish, coastal region, Lower Congo cultural area. In the René Vander Straete Collection, Brussels. Height 69.9 cm.
Louis Loose

bronze (using the lost-wax, or cire perdue, method of casting, which employs a wax model). None of the vitality of African art was sacrificed when tribal metalsmiths began to work with European subject matter. Instead, the Christian symbols could be seen in a new and interesting light.

Another object common to the Lower Congo area, produced primarily by the coastal tribes, is a wooden pot lid carved with pictorial narratives representing proverbs. The pot lid, which covered the meal served by a wife to her husband, illustrated a particular complaint about their marital relationship—a wife's displeasure with her husband, for example; when that lid was used, the husband was obliged to discuss and resolve the problem publicly with the help of mealtime witnesses. This manner of family arbitration was traditional, and each woman was given a variety of carved pot lids on the occasion of her marriage.

Coastal cultures

Ancestor figures and fetishes carved by the Kongo, Sorongo (Basolongo), Woyo (Bawoyo), Sundi (Basundi), Bwende (Babwende), and Vili (Bavili) peoples, who live along the coast and in the Mayombe forest of the Lower Congo, are more realistically expressive than the figures of other areas of the Congo. Every detail rendered, the deceased ancestor is portrayed standing, seated, or kneeling, each attitude revealing the dignity and pride with which he is viewed. The fetishes are less realistically portrayed (Figure 38); although the head is treated in great detail, the arms and legs are stylized, appearing to be of equal size, and often the sex of the figure is not indicated; the stance of the figure is usually symmetrical. Whereas the ancestor figure typically appears serene and calm, the countenance of the fetish differs according to its function, which can be protective or malevolent. Some, the *moganga* figures, exude the same peaceful quality as the ancestor figures; others, the *ngonde* (*konde*) figures, exhibit a furious, malefic anger.

The *nkongi*, a group of fetishes characteristic of the coast and the Mayombe forest, consist mainly of human figures; but there are some that combine the forms of a dog and a leopard, sometimes with two heads. The *nkongi* fetish is often completely covered by nails and other sharply pointed metal objects driven into its surface; these objects mark each appeal made to the spirit embodied in it. All fetishes, whether they represent humans or animals, whether made of wood, horn, ivory, or even calabash, must contain a number of magical substances such as blood, animal claws and feathers, nails, and grains. These ingredients, called *bilongo*, are placed in a cavity, usually in the figure's stomach but sometimes in the back or head. The opening of the cavity is covered by a shell or, in some modern fetishes, by a piece of mirror imported from the Europeans. The magical substances are believed to invest the fetish figure with power and make it possible for the devotee to establish contact with the spirit (*nkisi*).

Kwango River area

The Kwango River area is the home of the Yaka (Bayaka), Suku (Basuku), Mbala (Bambala), and Pende (Bapende), whose masks, figures, and other carved objects show a dynamic stylization (Figure 39). Characterized by geometric patterns formed by the relationship of stylized body parts, Yaka figures lack the organic integration of naturalistic forms typical of the far less abstract figures produced by the neighbouring Kongo coastal tribes. The turned-up (*retroussé*) nose is a characteristic of Yaka figures and masks. Large, life-size carved figures stand at the entrances of Yaka initiation huts, the inside walls of which are covered with painted bark panels. *Tudansi* masks, worn by the young men at their initiation into manhood and decorated with polychrome and raffia collars, are topped with animal figures. The dramatically painted *kakungu* mask worn by the leader of the initiation rite represents a gaunt face with exaggerated nose and cheeks. This mask embodies terrific powers and is kept in its own hut. Similar to the Yaka *tudansi* mask is the *hemba* mask of the nearby Suku, which is only slightly less grotesque. Carved Suku figures show more rounded forms than do the Yaka, and the arms are characteristically attached directly to the squared shoulder blades.

Mbala figures have three different types of faces: elongated, wide, and lozenge-shaped. The features (especially the forehead and chin) project forcefully, and the head is surmounted by a crestlike coiffure. Mbala mother-

By courtesy of the Museum Rietberg, Zurich, von der Heydt Collection



Figure 39: Pende chief's chair, wood and leather, Kwango River area, Lower Congo cultural area. In the Museum Rietberg, Zürich. Height 97 cm.

and-child figures are much more powerfully rigid than others in the Congo.

Pende masks, made in a realistic style, are among the most dramatic works of all African art. Like the Yaka, small Pende masks fit over the head, helmet style. Symbolizing the mysterious powers to which tribesmen are introduced at the time of their initiation, Pende masks are worn in comic entertainments performed during the ceremonies. The masks have facial forms that repeat the angular pattern established by the heavy triangular eyelids, and they are topped by a bushy coif of raffia. Smaller versions of these masks are made as amulets in ivory or wood. Other masks, made of raffia, are used to ward off women and children, as the initiation ceremonies are for men only. Like the Kongo coastal tribes, the Pende fashion their figures in a style identical to that of their masks. One type of figure, called *tungunlungu*, representing the female ancestry of the tribe, is placed in front of the chief's house.

Teke cultural region

The Teke tribes of Congo (Brazzaville) live on the banks of the Congo River. They are best known for their fetishes, called *butti*, which serve in the cult of a wide range of supernatural forces sent by the ancestors, who are not worshipped directly. Each figure has its own specific purpose, not related directly to its appearance. When a figure is carved for a newborn child, part of the placenta is placed in the stomach cavity of the figure while the rest is buried inside the father's hut (where the family's fetish figures are kept). The figure serves to protect the child until puberty. Figures of identical appearance serve also for success in hunting, trading, and other activities, each figure's specific purpose being known only to the owner. Another type of figure, called *nkiba*, without magical substances, apparently functions in a similar way, but a special ritual is needed to activate it. The angular geometric style and linear ornamentation of these rigid Teke figures is much more cubistic than realistic. Teke face masks, flat discs painted in bright polychrome, are highly abstracted forms bearing no naturalistic associations. Other carved objects, textile designs, and pottery are produced with the same degree of precision.

Kuba (or Bakuba) cultural area. The people who live in this central area of the Congo called themselves Bushongo, meaning Knife Throwers; and they were called Bakuba, People of Lightning, by neighbouring Luba tribes. Today, they are called Kuba, as is the whole cultural area. The kingdom of Kuba was as important in the history of the Congo as the kingdoms of Ife and Benin were to the history of Nigeria (see above *West African cultures*). The art of the Kuba is one of the most highly developed of all African visual arts, and significant cultural accomplishments are part of their heritage. Mucu Mushanga, their 27th king, was credited with the invention of fire, and he was the first to make clothing out of bark cloth. Shamba Bolongongo (*c.* 1600), the 93rd king, who introduced weaving and textile manufacture to his people, was also the first Congolese ruler to have his portrait carved in wood. Shamba Bolongongo's portrait established a tradition of such portraiture among the Kuba people. Originally, these statues were conceived as realistic portraits of individual kings; later, the simplification of certain forms led to a more conventionalized style. While the simplification led to a loss of individualism essential in portraiture, it established a compelling stylistic continuity from archaic to modern. The kings typically sit facing forward, with legs crossed, the left in front of the right; the right hand, with fingers extended, rests on the right knee and the left hand holds the royal dagger. Tattoo-like, stylized patterns cover the stomach and are continued on the back of the figure. The sculptures also include objects significant to each particular king, identifying his own personal accomplishments. Developing from the court style was a popular style, which utilized geometric forms instead of the well-modelled, full-volumed forms of the court figures. Kuba fetishes, emphasizing only essential organs, are highly stylized.

Development of a popular style

The popular style can also be found in the utensils and textiles produced by the Kuba. Four types of cloth are used for garments: plain (unadorned) cloth; cloth with ornamentation woven in; elaborately decorated cloth with soft pile (called Kasai velvet); and cloth decorated with dyed patterns (Figure 40). The ornamentation consists of highly complicated angular motifs especially

Frank Willett

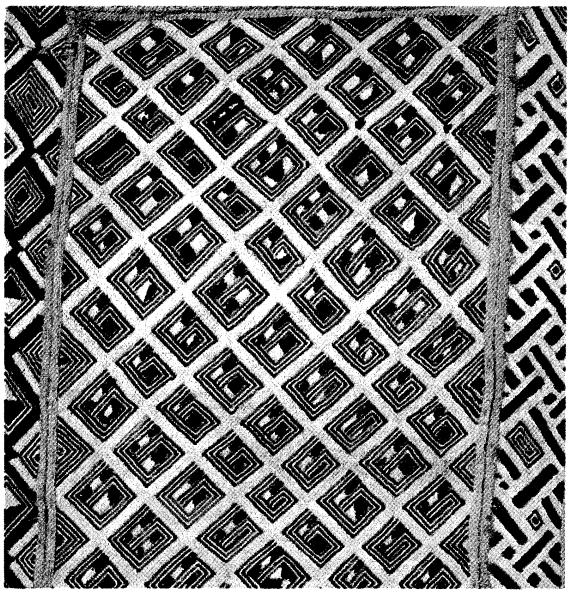


Figure 40: Kuba raffia pile cloth, Kuba cultural area. In the Hampton Institute, Virginia.

rectilinear interlace designs. The same repetitive geometric designs are used to decorate the handles and bases of "effigy cups," which are shaped like human heads and figures, and the surfaces of square, rectangular, and crescent-shaped wooden boxes. The Kuba metalsmith—so highly respected that his children were required to learn the same art—worked with copper, iron, and brass, making weapons and tools to be admired as well as used. In some cases, one metal was inlaid with another. *Mash-amboy*, or *mokenge*, masks (Figure 41), made of raffia and decorated with shells, beads, and even bells and feathers, were traditionally used by the chief to frighten women into obedience. They are now worn solely for dance entertainments. Another mask, the *bombo*, traditionally designed to symbolize the Pygmies and worn in initiation rituals, is a wooden helmet covered with copper that is ornamented with all-over linear patterns. The *shene-malula* mask, partially covered with shells and beads in geometric patterns, was worn to introduce members into a secret police society.

Luba cultural area. The tribal styles of the Luba cultural area are the Luba, Buli (Oli), Songe, Lulua, Kanioka, Jokwe, and Lunda. The Luba are the dominant cultural influence in this southeastern area of the Congo. The history of the Luba kingdom, founded in the 15th century, is filled with violent wars in which the Songe and Jokwe conquered and were conquered by the Luba and Lunda tribes, which led to the mixture of cultural traditions that make up the Luba people of today. Although the history of the Luba people is one of violence and warfare, their artistic style is characterized by harmonious integration of organically related forms by the most subtle expressions of balance and unity. Female figures are carved more often than are male figures. Some are freestanding, almost always in a frontal position with their hands on their breasts; others are kneeling, sitting, or standing figures whose upraised hands serve as supports for bowls, seats, and neck rests. A popular form consists of a kneeling or sitting female figure holding a bowl. Such mendicant figures are used to appeal to spirits for health and aid for pregnant women; neighbours, seeing the figure in front of a woman's hut, will fill it with gifts to help her avoid hardship in pregnancy. The

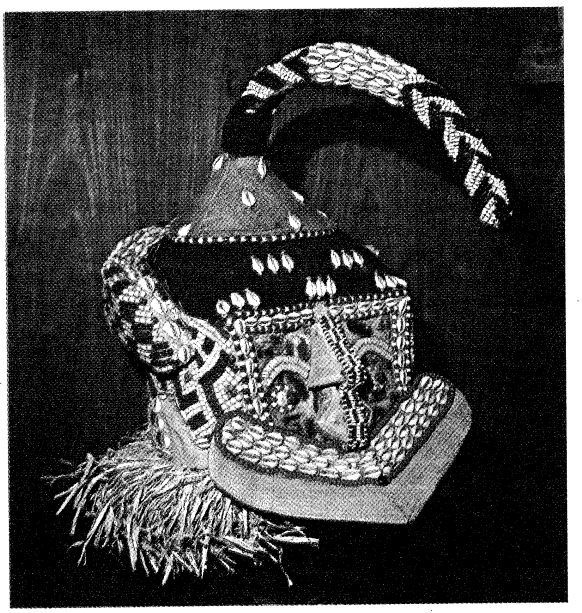


Figure 41: Kuba *mashamboy* mask, fibre, shells, and beads, Kuba cultural area. In the Hampton Institute, Virginia. Height 41 cm.
Frank Willett

female figures are modelled in rounded forms and have what is called *dodu;* that is, a stylistic tendency toward plumpness. The bodies are covered with a variety of tattoo patterns, on the basis of which tribal substyles can be differentiated. Although the naturalistic effect of these figures is simplified, it remains organically realistic.

"Long-face style" of the Buli

The major substyles are the classic Luba (discussed above) and the Buli style of the eastern area. What one scholar calls the "long-face style" of the Buli is distinctly unlike any other style in the Luba area, and it contrasts strongly with the realism found in the classic Luba figures. The features shown are more Hamitic than Negroid, the faces elongated, with angular, elegant features arising out of a geometric, cubist severity; angular cubist forms exaggerate the formal construction of the body, which is usually posed symmetrically. This style is exemplified in a stool (Figure 42; Musée Royal de l'Afrique Centrale, Tervuren, Belgium) made for a chief by the "Master of Buli"; female figures support the seat with their heads, which are exaggerated out of proportion and are stylized.

The Songe, members of a tribe who conquered and were conquered by the Luba, created a sculptural style of intense dynamism and vitality. Their figures—mostly male because they were a patrilineal people—are always symmetrical and shown in frontal positions. The style of their fetishes, carved from wood or horn and decorated with shells and polychrome, is not as realistic as the classic Luba style, and their integration of non-naturalistic, more geometric forms is impressive. The Songe also produce ceremonial axes, made of iron and copper and decorated with interlaced patterns. A subtribe of the Songe, the Kalebwe, are known for their *kifwebe* masks, which combine human and animal features. The masks, painted in either red and black or black and white, are meant to be seen in motion, their colourful forms interacting with the drum or vocal accompaniments.

The Lulua developed a style that epitomizes the Congolese expression of simplification and stylization. It is reflected in their compelling polychrome masks and in the elegant vitality of their freestanding figures, which show an emphatically symbolic treatment of form. The Lulua also made tobacco mortars carved in the form of a crouching human.

Kanioka figures are elongated, like those of the Lulua. Unusually asymmetrical, they are shown engaged in the various acts of daily life. The major distinctive feature of the Kanioka figures is their netted or crisscrossed hairdo.

In the 19th century, the Jokwe and Lunda conquered the Luba kingdom; today these hunters and farmers live in an area that includes part of northern Angola as well as southern Zaire. Their styles are often indistinguishable from one another. The forms they create are monumental and weighty, and both male and female figures are carved in an impressively vigorous style. Their large size in no way inhibits the carving of detail; the figures even have fingernails. Also made by these peoples are chairs decorated with figures posed in genre and legendary scenes. Zoomorphic motifs are found on all objects—even utensils such as combs and knives. In ceremonial rites of initiation, tribesmen wear painted barkcloth masks and net costumes.

Northern cultural area. The tribes of the northern Congo live in an area of rain forest, where there are comparatively few centres of artistic and cultural activity. Unlike the majority of peoples in the southern half of the Congo, the northern tribal systems are ordered on a patrilineal system and exhibit an affinity for pole sculpture.

Rega stylistic forms

The Rega (Lega), who inhabit the area between the Luba and the northernmost tribes, have produced an art somewhat akin to both traditions. Their figures and masks, largely carved from ivory, lack the subtle organic realism of those of the southern Congolese tribes. Rega forms are quite abstract, the simplification expressive but not smoothly developed. This simplicity of non-naturalistic forms creates what one scholar calls "reality in non-reality."

The influence of southern tribal subject matter is reflected in the postures of some Rega figures. In addition to a dance mask worn in front of the face, the Rega wear on the sides of the head small masks, which are generally white, with raffia beards.

Northeast cultural area. In the northeast area live tribes of Sudanic derivation: the Mangbetu and the Azande. Mangbetu pottery jars and carved ivory utensils are strongly decorative. The stylization of naturalistic forms is imaginative, as is their decorative use of the human figure on almost every object (Figure 43); the Mangbetu practiced skull deformation, and even the fibre-bound heads of children have been represented in



Figure 42: Luba chief's stool by the "Master of Buli," Buli "long-face style," wood, Luba cultural area. In the Musée Royal de l'Afrique Centrale, Tervuren, Belgium. Height 53.3 cm.
By courtesy of the Musee Royal de l'Afrique Centrale, Tervuren, Belgium

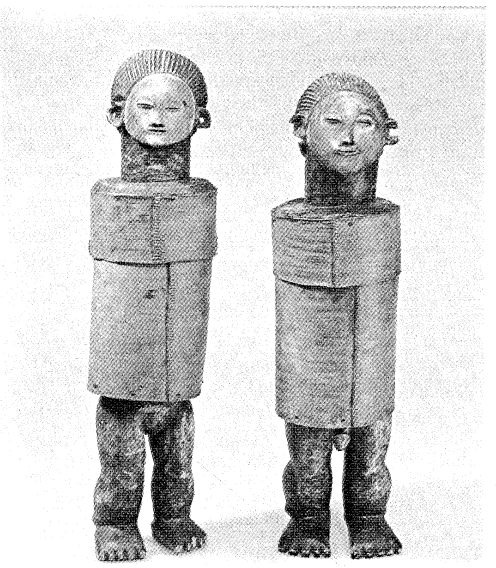


Figure 43: Mangbetu or Azande storage containers in human form, bark and wood, Northeast cultural area. In the Museum Rietberg, Zürich. Height 69.4 cm.
By courtesy of the Museum Rietberg, Zurich, von der Heydt Collection

sculpture. Their costumes are made from decorated bark and raffia. Some musical instruments made by the Azande resemble a woman's head on a long, necklike pipe. Their daggers are decorated with male heads, and box lids are frequently carved in the form of human heads. Only a few Azande figures exist, in wood and terra-cotta, and only a few crude masks have been identified as Azande. *Yanda* figures, used by members of the Azande secret Mani society, are extremely abstract; some even lack arms and have only stump legs. Made of clay or wood and often decorated with beads or brass, the *yanda* figures are believed by members of the Mani society to possess tremendous magical powers.

Northwest cultural area. The northwest Congo is a swampy river bed inhabited by non-Bantu tribes. A stylistic characteristic of the figures and masks carved by these tribes is a vertical plane of tattooed dots extending down the forehead to the nose. The Ngbaka (or Bwaka) and the Ngbandi are the tribes of major artistic importance. There is no single Ngbaka sculptural style: at times, the figures are fleshy and rounded; at other times they are considerably more angular. Small animal figures are used as fetishes in hunting. The masks used in circumcision ceremonies are roughly executed. Both the Ngbaka and the Ngbandi make clay images to be used in funeral rituals. The Ngbandi are also known for wooden fetishes and figures. Small carved ivory or wood figures were worn by Ngbandi warriors, who carried shields made of decorated woven fibre. It is often impossible to distinguish the few Ngbandi masks from those of the Ngbaka.

Gabon. West of the Congo, also in the equatorial rain forest of Central Africa, is the country of Gabon. Three major groups live in Gabon: the Pangwe, a people made up of several tribes including the Fang; the Ogowe group, consisting of the Mashango, Ashira, Galoa, and Mpongwe (also known as the Sumbo); and the Kota. The tribal life and art of these peoples is similar to that of the Congolese in many ways, suggesting a certain amount of interaction between the tribes.

Fang masks and figures

Fang masks and figures are characterized by simplicity and abstraction. Typical of Fang work are *bieri* boxes containing the skull and bones of deceased ancestors carved with an abstract figure intended to symbolize the spirit of the ancestor. Fang masks, such as those worn for hunting and punishing sorcerers by members of the secret Ngi society, are the most abstract and formally simple of all Fang art. The carved wooden face is painted white with kaolin, and facial features are outlined in black; the long line of the nose, the tiny eyes and mouth, and the unkempt raffia beard add to the foreboding quality of the mask.

The art of the Ogowe tribes, particularly the Mpongwe, is closely tied to tribal death rituals. Their masks, painted white to symbolize death, represent dead female ancestors, though they are worn by male relatives of the deceased.

Kota masks resemble those used by the Ogowe group. The Kota also create stylistically unique reliquary figures, called *mbulu-ngulu*, which are covered with a sheet of brass, iron, or copper. Like the Fang, the Kota keep the skulls and bones of ancestors in containers, which consist, in this tribe, of a basket surmounted by a carved figure. Evocation of the spirit housed in the figure is effected by hand clapping, after which a sacrificial offering is made so that the spirit will grant advice to the tribesmen. (P.Wa.)

BIBLIOGRAPHY. Hitherto, very little attempt has been made in the literature to bring together the arts of Africa as a whole, or even those of a single community. There is so far only one journal expressly devoted to all the arts of Africa: *African Arts*, published quarterly by the African Studies Center, University of California, Los Angeles, California. This specializes in fairly short articles of general interest and is well illustrated in colour and in black and white. More modest in scope but also wide-ranging though with an emphasis on literature was *Black Orpheus* (Ibadan, 1957–66). *Présence Africaine* (Paris) although primarily literary has articles on the other arts too. JOHN N. PADEN and EDWARD W. SOJA (eds.), *The African Experience* (1970), is a multidisciplinary approach to African studies. Vol. 1 includes essays on the separate arts; vol. 2 is a syllabus of lecture outlines; vol. 3a is an annotated bibliography grouped by lecture topic and vol. 3b a guide to resources. ROBERT PLANT ARMSTRONG, *The Affecting Presence: An Essay in Humanistic Anthropology* (1971), explores the aesthetic base of the Yoruba arts in a philosophical rather than a sociological manner. ULLI BEIER, *Contemporary Art in Africa* (1968), provides a most useful account of contemporary painting and sculpture, with some reference to the other arts. Otherwise information about the arts has to be culled from a variety of journals concerned either with African studies as a whole: *Odu* (semi-annual); *African Notes* (quarterly); *Ibadan* (3/year); *Africa; Notes Africaines* (quarterly); *Bulletin de l'Institut Fondemental d'Afrique Noire* (quarterly); and *Research Review* (quarterly); or with part of the continent only: *Nigeria Magazine* (quarterly); *Sudan Notes and Records* (semi-annual); *Tanganyika Notes and Records* (1936–55); *The Uganda Journal* (1934–45); and *Zaïre* (1946–61); or in journals devoted to a discipline: *Ethnomusicology* (3/year); *Journal of the International Folk Music Council* (annual); *Objets et Mondes* (quarterly); *Man* (1901–65); *Journal of the Royal Anthropological Institute* (quarterly); and *Zeitschrift für Ethnologie* (irreg.).

Literature: There is no general survey of African literature, but many anthologies contain useful introductions. The best general survey is ULLI BEIER (ed.), *Introduction to African Literature: An Anthology of Critical Writing from Black Orpheus* (1967). See also JANHEINZ JAHN, *Muntu: Umrisse der neoafrikanischen Kultur* (1958; Eng. trans., *Muntu: An Outline of Neo-African Culture*, 1961), and *Geschichte der neoafrikanischen Literatur* (1966; Eng. trans., *A History of Neo-African Literature*, 1968), covering also Afro-American, Caribbean, and American Negro culture and writings; and JOHN A. RAMSARAN, *New Approaches to African Literature* (1965). *(Bibliographies)*: JANHEINZ JAHN, *A Bibliography of Neo-African Literature from Africa, America and the Caribbean* (1965); MARGARET AMOSU, *A Preliminary Bibliography of Creative African Writing in the European Languages* (1963); J.M. CAMERON, *Pan Africanism and Négritude: A Bibliography* (1964). *(Oral traditions)*: An important early collection is LEO FROBENIUS, *Atlantis*, 12 vol. (1921–28), myths, folktales, sagas, fables, etc., in German translation. See also his selection of folktales, *African Genesis* (1937); MARCEL GRIAULE, *Dieu d'eau* (1948); MELVILLE J. and FRANCES S. HERSKOVITS, *Dahomean Narrative* (1958); GERMAINE DIETERLEN (ed.), *Textes sacrés d'Afrique noir* (1965); ULLI BEIER, *African Poetry* (1966), *The Origin of Life and Death* (1966), and with BAKARE GBADAMOSI, *Not Even God Is Ripe Enough* (1968); PIERRE VERGER, *Notes sur le culte des Orisa et Vodun à Bahia . . .* (1957); NGUMBU NJURURI, *Agikuyu Folk Tales* (1966); volumes in the "Oxford Library of African Literature," notably W.H. WHITELEY, *A Selection of African Prose*, vol. 1, *Traditional Oral Texts* (1964); H.A.S. JOHNSTON, *A Selection of Hausa Stories*

(1966); JOHN S. MBITI (ed.), *Akamba Stories* (1966); and E.E. EVANS-PRITCHARD (ed.), *The Zande Trickster* (1967). (*Modern literatures—anthologies*): The most wide-ranging of many collections since World War II are EZEKIEL MPHAHLELE (ed.), *African Writing Today* (1967); ULLI BEIER and GERALD MOORE, *Modern Poetry from Africa*, 2nd ed. (1968); and OSCAR R. DATHORNE and WILLFRIED FEUSER, *Africa in Prose* (1969). See also PEGGY RUTHERFOORD (ed.), *Darkness and Light* (1958); EZEKIEL MPHAHLELE and ELLIS A. KOMEY, *Modern African Stories* (1964); W.H. WHITELEY, *A Selection of African Prose*, vol. 2, *Written Prose* (1964); RICHARD RIVE, *Modern African Prose* (1964); OLUMBE BASSIR, *An Anthology of West African Verse* (1957); HENRY SWANZY (ed.), *Voices of Ghana* (1958); FRANCES ADEMOLA (ed.), *Reflections: Nigerian Prose and Poetry* (1962); D. COOK, *Origin East Africa: A Makerere Anthology* (1965); LEOPOLD S. SENGHOR (ed.), *Anthologie de la nouvelle poésie nègre et malgache* (1948); and *Nouvelle Somme de poésie du monde noir* (1966); LEONARD SAINVILLE, *Anthologie de la littérature négro-africaine*, 2 vol. (1963–68); CLIVE WAKE, *An Anthology of African and Malagasy Poetry in French* (1965); MARIO DE ANDRADE, *Antologia da poesia negra de expressão portuguesa*, prefaced by the essay, *Cultura negro-africana e assimilacão* (1958). (*Critical works*): EZEKIEL MPHAHLELE, *The African Image* (1962); CLAUDE WAUTHIER, *L'Afrique des Africains* (1964; Eng. trans., *The Literature and Thought of Modern Africa*, 1966); ROBERT W. JULY, *The Origins of Modern African Thought* (1968); ANTHONY C. BRENCH, *The Novelist's Inheritance in French Africa* (1967); and *Writing in French from Senegal to Cameroon* (1967); THOMAS MELONE, *De la Négritude dans la littérature negro-africaine* (1962); LILYAN KESTELOOT, *Les Écrivains noirs de langue française*, 3rd ed. (1967); GEORGE MOORE, *Seven African Writers* (1962); and *The Chosen Tongue* (1969); JUDITH GLEASON, *This Africa: Novels by West Africans in English and French* (1965). (*Bantu literature*): H. DIETERLIN and F. KOHLER, "Les Bassoutos d'autre fois," and "Les Bassoutos d'aujourd'hui," in *Livre d'or de la mission du Lessouto* (1912); G.H. FRANZ, "The Literature of Lesotho," *Bantu Studies* (1930); A.C. JORDAN, "Towards an African Literature," 12 papers on literature in indigenous African languages spoken in the Republic of South Africa and neighbouring regions, in *Africa South* (1957–60); DANIEL P. KUNENE and RANDAL A. KIRSCH, *The Beginnings of South African Vernacular Literature* (1967); R.H.W. SHEPHERD, *Bantu Literature and Life* (1955); JAMES STEWART, *Dawn in the Dark Continent; or, Africa and Its Missions* (1903). (*Yoruba literature*): ULLI BEIER and BAKARE GBADAMOSI (comp. and trans.), *Yoruba Poetry* (1959); S.A. BABOLOLA, "La Poésie Yoruba," in *Présence Africaine*, 47:211–217 (1963); and *The Content and Form of Yoruba Ijala* (1966). (*Swahili literature—epic poetry*): J.W.T. ALLEN (comp.), *Tendi* (1971), the great work by the foremost authority on Swahili; ERNST DAMMANN, *Dichtungen in der Lamu-Mundart des Suaheli* (1940), still the most important work on Swahili epic poetry; JAN KNAPPERT, *Swahili Sagas and Epic Poetry* (1973), summarizes the present state of our knowledge on the subject. (*Lyrical poetry*): JAN KNAPPERT, *A Choice of Flowers* (U.S. title, *An Anthology of Swahili Love Poems;* 1972). (*Didactic poetry*): JAN KNAPPERT, *Traditional Swahili Poetry* (1967); and *Swahili Islamic Poetry*, 3 vol. (1971–72). (*Traditional stories*): JAN KNAPPERT, *Myths and Legends of the Swahili* (1970); and *Three Centuries of Swahili Verse* (1973), deals with the poetry in all its aspects from the beginning till the present day. There is no comprehensive work on Swahili prose literature. (*Proverbs and sayings*): WILLIAM E. TAYLOR, *African Aphorisms; or, Saws from Swahili-Land* (1891), still the only work of size on the subject, and easily the best.

Music: E. AMU, *Twenty-five African Songs in the Twi-Language* (1932), original compositions in the African style that set the tone for the modern development of African music in Ghana; FRANCIS BEBEY, *Musique de l'Afrique* (1969), a popular description of African music and instruments, mostly West African, with many illustrations; ULLI BEIER, "Yoruba Folk Operas," *African Music*, 1:32–34 (1954), mention of the birth of a new Yoruba tradition. (*Nigeria*): JOHN BLACKING, *Venda Children's Songs* (1967), a thorough social and musical description and analysis of Venda children's songs that provides a key to the understanding of other Venda music. (*South Africa*): JOHN BLACKING, "Songs, Dances, Mimes and Symbolism of Venda Girls' Initiation Schools," *African Studies*, 28:3–35 (1969), sociological and musical analysis of a major Venda ritual; PETER COOKE, "Ganda Xylophone Music: Another Approach," *African Music*, vol. 4, no. 4 (1970), stresses the need for understanding the songs behind instrumental performances on the *amadinda* xylophone. (*Uganda*): MAURICE DJENDA, "Les Pygmées de la Haute Sangha," *Geographica*, 4:27–43 (1968), brief description of musical practices and instruments, illustrated in colour. (*Central African Republic*): NICHOLAS M. ENGLAND, "Bushman Counterpoint," *Journal of the International Folk Music Council*, 19:58–66 (1967), description of polyphony and the independence of vocal lines in Bushman song. (*South West Africa and Botswana*): AKIN EUBA, "Multiple Pitch Lines in Yoruba Choral Music," *Journal of the International Folk Music Council*, 19:66–71 (1967); A.M. JONES, "African Drumming," *Bantu Studies*, 8:1–16 (1934); and *African Music in Northern Rhodesia and Some Other Places*, rev. ed. (1958), introduces some basic concepts of African rhythm, including multiple main beat and description of some Zambian musical instruments; *Studies in African Music*, 2 vol. (1959), a comprehensive discussion of African rhythm, singing, and dancing, with special but not exclusive reference to the Ewe people in Ghana, and the Lala in Zambia—vol. 2 contains the full score of seven dances and also children's and fishing songs; and *African Rhythm* (1961), a general survey with musical scores of songs with their background rhythms and of drumming in Zambia; PERCIVAL R. KIRBY, *The Musical Instruments of the Native Races of South Africa* (1953), the standard reference work on South African musical instruments; and "Physical Phenomena Which Appear To Have Determined the Bases and Developments of an Harmonic Sense Among Bushmen, Hottentot and Bantu," *African Music*, 2:5–9 (1961); GERHARD KUBIK, "The Structure of Kiganda Xylophone Music," *African Music*, 2:6–30 (1960); "Xylophone Playing in Southern Uganda," *Journal of the Royal Anthropological Institute*, 94:138–159 (1964); and "Composition Techniques in Kiganda Xylophone Music," *African Music*, 4:22–72 (1969); A.A. MENSAH, "The Impact of Western Music on the Musical Traditions of Ghana," *The Composer*, 19:19–22 (1966), complete analyses of an African instrumental technique, including repertoire. (*The amadinda xylophone of Uganda*): GERHARD KUBIK, "The Phenomenon of Inherent Rhythms in East and Central African Instrumental Music," *African Music*, 3:33–42 (1962), article in which the concept of inherent rhythms was introduced and demonstrated for the first time; "The Traditional Music of Tanzania," *Afrika*, vol. 8, no. 2 (1967); *Mehrstimmigheit und Tonsysteme in Zentral und Ostafrika* (1968), an important work introducing the basic concepts of African harmony and vocal polyphony; and *Musica Tradicional e Aculturada dos !Kung' de Angola* (1970), description and analysis of the !Kung' bushmen of southern Angola; J.H.K. NKETIA, "Modern Trends in Ghana Music," *African Music*, 1:13–17 (1957), a description of the musical situation in Ghana by the leading African musicologist; and *African Music in Ghana* (1962), a general introduction to the music of Ghana—the standard work. (*Tanzania*): EDUARD PECHUEL-LOESCHE, *Die Loango-Expedition* (1907); DAVID RYCROFT, "Nguni Vocal Polyphone," *Journal of the International Folk Music Council*, 19:88–103 (1967), a description of Nguni singing (South Africa) showing its circular form, non-simultaneous entry; CURT SACHS, *The History of Musical Instruments* (1940), standard work on world musical instruments, from which the classification of African instruments has been derived; ANDREW TRACEY, "The Mbira Music of Jege A. Tapera," *African Music*, 2:44–63 (1961); "The Matepe Mbira Music of Rhodesia," *African Music*, vol. 4, no. 4 (1970), analysis of how to make and play the *matepe*, a Rhodesian *mbira;* HUGH TRACEY, *Chopi Musicians* (1948, reprinted 1970), a player's instruction manual for another type of Rhodesian *mbira;* K.M. TROWELL and K.P. WACHSMANN, *Tribal Crafts of Uganda* (1953), standard work on Ugandan material culture, containing a long section on musical instruments; ERICH VON HORNBOSTEL, "The Ethnology of African Sound-Instruments," *Africa*, 6:129–157 (1933); W.E.F. WARD, "Music in the Gold Coast," *Gold Coast Review*, vol. 3, no. 2 (1927). (*Bibliographies*): L.J.P. GASKIN, *A Select Bibliography of Music in Africa* (1965); A.P. MERRIAM, "An Annotated Bibliography of African and African-Derived Music Since 1936," *Africa*, 21:319–329 (1951); D.L. THIEME, *African Music: A Briefly Annotated Bibliography* (1964); D.H. VARLEY, *African Native Music: An Annotated Bibliography* (1936, reprinted 1970).

Dance and theatre: There is little opportunity to witness the performing arts of Africa outside the continent itself, though dance troupes (such as Ballets Africains) and theatre groups (such as Natal Theatre Workshop Company, from South Africa) occasionally visit the West. Literature on drama and dance is meagre, especially in the English language. Of use to those generally interested in Negro-African writing (including drama), is JOHN A. RAMSARAN, *New Approaches to African Literature* (*op. cit.*). This contains a bibliography of plays, published and unpublished, by African dramatists of today. For anyone approaching 20th-century

African drama for the first time, COSMO PIETERSE (ed.), *Ten One-Act Plays* (1968), is a good introduction. The dramatists represented in this work come from West, East, and South Africa. On traditional theatre and dance forms, KEITA FODEBA, "La Danse africaine et la scène," *Présence Africaine*, 14–15:202–209 (1957), is of special interest; its author is primarily associated with the Ballets Africains. (*Dance*): GERTRUDE P. KURATH, "Panorama of Dance Ethnology," *Current Anthropology*, 1:233–254 (1960), a clear account of the training, techniques, and aims of dance ethnologists; CURT SACHS, *Eine Weltgeschichte des Tanzes* (1933; Eng. trans., *World History of the Dance*, 1937, reprinted 1963), a general history of the dance, with valuable descriptions of themes, styles, forms, and types of dance from earliest times to the present; WILFRID DYSON HAMBLY, *Tribal Dancing and Social Development* (1926), contains material on dances of South, East, and West Africa; GEOFFREY GORER, *Africa Dances* (1935, reprinted 1962), a somewhat discursive study of the dances of West Africa, especially of Dahomey; HUGH TRACEY, *Chopi Musicians: Their Music, Poetry and Instruments* (1948, reprinted 1970), an excellent descriptive account; and *African Dances of the Witwatersrand Gold Mines* (1952), a copiously illustrated description of Bantu dances, as performed for recreation by workers in the gold mines of South Africa; ARCHIBALD N. TUCKER, *Tribal Music and Dancing in the Southern Sudan (Africa), at Social and Ceremonial Gatherings* (1933), a useful study of music and rhythms in the dance; ROSE BRANDEL, *The Music of Central Africa* (1961), containing a section on mimetic and abstract patterns in the dance. (*Drama*): Very few works exist on the drama of Africa as a special study, and most of them are in French. Apart from the necessarily brief selection listed here, additional descriptive material can often be found in more general works by anthropologists, cultural historians, and travellers. BAKARY TRAORE, *Le Théâtre Négro-Africain et ses fonctions sociales* (1958), the basic study; WILLIAM RIDGEWAY, *The Dramas and Dramatic Dances of Non-European Races in Special Reference to the Origin of Greek Tragedy* (1915); and LOOMIS HAVEMEYER, *The Drama of Savage Peoples* (1916), both general works on the impulse and nature of dance-drama among primitive peoples, now dated; EDWARD G. PARRINDER, *African Traditional Religion*, 2nd ed. (1962), a lucid account of African religious rituals and their early occasions; J.C. MESSENGER, "Anang Art, Drama, and Social Control," *African Studies Bulletin*, 5:29–35 (1962), includes a fascinating and detailed account of the Ekong association of the village of Ikot Obong, Calabar Province, in South-Eastern State, Nigeria; JOSEPH GREGOR, *Masks of the World* (1937), on ritual disguises and transfiguration; PERCY A. TALBOT, *Life in Southern Nigeria* (1923, reprinted 1967), valuable for its description of a form of marionette theatre found in Nigeria; PIERRE VERGER, *Dieux d'Afrique* (1954), reveals the nature and forms of ritual and of mime.

Visual arts: The literature on the visual arts is too extensive to catalog in detail here. The following are references available in the English language, almost all of which have detailed bibliographies. The primary key to the available literature is L.J.P. GASKIN, *A Bibliography of African Art* (1965). General works that survey the varieties of art of the whole continent include WILLIAM R. BASCOM, *African Arts* (1967); ELIOT ELISOFON and WILLIAM B. FAGG, *The Sculpture of Africa* (1958), *Tribes and Forms in African Art* (1965), *African Tribal Images* (1968), and *African Sculpture* (1970); JEAN LAUDE, *Les Arts de l'Afrique noire* (1966; Eng. trans., *The Arts of Black Africa*, 1971); MICHEL LEIRIS and JACQUELINE DELANGE, *African Art* (1968); ELSY LEUZINGER, *Afrika: Kunst der Negervölker* (1959; Eng. trans., *The Art of Africa: The Art of the Negro Peoples*, 1960, reprinted 1967), *African Sculpture: A Descriptive Catalogue* (1963), *Die Kunst von Schwarz-Afrika* (1970; Eng. trans., *The Art of Black Africa*, 1972); ROY SIEBER and ARNOLD RUBIN, *Sculpture of Black Africa: The Paul Tischman Collection* (1968); KATHLEEN M. TROWELL, *Classical African Sculpture* (1954); FRANK WILLETT, *African Art: An Introduction* (1971); P.S. WINGERT, *The Sculpture of Negro Africa* (1950). The following deal with more restricted areas or subjects: P.A. ALLISON, *African Stone Sculpture* (1968); DANIEL P. BIEBUYCK (ed.), *Tradition and Creativity in Tribal Art* (1969); ULLI BEIER, *Contemporary Art in Africa* (1968); K.C. CARROLL, *Yoruba Religious Carving: Pagan and Christian Sculpture in Nigeria and Dahomey* (1967); W.B. FAGG, *Nigerian Images* (1963), *The Art of Western Africa* (1967), and *The Art of Central Africa* (1967); W. and B. FORMAN and P. DARK, *Benin Art* (1960); ADRIANUS A. GERBRANDS, *Art as an Element of Culture, Especially in Negro Africa* (1957); NEW YORK MUSEUM OF PRIMITIVE ART, *Bambara Sculpture from the Western Sudan* (1960); *Senufo Sculpture from West Africa* (1964); MARCEL GRIAULE, *Folk Art of Black Africa* (1950; British title, *Arts of the African Native*), a most useful account of the visual arts of the Dogon; LADISLAV HOLY, *Masks and Figures from Eastern and Southern Africa* (1967); ROBIN HORTON, *Kalabari Sculpture* (1965); J.D. LAJOUX, *Merveilles du Tassili n' Ajjer* (1962; Eng. trans., *The Rock Paintings of Tassili*, 1963); HENRI LHOTE, *À la découverte des fresques du Tassili* (1958; Eng. trans., *The Search for the Tassili Frescoes*, 1959); BRIGITTE MENZEL, *Goldgewichte aus Ghana* (1968; *Goldweights from Ghana*, the text is in English and German); MARGARET W. PLASS, *African Miniatures* (1967); M.W. SMITH (ed.), *The Artist in Tribal Society* (1961); KATHLEEN M. TROWELL, *African Design*, 3rd ed. (1971); WALKER ART CENTER, *Art of the Congo* (1967); A.R. WILLCOX, *Rock Paintings of the Drakensberg* (1956), and *The Rock Art of South Africa* (1963); FRANK WILLETT, *Ife in the History of West African Sculpture* (1967).

(F.Wi./H.Ty./G.K./A.M.J./A.M.D./D.P.K./U.Be./Ja.Kn./Ge.Mo./Ma.A.C./P.Wa./Si.M.)

African Peoples and Cultures

The continent of Africa, an area of almost 12,000,000 square miles (31,000,000 square kilometres), has over three-quarters of its land in the tropics. Its main climatic zones include the fertile, narrow Mediterranean littoral, with the Lower Nile Valley; the Sahara; the savanna zone along the southern borders of the desert, running from Senegal to Ethiopia and extending into the semi-arid areas of Somalia; the West African forest belt along the Atlantic coast, extending into the Congo Basin; the Ethiopian and East African highlands, extending into Rwanda; the East, Central, and southern African savannas; the fertile southern African coastline; and the island of Madagascar. There are few large mountain ranges, few large navigable rivers, and few natural harbours. Rainfall varies considerably from an annual precipitation of over 60 inches (1,500 millimetres) in parts of the tropical forest belt to under five inches (130 millimetres) in the various arid areas. Most of the African continent therefore comprises plateaus supporting savanna bush and grasslands, mostly in areas of relatively low fertility and traditionally with few easy long-distance routes of communication.

Basic African economies

In the most arid areas the basic economy has been and still is based upon hunting and gathering only, without agriculture or animal husbandry but usually with some form of petty trade with neighbouring agricultural groups. These hunting peoples are found today only in southwestern Africa (Bushmen), in parts of the Congo forest (Pygmies), and in a few parts of northeast Africa. At one time, they had a distribution that was considerably wider, but they have been driven by other groups into these refuge areas.

Pastoralism has been traditionally a basic livelihood mainly in the semi-arid regions along the southern edge of the Sahara (peoples such as the Fulani and Tuareg) and in the Horn of Africa (Somali and Galla). The most widespread animals are cattle, varieties of crossbreedings between Mediterranean and Indian stocks. Other livestock include goats and sheep, both widely distributed; camels and horses in the arid areas; donkeys; pigs in a few areas; dogs and fowl in virtually every part of the continent.

The traditional economy of 90 percent of the population is agriculture, usually in association with animal husbandry and hunting. Crops comprise both indigenous and imported trees and plants. In the savanna zones grains provide the staples, whereas in the forest areas the staples are mainly root crops. There are several important indigenous plants: the chief food crops are millets and sorghums, yams, plantains, and many lesser vegetables; the chief oils are sesame, castor, and the oil palm; the chief spices and condiments are kola, coffee, and kat; and the chief plant textiles are cotton and bark cloth. There are also many indigenous trees used for fruit, building, and fuel. The more important imported crops (most of considerable antiquity) include barley, wheat, rice, peas, many beans, some varieties of yams, taro, the date palm, banana, sugarcane, and coconut palm, all of which come from Asia; and also maize (corn), manioc (cassava),

sweet potatoes, groundnuts (peanuts), and tobacco, all coming from the Americas.

Methods of agriculture vary widely. Throughout the savanna areas forms of shifting agriculture and crop rotation are practiced, fields being burned and cleared, used for a few years until the natural fertility is lost, then left as fallow. Consequently, small local groups need a larger area of land than they use at any one time, and settlements are rarely permanent or even long lasting. In the more fertile areas more fixed methods of farming are usual, and settlements tend to be permanent and larger. There are many exceptions to this generalization, and the spread of cash crops and trees has markedly altered the traditional pattern.

Demographic patterns

The demographic situation in Africa is one of marked inequality of population distribution. Censuses in most of Africa have been unreliable and uneven as far as information for the constituent ethnic groups of modern nation-states is concerned. The total population of Africa in the early 1970s was estimated to be about 335,000,000, distributed approximately among the major regions as follows:

northern Africa	65,500,000
western Africa	107,000,000
eastern Africa	72,500,000
west central Africa	35,800,000
southern and central Africa	47,700,000
Madagascar	6,500,000
	335,000,000

Certain areas have very dense populations, while others are almost empty. Among the most heavily populated are the Mediterranean littoral, much of the West African coastal forest region, the areas of the Great Lakes and the highlands of eastern Africa; whereas the Sahara, the Kalahari, the Horn of Africa, and the dense Congo forests are all lightly inhabited with many areas devoid of any fixed population.

This article is divided into the following sections:

Cultures and peoples
  Cultural areas
  Races and languages
Origins and early history of African peoples
  The known beginnings
  Medieval Africa
  The colonial era
The nature of African societies and their cultures
  Economies
  Patterns of settlement: village and town
  Domestic groupings
  Political patterns
  New associations and social stratification
  Religions

CULTURES AND PEOPLES

Cultural areas. The basic units of African society have been—and in much of modern Africa still are—the ethnic groups: tribes, or peoples. They number almost 3,000. Most are not politically or economically independent social units but rather groups that have a common sense of culture and identity, especially in terms of a distinct language and religion. Boundaries between them are usually not very clearly defined, and it is rather frequently difficult to know whether or not a congeries of very similar communities should be regarded as one tribal group or as many tribal groups. It is convenient to classify these peoples by languages and by the main outlines of their cultures.

Classifications of culture areas

There have been many attempts at cultural classifications based on the concept of a culture area consisting of a number of societies with certain common features. Although to some extent useful, the generally accepted culture areas for Africa are too large to be very meaningful because each includes widely differing forms of social and political organization. The best known of many older classifications is that by the American anthropologist Melville J. Herskovits (from 1924), which is still referred to despite its inadequacies. A later classification (1959), also by an American, G.P. Murdock, takes as units relatively small clusters of peoples, who are grouped largely on linguistic and ethno-historical grounds. He divides the continent (including northern Africa) into 55 clusters of tribes, forming 11 main groupings, some of which are defined by livelihood (such as hunting or pastoralism) rather than by ethnic or linguistic affiliation. If one chooses to ignore these rather numerous divisions, then the principal regions (excluding northern Africa) are western, west central, eastern, and central and southern Africa.

Western Africa. Western Africa comprises two principal zones, the savanna along the southern Saharan borderland and the rain forest along the Atlantic (western) coastline.

Principal crops of the savanna zone include millets, sorghums, maize (corn), sesame, and many legumes; in the western part, from The Gambia to Liberia, wet and dry rice are important. In the savanna, cotton is grown, and the loom is traditional. Indigenous crafts include casting in brass and bronze, silver and gold filigree work, leatherworking, and building in mud brick with vaults and arches. Cattle and horses are important, the latter being used in trade and formerly in warfare, communications being relatively easy east and west along the open plains.

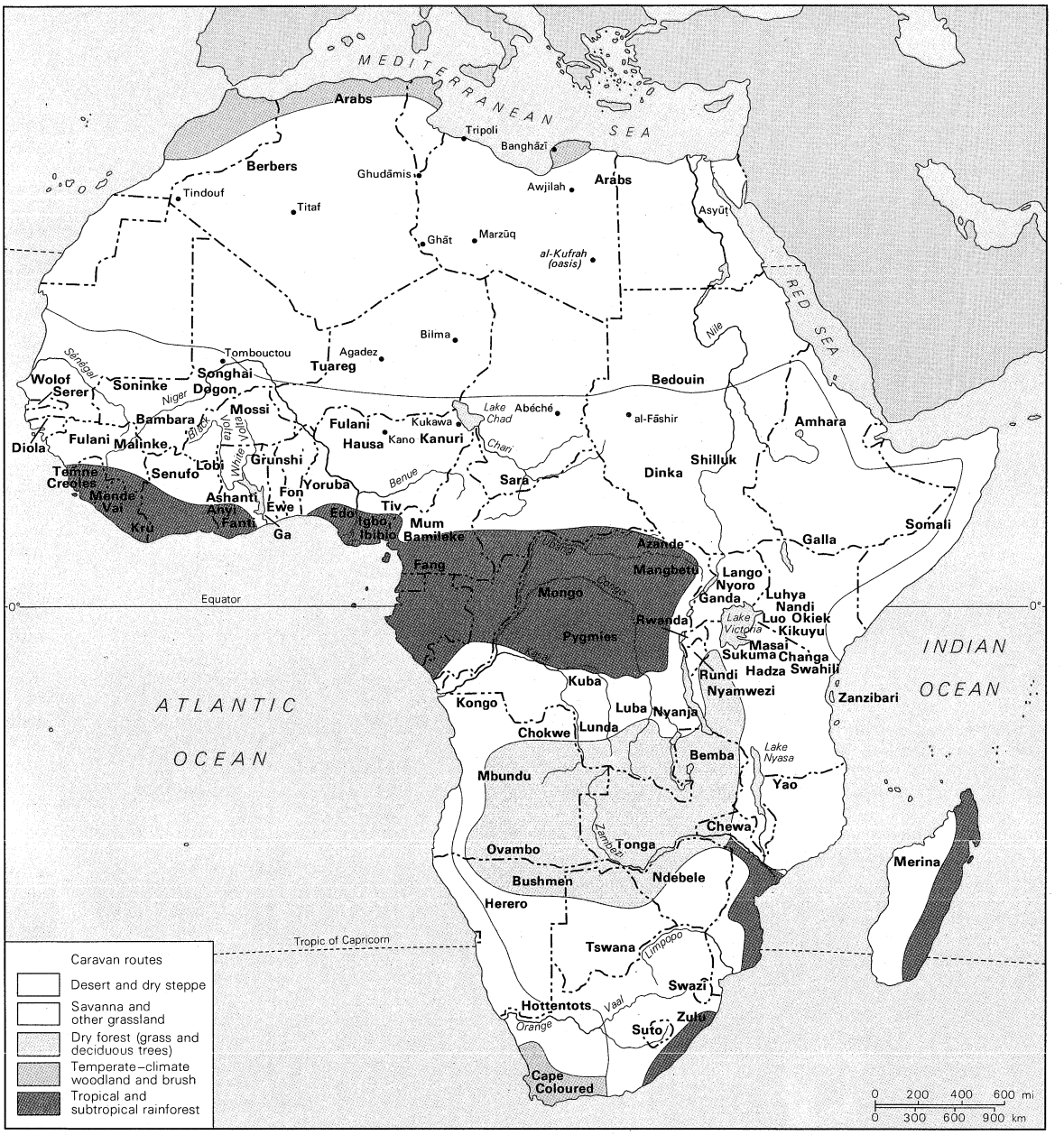
In the forest zone root crops are basic (yams, taro), and oil and raffia palm, fruits, cocoa, pepper, and kola are valuable for both subsistence and external trade. Livestock comprise goats, pigs, fowl, and in a few areas dwarf cattle. The forest and coastal areas contain important minerals (bauxite, diamonds, and gold). There is much wood carving, pottery making, metal casting, and iron forging.

Kingdoms of West Africa

Most of the larger and more powerful of the traditional societies are kingdoms, each surrounded by lesser and politically weaker communities. Among the more important of the savanna peoples are the three main clusters known as Mande (in Senegal and Mali, and including the Bambara, Malinke, Soninke, and many more); the Voltaic group in the savanna zone to the east (including the Senufo, Lobi, Grunshi, Dogon, and Mossi); and in northern Nigeria, Niger, and Cameroon the many small mainly non-Muslim tribes of the plateau and highland areas. Throughout this region live the many groups of the Fulani (known also as Fula and Peul), at least 6,000,000 strong, a cattle-keeping Muslim people who have either conquered indigenous peoples or live in a symbiotic relationship with agricultural peoples. In the Sahara fringe are the many Berber-speaking groups collectively known as Tuareg, the Kanuri of Lake Chad, and many Semitic-speaking Bedouin Arab peoples. Many of the kingdoms are successor states to those of the Middle Ages (see below).

The larger societies in the coastal zone are also mostly kingdoms. In Nigeria are the Igbo and Ibibio, organized into many autonomous polities; the Tiv; the Edo of Benin; and the several powerful kingdoms of the Yoruba. Westward are the Fon of Dahomey, the various peoples of the Akan confederacy, the largest being the Ashanti; the Ewe, Ga, Fanti, and Anyi of the coast; the Mende and Temne of Sierra Leone; the Kru of Liberia; the Wolof, Serer, Dyola, and others of Senegal; and the Creoles of Sierra Leone and Liberia, descendants of freed slaves from the New World.

West central Africa. Geographically, west central Africa may be considered as an easterly extension of western Africa: in the north are the savannas of Chad, Ubangi, and the Sudan, stretching to the Nile, and in the south is the largely forested area of the Congo. It is ethnically very mixed, with Arabs in the north, Pygmies in the Congo and Gabon, and Sudanic- and Bantu-speaking peoples in the more southern areas. Agriculture is the basic economy of virtually the entire region. The northern savanna includes many peoples of the Cameroon area, divided into many small kingdoms, of which the Bamileke tribes are the most numerous. Between Cameroon and the Nile are many large Sudanic-speaking peoples, such as the Sara, Mangbetu, and Azande. Southward are Bantu peoples, of which the more important in-



Major peoples and vegetational zones of Africa.

clude the Fang of Gabon and the Kongo, Mongo, Kuba, Luba, Lunda, and Chokwe of the Congo and Angola.

Eastern Africa. Eastern Africa contains several ecological and cultural areas. In the north and east are the arid Sudan and Somalia separated by the Ethiopian highlands; in the centre are the fertile areas of the Great Lakes (Victoria, Albert, Tanganyika, Nyasa) and the highlands around Mounts Kenya and Kilimanjaro. The remainder consists of savanna with the depression of the Great Rift Valley running from north to south. The areas of densest population are the fertile highlands, which produce the main cash crops today, mainly cotton and coffee.

Throughout the area the traditional staples are grains (millets, maize [corn]), except for the Lake Victoria region, where it is the plantain. Root crops are not as important as in West Africa. Livestock (cattle, goats, sheep, fowl, donkeys, dogs) are found throughout the region; pastoralist societies are found in The Sudan, Somalia, and the central plains of Kenya and Tanzania.

This ethnically complex region includes the Para-Nilotic-speaking pastoralists of the Nile Valley (Shilluk, Dinka, Luo, Lango, and others, formerly called Nilotes), those of the central plains (Masai, Nandi, and others, formerly known as Nilo-Hamites), and the Cushitic-speaking Somali and Galla of the Horn of Africa. In Ethiopia also are the Semitic-speaking Amhara and others. Most of the remaining peoples of the region are Bantu speakers varying widely in political and other cultural features but all being peasant farmers. In the Great Lakes region are the several formerly powerful Bantu kingdoms (Ganda, Nyoro, Rwanda, Rundi, and others). In the Kenya highlands are the Kikuyu, Luhya, and others. On the coast are the various Swahili-speaking tribes; in Tanzania are the Chaga, Nyamwezi, Sukuma, and many more. There are also small remnants of other ethnic groups: the hunting Okiek or Dorobo, Hadza, and some Pygmies. And on the coast are the remnants of the once politically powerful Arabs, formerly based on Zanzibar.

Ethnic complexity of East Africa

Central and southern Africa. Central and southern Africa may be considered as forming a single large culture area. Most of it consists of open and dry savanna grasslands: to the northwest are the edges of the Congo forests; the southwest is very arid; and the coastline of South Africa and Mozambique is more fertile, much of it with Mediterranean climate.

The region was once populated by Bushmen and Hot-

tentots, both Khoisan-speaking peoples, the former being hunters and the latter pastoralists. The Bushmen are today restricted to the arid areas of southwest Africa and Botswana, and most of the Hottentots are merged into the racially mixed Cape Coloured people of the Cape Province of the Republic of South Africa with European and Asian admixtures. The other indigenous groups are all Bantu-speaking Negroid peoples, originally from the area of Cameroon, who dispersed across this region some 2,000 years ago. The vanguard, known linguistically as the Southern Bantu, drove the Khoisan before them and adopted some of their click gutturals into their own languages. After contact with Europeans in the 18th and 19th centuries, the Bantu turned upon one another, establishing the conquest states of the Zulu, Swazi, Tswana, Ndebele, Sotho, and others. To the north are the Central Bantu from the Congo to Malawi, as well as farmers with many kinds of political systems and differing markedly from the Southern Bantu in their kinship systems; they include the Bemba, Thonga, Chewa, Nyanja, Yao, and many others. In southwest Africa are the Southwestern Bantu, including the Ovambo and Herero.

Madagascar. Finally, the island of Madagascar forms a distinct culture area. The various Malagasy tribes, of which the politically most important is the Merina, are of Indonesian origin, having migrated across the Indian Ocean probably during the 5th and 6th centuries AD.

Population types in Africa

Races and languages. The racial and ethnic compositions of Africa are complex. The usually accepted classification of modern populations is based on phenotypic characteristics such as stature, skin colour, and hair type rather than on genetic and blood-type features. The fivefold classification of those groups in existence at the present time is as follows: (1) The Bushmanoid peoples, who, though numerically few today, were once widespread over eastern and southern Africa, are short and have yellow-brown skins and hair of a distinct "peppercorn" type; often there is steatopygia (excessive development of fat on the buttocks) in the women. Today they are represented only by the Bushmen and Hottentot remnants in southwestern Africa and by a few probably Bushmanoid types in remote areas of Tanzania. (2) The Pygmoid peoples are also short and brown skinned, though today they are much affected by Negroid admixture. (3) The Negroid peoples are the most widespread on the continent and have many diverse physical features: the skin colour varies from almost black in much of western Africa to light brown in eastern and southern Africa; in height, the range is from short people in much of western Africa to the very tall and slender Nilotes in the southern Sudan. (4) The Caucasoids of northern and northeastern Africa are related to the non-African Mediterranean Caucasoids; their skin colour varies between very light brown in northern Africa and almost black in Somalia; the hair type is usually wavy. (5) The Mongoloid peoples are found in Madagascar as a relatively recent immigrant group to the continent.

Linguistic classifications

Many attempts to classify the languages of Africa have been inadequate because of a confusion between language, race, and economy; for example, there was once a spurious view of pastoralism as related to cultures whose members spoke Hamitic languages and were descendants of ancient Egyptians. Today the knowledge of most of the individual languages of the continent is still very inadequate, but there are known to be at least 1,000 distinct languages (see also AFRICAN LANGUAGES; HAMITO-SEMITIC LANGUAGES).

One of the most recent and accurate attempts to classify African languages, prepared by the American linguist Joseph Greenberg, is based on the principles of linguistic analysis used for Indo-European languages rather than on geographical, racial, or other nonlinguistic criteria. The main language families of the continent according to this classification (not accepted in totality by all specialists) are the following: the Niger-Kordofanian is the most widespread family and comprises two subfamilies, Niger-Congo and Kordofanian. These languages are found from Senegal to the Congo, with a geographically widespread extension due to recent migrations over eastern and southern Africa. Niger-Congo includes the West Atlantic, Mande, Voltaic, Kwa, Benue-Congo, and Adamawa-Eastern subgroups; Kordofanian includes 15 subgroups all spoken within a very small area of the Republic of the Sudan. The most original point in this new classification is the group called Benue-Congo, which linguistically subsumes all the Bantu languages found dispersed over most of eastern, central, and southern Africa. This dispersal is due to the rapid expansion of people from the area of the Bight of Benin from the beginning of this millennium onward: the vanguard, the Southern Bantu, had not reached the Cape of Good Hope when the Dutch arrived there in the 17th century. The close linguistic similarity between all the Bantu languages points to the speed of this vast migration.

The Nilo-Saharan family comprises languages spoken along the savanna zone south of the Sahara from the middle Niger to the Nile, with outlying groups among the Para-Nilotic pastoralists of eastern Africa. Its subgroups are: Songhai, Saharan, Maban, Fur, Chari-Nile (including the large clusters of Eastern and Central Sudanic), and Koman.

The Hamito-Semitic (Afro-Asiatic) family includes Semitic, Ancient Egyptian, Berber, Cushitic, and Chad.

The Khoisan, or Click, family comprises the languages of the Bushmen and Hottentots, today limited to southwestern Africa and the outlying Hadza and Sandawe of northern Tanzania.

The Austronesian (Malayo-Polynesian) family is represented by the various dialects of Malagasy in Madagascar.

Lingua francas of Africa

There are also many widespread trade languages and lingua francas. Some were imported and used by administrators, missionaries, and traders during the colonial period. They include English, French, Arabic, and other languages of the former colonial powers; some of them are today the national languages of newly independent nation-states. Swahili, grammatically Bantu but with much Arabic in its vocabulary, is widely used in eastern Africa; as the language of the people of Zanzibar and the east coast, it was spread by 19th-century Arab slavers in the hinterland as far as the Congo. In western Africa forms of Creole (Krio) and Pidgin are widespread in the coast towns of very heterogeneous ethnic composition, and Hausa and Fulani are widespread in the savanna zone. In southern Africa, Fanagalo, a mixture of English and local Bantu tongues, is spoken in mining areas.

The great majority of African languages have no indigenous forms of writing, although today governments and missions have been reducing them to writing. Many African languages (such as Swahili) have for centuries been written in Arabic script. There are exceptions, however. The most famous are those of the Vai of Sierra Leone, the Mum of Cameroon, and the Tuareg and other Berber groups of the southern Sahara, all of whom invented their own scripts.

ORIGINS AND EARLY HISTORY OF AFRICAN PEOPLES

The known beginnings. In much of Africa about 15,000,000 to 25,000,000 years ago there were various forms of the dryopithecine apes whom some authorities claim to be common ancestors of man and the recent apes; and about 14,000,000 years ago a hominid form, *Ramapithecus*, developed in East Africa. From about 5,000,000 years ago forms of *Australopithecus* were also found, particularly in the Transvaal and at the Olduvai Gorge in Tanzania. These last were tool-making hominids comprising two different species, the slender *Australopithecus africanus* and the larger *Australopithecus robustus*, or *Paranthropus* (including a form found at Olduvai that has been called *Zinjanthropus*). These would seem to have evolved into the hominids known as *Homo habilis* and *Homo erectus*, found in East, South, and Northwest Africa from about 1,000,000 years ago, gradually evolving into a Neanderthal form by about 60,000 years ago. Neanderthals (found also in Europe, China, and Indonesia) were hunters who used stone tools and were without significant regional specialization.

There then occurred a lowering of much of the conti-

nent's climatic temperatures and the appearance of *Homo sapiens* of modern type, who were hunters living in both forest and more arid regions. There is evidence for continuous occupation of sites from many areas, and by the Middle Stone Age, about 35,000 years ago, man was spread throughout Africa, with marked regional specialization of tools consistent with environmental variations. The earliest *Homo sapiens* fossils probably represent the basic stock from which have developed the modern Bushmanoid and Negroid races.

Appearance of indigenous African races

By the end of the Pleistocene Epoch (about 10,000 years ago), the emergence of the distinct indigenous African races may be discerned. Bushmanoid sites have been found from the Sahara and The Sudan to South Africa. It seems likely that the homeland of the Negroids was in western Equatorial Africa. A burial site at Iwo Eleru, near Benin in Nigeria, shows a Negroid physical type with stone-bladed tools that were typical of the period and can be dated to about 11,000 years ago. Neolithic people—that is, people with crops and domesticated livestock—both Bushmanoid and Negroid, occupied the Sahara, eastern and southern Africa, and elsewhere from about 5500 to 2500 BC, with the latter part of the period marked by extensive rock paintings such as at the Tassili site in the Sahara. In the West African forests Neolithic peoples can be dated from the 3rd millennium BC, and by the 1st millennium BC cultivators were clearing the forests of the Great Lakes region. In West Africa are found relics of the Nok Culture of northern Nigeria, dating from about 900 BC to AD 200, with terra-cotta figures and traces of ironworking; it was transitional between a Stone Age hunting culture and a food-cultivating one. Saharan-type people occupied the East African highlands from about 2500 BC, to be replaced by Iron Age Negroid peoples about the beginning of the present millennium. By this time the knowledge of ironworking was spreading throughout Africa, perhaps by borrowing directly from Meroe (see below), perhaps in some areas by independent invention. This period also marked the dispersal of the Bantu-speaking peoples, first from a centre in the Cameroon–Chad region to a centre in southeastern Congo and thence throughout eastern, central, and southern Africa.

Ancient Egypt and Nubia

The first culture of which a great deal is known is that of Egypt. By 4000 BC the Egyptians were cultivating crops and were soon to establish the kingdoms of Lower and Upper Egypt, to be united by 3000 BC. By 2000 BC they were sending expeditions to the far south (Nubia and Ethiopia) and to the land of Punt by the Indian Ocean. Details of Egypt need not be given here, but by 1000 BC its power was in decline, and the centre of royal authority moved southward into Nubia (see EGYPT, HISTORY OF; NILOTIC SUDAN, HISTORY OF THE).

The main successor state was that of Kush (Cush) in Nubia. Its first centre was the city of Napata on the Nile, which controlled the sources of the Nubian gold exported to Egypt. In the 8th century BC the Kushites invaded Egypt and established a dynasty in Upper Egypt for about a century. At this time the importance of iron was recognized as superior to bronze, and the centre of Kush was moved southward to Meroe, about 100 miles north of the present-day city of Khartoum. Meroe, the centre of Kush for some 600 years, was a trading centre for the region between the Sahara and the Red Sea. It was a source of iron and ironworking; it had its own alphabetic script and may have been the earliest site to cast metal by the lost-wax process.

The empire of Aksum to the east of Kush defeated its rival Meroe about the 4th century AD. The Aksumites, who spoke a Semitic language from southern Arabia, became Christian and were later absorbed into the kingdom of Ethiopia. The Nubian Christian kingdoms of the Sudan maintained their independence another ten centuries until they were overthrown by Muslim Arabs from the north.

Medieval Africa. The African Middle Ages was a period of large trading empires or kingdoms intermediate between the beginnings of the Iron Age and the advent of colonial rule by European and other powers.

In western Africa there was a succession of powerful trading empires that controlled the trans-Sahara trade in gold, kola, and slaves from the south and in salt and metals from the north; they were almost all Muslim. All had similar histories of development and decline: from towns controlling trade routes they grew into large, wealthy, and sophisticated empires; they then were conquered by rival groups and underwent a final decline and eclipse by the forest states of the Atlantic coast at the beginning of the colonial era.

Empire of Ghana

The first empire of which there is extant knowledge was Ghana. Its capital at Kumbi (some 200 miles north of present-day Bamako, capital of Mali) controlled trade between Morocco and Mauretania to the north and Senegal and the coastal forest to the south. Arising in the 5th and 6th centuries AD, it was by 1000 a powerful empire with an army using iron weapons. Kumbi is estimated to have had a population of 30,000, with an army throughout the empire of 20,000 (larger than European armies of the same period). About 1000 the Berbers of the northwest Sahara expanded into Spain to the north and into Ghana to the south, conquering Kumbi in AD 1076. Although they withdrew, they weakened the state, and it was finally destroyed by Fulani attacks in 1230.

The same period saw the rise of the empire of Kanem-Bornu around Lake Chad, the ruling dynasty of which lasted from the 8th century until 1846. It controlled the eastern Saharan routes to Sudan and Egypt, dealing in cotton, kola, and spices from the south for salt and copper from the north. The 17th century marked the height of its power, with the state maintaining diplomatic missions in Cairo and Tripoli (Tarābulus); after that there was a slow decline.

From the 11th century onward the Hausa states to the west of Kanem-Bornu acquired international power, first centred at the city of Katsina and later at Kano (both in present-day Nigeria). Always Muslim, they were captured by the Muslim Fulani in the early 19th century and are still in existence as Fulani emirates, with that of Sokoto exercising formal primacy.

Empire of Mali

The greatest of the West African medieval empires was the Muslim state of Mali, which about 1250 filled the vacuum left by Ghana. The height of its power was under Emperor Mansa Mūsā, who reigned from 1307 to 1332 and controlled the area from the Atlantic almost to Lake Chad. Mansa Mūsā passed through Cairo on his way to Mecca in 1324, astonishing the Muslim world by his wealth. The capital was Niani (some 150 miles southwest of the modern Bamako); Timbuktu was its greatest city, renowned for its university, its book trade, and its wealth. The empire declined after Mūsā's death (though the Malinke people still occupy the modern state of Mali), giving way in about 1400 to the empire of Songhai (based on the town of Gao, on the Niger some 200 miles east of Timbuktu). The height of Songhai power was under the emperor Mohammed I Askia of Songhai (Mohammed Towri), who visited Mecca in 1496 and was deposed in 1528. In 1591 Songhai was invaded by Moroccan Berbers. The Songhai still number about 1,000,000 people in the bend of the Niger River.

The prehistory of the same period in eastern and central Africa has a greater emphasis on trade with countries beyond Africa, including Arabia, Persia, India, and even China. It is known that about AD 100 an Arab-controlled trade was established between these countries and the East African coast, based on the export of ivory, gold, and slaves. By the 13th century the Arab–Swahili settlements had developed into powerful city-states: Mogadishu, Mombasa, Lamu, Kilwa, and others. In the 16th century they were destroyed by the Portuguese, who were in turn ousted by the Omani Arabs who established the slave-trading state of Zanzibar.

To the south was the empire of Monomatapa, with the citadel of Zimbabwe (near the present Fort Victoria in Rhodesia) as its greatest monument. Unlike the semi-Arab states of east Africa, this was a Bantu kingdom. Its height was in the 15th century, but in the 17th century it too was weakened by the Portuguese and finally destroyed by other Bantu groups in the 19th century.

Other large protohistoric and medieval societies in Africa are less well known. Inland from the East African coast the states of the so-called Interlacustrine Bantu (Bantu living between the Great Lakes) were established about AD 1300 by the invasion of light-skinned peoples from the north, followed by later Negroid invasions about 1500. The dynasties of Nyoro, Ganda, Rwanda, and others lasted until the 1960s, when the rulers were deposed after the withdrawal of colonial rule.

The kingdoms of the Congo region were important in establishing trade routes across tropical Africa. The kingdom of Kongo near the mouth of the Congo (Zaire) River was a trade state (or congeries of small states) that was converted to Christianity by the Portuguese about 1500; it was later weakened and destroyed by Portuguese demands for slaves. Further south were the savanna kingdoms of Mbundu, Luba, Lunda, and others, all likewise weakened by the Portuguese but lasting in one form or another until the present day.

The colonial era. All the empires noted above were based on long-distance trade that involved the development of specialized economic institutions (markets, forms of coinage, tribute, and taxation) and forms of government (military protection and administrative bureaucracies). They were weakened and in many cases destroyed by the early colonial powers from Europe and Arabia, who coveted the trade and supplanted them, turning much of the trade to external rather than internal markets and creating new trade goods. The most important of these trades was in slaves, which increased in importance until the 19th century, when it was abolished and replaced by trade in such export crops as palm oil. Estimates of the slaves shipped across the Atlantic run from 30,000,000 to 100,000,000; figures for the Arab trade in eastern Africa are not known but certainly run into many millions.

The slave trade

The West African trade involved the capturing of slaves by African rulers who traded them for guns and other goods supplied by the European shippers; the East African trade was carried on more directly by the Arab slavers themselves. There were several main consequences of this trade in the historical development of Africa: great depopulation, especially in Angola and in East Africa; the import, especially into West Africa, of such European trade goods as cloth, guns, and metalware, which inhibited indigenous African production; the bringing of Africa into closer contact with the outside colonial world; and a change in the nature and distribution of African trading states. In West Africa, particularly, the Saharan states declined in favour of the growth of the southern forest states that dealt in gold and slaves, and in East Africa Zanzibar grew while the hinterland kingdoms declined.

In West Africa the early colonial period saw the growth of the forest kingdoms at the expense of those of the savanna. These forest states had existed during the Middle Ages, but, though they left impressive works of art, they had been politically unimportant. During the 17th and 18th centuries, they flourished, competing among themselves for the lucrative Atlantic trade with the European traders whom they wished to keep out and so preserve their own monopolies. The forest states have been mentioned above: the kingdoms of Yoruba, Benin, Dahomey, Ashanti, and many others.

Social results of European colonialism

This article is not the place for an extended account of the European colonial period in Africa, but it is useful to mention the principal social consequences. They have included the introduction of a cash economy and the bringing of African farmers into a worldwide capitalistic system; part of this process has been the building of widespread and efficient communications systems. The power of indigenous kings and chiefs has been weakened and in places destroyed as the colonial powers appropriated ultimate political authority for themselves. By establishing inter-territorial and other boundaries, the colonial powers set up arbitrary barriers between groups that had traditionally been in a far more fluid situation. This era witnessed the weakening of the tribe and the establishment of modern nation-states with the widening of the social scale and resultant changes in personal and group affiliations and loyalties. In the religious sphere, the introduction of Christianity and Islām weakened the strength of small-scale tribal religions: again, Africa has been brought into the orbit of worldwide religious systems. The introduction of medical services has altered much of the traditional demographic balance. (For more detailed histories, see WEST AFRICA, HISTORY OF; EAST AFRICA, HISTORY OF; ETHIOPIA, HISTORY OF; CENTRAL AFRICA, HISTORY OF; and SOUTHERN AFRICA, HISTORY OF.)

THE NATURE OF AFRICAN SOCIETIES AND THEIR CULTURES

It is useful here to present the main aspects of African social organizations and cultures as they exist at the present time. All societies in Africa have always been in a state of slow change, but in the second half of the 20th century the rate of development has increased immensely. Most societies are in transition from remembered traditional, small-scale, narrow-range forms to wide-scale ones that are part of worldwide economic and political systems; this change is often referred to as "breakdown" but is also part of the process of modernization. Perhaps nowhere in the world is this process so rapid and so all pervasive as in Africa.

Economies. Almost all Africans are small-scale farmers, who have traditionally relied on simple resources, tools, and techniques, have been hampered by poor communications, and have not engaged in much widespread trade. Traditionally most production has been for subsistence only. The unit of production and distribution has been typically the community or family—nonspecialized groupings that are not primarily economic but equally domestic, political, and religious units. Traditionally, too, standards of living have been more or less uniform within a given region; in the past there were few markets and rarely any forms of money, exchange being accomplished between kin and neighbours by barter or gift. Wage labour was unknown, though informal cooperation in field work and other tasks was almost universal. The principal exceptions were those few societies that had engaged for many centuries in long-distance trade and had elaborate exchange and craft facilities, communications, and the political superstructure to maintain their trade routes.

Traditional farming economies

During the past century this traditional picture has changed with the establishment of colonial rule and later political independence over most of the continent. Africa has become part of the total worldwide system of economic production and exchange. The former empires early lost their importance in controlling the main external markets as other markets were developed in Europe, the New World, and across the Indian Ocean. Among the more important developments were the slave trades to the Americas and the East; the discovery and exploitation of minerals (such as gold, diamonds, tin, and bauxite) and of profitable export crops (such as cocoa, coffee, palm oil, cotton, and tea); the establishment of taxes in most colonial countries, which were payable only in cash and so forced people to grow cash crops or sell their labour; and the setting up of industrial and administrative cities and ports. The consequences include the growth of peasant production and of the import of consumer goods; a change from a traditional equality in living standards to a marked distinction between rich and poor, especially in the cities; the growth of elites and class differentiation in both political and economic contexts; the growth of wide-scale exchange systems based on money and markets; the establishment of road and rail transportation for movement of peoples, export crops, and consumer goods; the migrations of wage labourers from the more distant and impoverished rural areas to the cities and the regions growing export crops (as in the migrations to the cocoa areas of Ghana and Nigeria, the cotton areas of Uganda, the industrial areas of Zambia and South Africa); the development of large-scale economic and political regions controlled largely from the cities, with the loss of autonomy for the traditionally smaller tribal units; the growth of a poor and typically exploited urban proletariat, with low skills and

Contemporary economic systems

serious underemployment (especially of younger people with some education for whom there is little but unskilled work available); and the decline in the number of able-bodied younger men in the rural areas, creating a vicious cycle of impoverishment in both city and countryside.

Patterns of settlement: village and town. Traditional African patterns of settlement vary with differences in landscape and ecology, communications, and warfare. The most widespread pattern has been that of scattered villages and hamlets, the homesteads of joint and extended families (see below), large enough for defense and domestic cooperation but rarely permanent because of the requirements of shifting cultivation and the use of short-lived building materials. In much of the West African savanna, large mud-adobe villages are traditional, but over most of Africa housing consists of mud and wattle with roofs of thatch or palm leaves.

Characteristics of old and new towns

Large and permanent towns have always been found, however, in many parts of the continent. Towns dating from precolonial times are found mainly in West Africa, in both forest and savanna zones, where they were the seats of governments of kingdoms. Timbuktu, Jenne, Kano, and others date from the Middle Ages; others such as Benin and the Yoruba cities of Ibadan and Oyo are more recent but certainly many centuries old. In East Africa there are such medieval Arab coastal towns as Mombasa. Almost all these towns still have an indigenous and fairly homogeneous population; there are many occupations, including farming and crafts; there is little or no industry and a wide range of wealth and power culminating in the ruler, whose authority extends over the neighbouring countryside, so that the town is really the nucleus of a wide area rather than an isolated unit. On the other hand, the towns and cities established under colonial rule as administrative, trading, or industrial centres and ports are very different. They are found throughout Africa and include Johannesburg, Lusaka, Salisbury, Kinshasa, Lubumbashi, Nairobi, Dakar, Freetown, Abidjan, and many others; often, as in the case of Lagos or Accra, they are built onto traditional towns. Typically the focus of immigration from an impoverished hinterland, they are ethnically heterogeneous, and there is a sharp break between the city and the surrounding countryside. There is usually a marked differentiation between skilled and unskilled labour, the former (together with members of the urban administration) often being alien and, in colonial days, racially distinct from the mass of the population.

Domestic groupings. The forms of the family found in Africa are consistent with the forms of economic production. Throughout most of the rural areas the typical domestic group is the joint or extended family consisting of several generations of kin and their spouses, the whole being under the authority of the senior male. The size of the group varies, but it typically consists of three to five generations of kin. It provides a stable and long-lasting domestic unit able to work as a single cooperative group, to defend itself against others, and to care for all its members throughout their lifetimes. Polygyny is traditionally widespread as an ideal, its extent depending on the status and wealth of the husband: chiefs and rulers need many wives to give them a mark of high position and to enable them to offer hospitality to their many subjects.

Clans and lineages

In most of Africa these residential groups are based upon descent groups known as clans and lineages, the latter being segments of the former. The significance given to descent groups varies, but they are widely important in providing arrangements for heirs, successors, and marital partners.

In the second half of the 20th century this pattern has been changing, rapidly in the urban and poverty-stricken areas, more slowly in those areas less affected by economic and political development. In cities and in major labour-supplying areas such as most of southern Africa, the joint or extended family is giving way to the independent elementary family of man, wife, and children. There is also a tendency toward the breakdown of family structure because of labour migration—the younger men moving to the cities, leaving women, older men, and children in the impoverished homelands.

Political patterns. There is wide variation in both traditional and present-day systems of government. There are three main types of traditional government. One is the simple organization of the hunting band, as among Bushmen and Pygmies, in which the political group, the band itself, is essentially a single kinship unit linked with neighbouring bands by ties of intermarriage and consciousness of common cultural identity. What authority exists is held by the oldest men. The second type is the most widespread and comprises those societies that lack centralized political government and maintain political order and identity by other means. Political authority in this case is typically not specialized and is not clearly distinguished from kinship and religious authority. Many such societies are based on clans and lineages, with most political authority being held by genealogically senior men in the clans and lineages; disputes, though, were traditionally settled by organized force (feud and warfare). In other kinds of politically uncentralized societies, authority is held by the senior men by age, the whole society being divided into formally recognized age generations. Here councils of elders representing local groups over a wide area hold governmental authority, substituting discussion for feud. They are found especially among such societies traditionally based on cattle raiding and warfare as the Masai and some other tribes of eastern Africa and the Zulu and other south African peoples. In yet other societies, village associations (as among the Igbo of Nigeria) or "secret societies" (as among the Mende of Sierra Leone) hold governmental authority over local communities, the associations or societies being composed of representatives of local families and villages.

Traditional African states

The third main type of political system is that of the traditional African states. These states vary in many ways: in their size (from a few thousand to the many million of the Ashanti or Yoruba), in the degree of political power held by the king (from that of the ruler of the Shilluk of the Sudan, who possesses the mystical power of the tribal culture hero yet lacks any real authority himself, to the traditionally despotic kings of Rwanda, Ganda, and the other Interlacustrine Bantu kingdoms); and in the ways in which the king is related to his subjects. In this last respect, he may be regarded as simply a senior kinsman, as among the Zulu or Swazi; or he may be *primus inter pares* ("first among equals") in a large federation of many rulers, as among the Ashanti of Ghana; or he may be a member of a conquering autocracy that regards itself as racially distinct from the majority of its subject people, as in Rwanda, Rundi, or the former Arab sultanate of Zanzibar.

Many of these polities had similarities to medieval European feudal states. They were highly organized, powerful kingdoms with complex institutions to define and control the power of the ruler, who had both political and religious authority. Many were controlled by dynasties of many centuries with aristocratic courts and administrative bureaucracies that made them comparable to states in other parts of the world.

Conditions in the new nation-states

Today most of these polities still exist but as dependent parts of new nation-states and with the former authority of their chiefs and rulers greatly curtailed or even abolished. All of them were weakened during the colonial era because the colonial governments monopolized political power, and others have been destroyed by the postcolonial independent national governments, which have often considered them reactionary and anti-nationalistic. Traditional small-scale tribal polities have been able to command the loyalty of their members—a phenomenon generally known as tribalism—but are politically and economically unviable in the modern world. The new nation-states are almost all those formerly created and given ethnically artificial boundaries under colonial rule; they are mostly multitribal, with varying forms of government based on democracy, "African Socialism," military rule, and so on. The contradiction between loyalties based on tribal and language group

on the one hand and the new nationalism on the other is of considerable importance in many parts of the continent—as, for example, in Nigeria, Zaire, and Kenya.

New associations and social stratification. African societies seem traditionally to have been relatively homogeneous in both ethnic and class terms: they were stable as far as the ranges of social scale were concerned; they permitted little social mobility; and they admitted relatively little radical or rapid social change and adaptation.

In recent years, however, there has been marked social mobility both vertically and horizontally, and there have appeared new forms of social organization. New voluntary associations of many kinds have arisen to help urban immigrants cope with city life and still retain some measure of traditional loyalty to their rural homelands. The first associations to appear are various kinds having primarily recreational, religious, or educational aims; at a later stage there arise trade unions and political parties. At first these associations are usually tribally oriented, linking fellow tribesmen together in areas to which they migrate; later, common occupational and political interests assume greater importance.

The new social classes

Most traditional African societies had only a modicum of social stratification measured on the basis of wealth and power. But today economic and political differences are growing increasingly more significant with the increase in African entrepreneurs and traders, new politicians and religious leaders, and with the growth of inequalities in education. Economic and political elites have appeared in all areas, and they have tended to drop traditional tribal and familial affiliations and become members of new upper and middle classes.

Religions. The indigenous religions have certain common features. All known include the notion of a High or Creator God, remote from men and beyond their comprehension or control. This God is typically not attributed a sex but in some cases is male or female; often God is given an immanent and visible aspect as well. The most important "spiritual" powers are usually associated with things or beings with which people have day-to-day contact or which they know from the past. Thus there may be many kinds and levels of spirits of the air, of the earth, of rivers, and so on. There may be ancestors and ghosts of the dead who have achieved a partial divinity. There may be mythical heroes who led the people to their present land and founded their society as it is known today.

Varieties of religious leaders and functionaries

The ritual functionaries found in most African societies include priests, lineage and clan elders, rainmakers, diviners, prophets, and others. Very few of these are specialists; typically they hold ritual authority by virtue of age, genealogy, or political office and are primarily responsible for the ritual well-being only of the members of the social groupings that they head; their congregations consist of their joint families, lineages, clans, local village communities, chieftaincies, or the like. Their ritual authority is thus a sanction for their secular and domestic authority.

A central element of every indigenous African religion is its cosmology—which tells of tribal origins and early migrations and explains the basic ideological problems of any culture, such as the origin of death, the nature of society, the relationship of men and women, of living and dead, and so on. Social values are typically expressed in myths, legends, folktales, and riddles; the overt meanings of these various oral statements frequently conceal sociological and historical meanings not easily apparent to outsiders.

In the past, witchcraft and sorcery were given widespread credence and served to explain or control the misfortunes of people who were aware of their lack of mastery and understanding of nature and society. Travellers' tales of African people living in fear of witchcraft, however, were of course grossly exaggerated; the colonial powers usually assumed (incorrectly) that witch doctors were socially harmful and prohibited them. Today, though belief in witchcraft is receding, it is still important in both rural and urban areas, often serving as a means of explaining the misfortunes that beset urban dwellers and labour migrants who find themselves in new and confusing social milieus. There have been many cases in recent times of "epidemics" of beliefs in witches, and there have arisen a number of evanescent cults led by various kinds of prophets and evangelists. Such manias arise in periods of radical change and its resultant uncertainty and stress.

Nativistic movements

Social and religious changes in Africa have often been accompanied by the appearance of prophets prophesying the expulsion of the Europeans or the eradication of epidemic diseases threatening the traditional ways of life. Recently, also, the spread of Christianity and, to a lesser extent, Islām has given rise to Christian prophets and to leaders of separatist movements repudiating European-controlled mission churches for nationalistic, tribalistic, or racial reasons. Such prophets lead their own cults and establish their own churches, typically gaining new political power sanctioned by their presumed direct links with God. These churches have been reported in almost all parts of the continent (see also PRIMITIVE RELIGION; TRIBAL RELIGIOUS MOVEMENTS, NEW).

BIBLIOGRAPHY. Four books giving a general background picture are LORD HAILEY, *An African Survey* (1957); MELVILLE J. HERSKOVITS, *The Human Factor in Changing Africa* (1962); GUY HUNTER, *The New Societies of Tropical Africa* (1962); and PAUL BOHANNAN and PHILIP CURTIN, *Africa and Africans*, rev. ed. (1971). Accounts of both traditional and modern African societies and their cultures are in SIMON and PHOEBE OTTENBERG, *Cultures and Societies of Africa* (1960); JOHN MIDDLETON, *Black Africa: Its Peoples and Their Cultures Today* (1970); JAMES L. GIBBS, *The Peoples of Africa* (1965); and PETER MCEWAN and ROBERT B. SUTCLIFFE, *Modern Africa* (1965). More ethnologically oriented surveys include GEORGE P. MURDOCK, *Africa: Its Peoples and Their Culture History* (1959); and C. DARYLL FORDE (ed.), *Ethnographic Survey of Africa* (1950–), with 56 volumes to date.

J.D. CLARK, *The Prehistory of Africa* (1970); and ROLAND A. OLIVER, *The Dawn of African History* (1961), are good accounts of prehistorical beginnings. Books dealing with medieval and later history include BASIL DAVIDSON, *The African Genius* (1969) and *Africa in History* (1968); ROLAND A. OLIVER and J.D. FAGE, *A Short History of Africa* (1962); and ROLAND A. OLIVER, *The Middle Age of African History* (1967). JOSEPH H. GREENBERG, *Studies in African Linguistic Classification* (1955), provides the most generally accepted account of African languages. Books on African economic systems include KENNETH BARBOUR and R.M. PROTHERO (eds.), *Essays on African Population* (1961); PAUL BOHANNAN and GEORGE DALTON, *Markets in Africa* (1962); WILLIAM A. HANCE, *African Economic Development* (1967); and RENE DUMONT, *L'Afrique noire est mal partie* (1962; Eng. trans., *False Start in Africa*, 1969). Traditional forms of government are considered in JACQUES MAQUET, *Pouvoir et société en Afrique* (1971; Eng. trans., *Power and Society in Africa*, 1971); LUCY MAIR, *Primitive Government* (1962); and MAX GLUCKMAN, *Order and Rebellion in Tribal Africa* (1963). Among the many works on more modern political developments are MARGERY PERHAM, *The Colonial Reckoning* (1961); T.R. ADAM, *Government and Politics in Africa South of the Sahara*, 3rd ed. rev. (1965); THOMAS L. HODGKIN, *Nationalism in Colonial Africa* (1956); IRVING L. MARKOVITZ, *African Politics and Society* (1970); RUPERT EMERSON and MARTIN KILSON, *The Political Awakening of Africa* (1965); ROBERT W. JULY, *The Origins of Modern African Thought* (1968); GWENDOLEN M. CARTER, *National Unity and Regionalism in Eight African States* (1966); JAMES S. COLEMAN and CARL G. ROSBERG (eds.), *Political Parties and National Integration in Tropical Africa* (1964); and LEOPOLD S. SENGHOR, *Rapport sur la doctrine et le programme du parti* (1959; Eng. trans., *African Socialism: A Report to the Constitutive Congress of the Party of African Federation*, 1959). Other aspects of changing society in Africa include PETER C. LLOYD, *The New Elites of Tropical Africa* (1966); HORACE MINER, *The City in Modern Africa* (1967); and ARTHUR PHILLIPS, *Survey of African Marriage and Family Life* (1953). Religion, both traditional and modern, is dealt with by C. DARYLL FORDE, *African Worlds: Studies in the Cosmological Ideas and Social Values of African Peoples* (1954); EDWARD G. PARRINDER, *Religion in Africa* (1969); PLACIDE TEMPELS, *La Philosophie Bantoue* (1949; Eng. trans., *Bantu Philosophy*, 1969); and J.S. TRIMINGHAM, *The Influence of Islam upon Africa* (1968). The most useful bibliography is published by the International African Institute, *Select Annotated Bibliography of Tropical Africa* (1969).

(J.F.M.)

Africans, Prehistoric

The reconstruction of the prehistory of human or human-like (hominid) animals in sub-Saharan Africa is made easier than in many other parts of the world by a unique combination of circumstances. The tropical climate and environment suited hunter-gatherers so well that the area has been populated from the time of the earliest known appearance of hominids. Secondly, conditions for the preservation of bone are often favourable, except in the equatorial forests; in the Rift Valley, with its ancient lake beds and volcanic ash, these conditions are exceptionally good. Thirdly, artifacts are often found *in situ*, and undisturbed living floors sometimes survive intact, sealed in by later deposits. With skillful excavation these floors reveal distribution patterns of bones and implements, from which the activities of prehistoric man may be inferred (see also HOMINIDAE for illustrations of skeletal features discussed here). Living sites of the earliest hominids are unique to Africa and consequently have received much attention, but there are also camping places of later Acheulean and Middle Stone Age times, the period under consideration here. As the earliest representatives of *Homo sapiens* in Africa, these latter groups are of particular interest.

AGES AND DISTRIBUTION OF FOSSIL REMAINS

Radiocarbon dating suggests that the Earlier Stone Age (ESA), with its immensely long Acheulean tradition (characterized by bifacial stonecutting tools), may have lingered on until about 55,000 BC—long after the Acheulean had come to an end in western Europe. It was followed by a relatively short transitional period and then by the Middle Stone Age (MSA), which lasted from about 40,000 to 10,000 BC. Beyond the limit of radiocarbon dating, it is sometimes possible to date human remains by the potassium-argon method if they are contained in volcanic tuffs or overlain by suitable lavas; other radioactive materials are also used (see DATING, RELATIVE AND ABSOLUTE). In other cases attempts at correlation must be made on the basis of the accompanying remains of fauna or on evidence of tool industry. A great many sites have a profusion of stone tools and some also have fossil fauna, but the remains of the men who made the implements and hunted the animals are exceedingly rare.

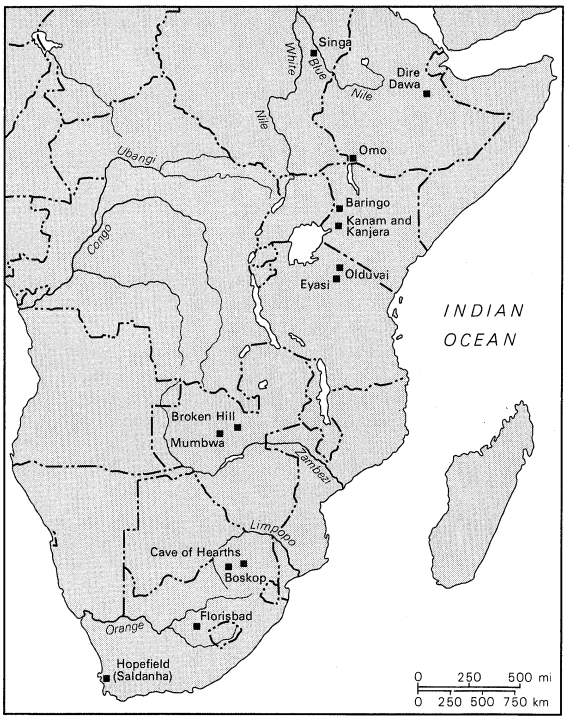
Relationships of the genus *Homo* in sub-Saharan Africa

Until very recently, there was a long gap in the record between the *H. erectus* skull from the upper part of Bed II at Olduvai Gorge, Tanzania (Chellean man), dated by potassium-argon to about 500,000 years ago, and the late stages of the ESA. New discoveries from the Omo Valley in southern Ethiopia, tentatively dated by uranium-thorium to about 130,000 years ago, are important in being the earliest known *H. sapiens* in Africa, representing a stage at which man had just become distinct from *H. erectus*. This type may have been ancestral to both the Rhodesioids, *H. sapiens rhodesiensis*, who were widespread in sub-Saharan Africa from about 50,000 to 35,000 years ago, and to the Negro-Bushmanoid stock, *H. sapiens afer*. There are plenty of examples of Bushmanoid remains after about 30,000 BC, but the ancestors of this stock may already have been present in eastern Africa before the end of the ESA.

From the period between 100,000 and 20,000 years ago, human bones have been found at only about a dozen sites, stretching from the Sudan to the Cape of Good Hope. Most come from East Africa, where they are associated with volcanic centres and ancient lake deposits in the Rift Valley; others come from caves in Ethiopia, Zambia, and the Transvaal; from a mineral spring in the Orange Free State; and from dune sand near Saldanha Bay, Cape Province. Although artifacts show that West Africa was inhabited during the MSA, no human remains are known before the end of the Pleistocene (perhaps 10,000 years ago).

MORPHOLOGICAL CHARACTERISTICS AND POPULATION VARIATION

Except for small fragments, less than a dozen skulls and only five mandibles are known from the late ESA and the greater part of the MSA (a few more appear toward the end of the MSA, after 20,000 BC). Except in one case (Saldanha Bay), none of the skulls is associated with mandibles (lower jaws); the other jaws are isolated. With such meagre material, only very broad generalizations can be made about the inhabitants of sub-Saharan Africa during this period.



Sites of fossil man in sub-Saharan Africa dating from about 130,000 to 20,000 BC.

Early *Homo sapiens* in sub-Saharan Africa

As in Europe and Asia, after the *H. erectus* stage man appears to have been in a "fluid" state, evolving in different directions. One extreme of this range is represented by the gerontomorphic ("adult") Rhodesioids, which are typified by a rugged, heavy-browed skull from Broken Hill (now Kabwe, Zambia). At the other end of the evolutionary spectrum are modern-looking, smooth-browed, pedomorphic ("infantile") people from Kanjera, western Kenya. These may have been the forerunners of the Bushmanoid stock, of which many examples appear at the end of MSA times.

Even from the few instances in which more than one individual has been preserved at a single site, it is clear that there was great variation within small populations. Male skeletal remains often are readily distinguishable from female; that is, sexual dimorphism was evidently very marked. As more human remains are found, they demonstrate how unwise it is to regard a single specimen as a "type." The picture is further complicated by the possibility that several of the known fossils were distorted by pathological conditions (*e.g.*, diseases of the bone).

The Omo skulls. There is no more striking example of individual variation than in three skulls discovered in 1967 in the Omo Valley, north of Lake Rudolf, Ethiopia. They were accompanied by typical Middle Pleistocene fauna; uranium-thorium dating gave an age of 130,000 years. The skull known as Omo II, an elderly male, has a low vault and receding forehead, with a massive occipital torus above a very flat nuchal area (nape) bearing strong impressions where the powerful neck muscles were attached. Omo I, almost certainly a female, is far more modern in appearance. The bones, though still robust, are more lightly built, and the occipital (back of the skull) region is more rounded than that found in Omo II. Rather surprisingly, the chin is fairly well developed. The Omo skulls, although certainly a subspecies of *H. sapiens*, show definite traces of their *H. erectus* ancestry and may be on a direct evolutionary

line from the skull in Bed II at Olduvai Gorge. In certain features they also foreshadow the Rhodesioids, but most of all they resemble early *H. sapiens* fossils from sites at Vértesszőllős, Hungary; Swanscombe, England; and Steinheim, Germany.

Two other isolated mandibles may also be examples of this evolutionary stage; one may be roughly the same age as the Omo skulls, and the date of the other is unknown. A robust jaw with receding chin from Middle Pleistocene beds west of Lake Baringo, Kenya, resembles *H. erectus* mandibles from Ternifine in Algeria. The Kanam jaw, discovered in 1932 at Kanam on the shores of the Kavirondo Gulf in western Kenya, has for long been a puzzle. Its apparent association with Lower Pleistocene fauna seems to be contradicted by the incipient chin. This prominence in the chin region is due in part to a pathological condition, but it is known from Omo I that hominid jaws of the Middle Pleistocene were not invariably chinless. The uranium content of the Kanam mandible was found to be very low in comparison with that of accompanying Lower Pleistocene mammal fossils; hence the jawbone may perhaps date from the Middle Pleistocene.

Kanjera man. Four fragmentary skulls found in 1932 at Kanjera, near Kanam, are also something of an enigma. Uranium tests confirm that they are contemporary with Middle Pleistocene fauna from the area and hence presumably with the Acheulean hand axes accompanying them. The skulls probably date back to at least 55,000 BC, the termination date in Africa for the Acheulean culture, but there is no way of telling how much older they are than that. Unlike others of this age in Africa, two of the skulls are completely smooth browed. In one, however, the brow ridges are more pronounced. Again, there is individual variation within a small group. The skulls, as reconstructed from many small fragments, are extremely long and narrow (ultradolichocephalic). One piece suggests that the facial skeleton was small, as that found in the modern Bushmen, and it seems possible that the Kanjera people were ancestral to the Negro-Bushmanoid stock.

The Rhodesioids. Perhaps not much later than the Kanjera people were the first of the very different Rhodesioids, *H. sapiens rhodesiensis*. The best known, though not the earliest, is a skull found in 1921 in a deep cave at lead and zinc mines at Broken Hill. It was so unlike any other known skull that it gave some anthropologists, no less than the miners who found it, quite a shock. It is now known that this general type is not exceptional in Africa, although this particular individual may have been rather extreme.

Examples of Rhodesioid *Homo sapiens*

The Broken Hill skull is one of the most massive known, with an enormous bony ridge above the eyes (supraorbital torus) and an unusually long face. The forehead is receding and there is a marked constriction behind the brow ridges. The nuchal area (for attachment of neck muscles) below the prominent ridge at the back of the skull (occipital torus) is flat, as in Omo II, and the foramen magnum (the hole at the base of the skull that accommodates the spine) is placed well forward and faces downward, indicating that this type of man carried his head erect. The limb bones, which were found separately, are stout and long, indicating that the individual was tall and walked like modern man. Small pieces of bone from the skulls of two other individuals were also found; one of them is less robust and may have been a female. The accompanying fauna consists of modern species, with the exception of a giant buffalo, and the associated tool industry (Sangoan–Proto–Stillbay) is typical of the early MSA, about 40,000 BC.

A rather similar industry was associated with two fragmentary crania from old lake deposits by the northeastern shore of Lake Eyasi, Tanzania. The more complete remains, suggesting a female of the Rhodesioid type, were lost in Germany during the 1939–45 war. Some fragments, including a skull cap, from Olduvai Bed IV may also be Rhodesioid; the industry of this bed is Acheulean, earlier than the industry at Eyasi.

Probably earlier than the Broken Hill and Eyasi specimens is another Rhodesioid from Hopefield, near Saldanha Bay. Pieces of the skull of a young adult were found in 1953, and others, including a mandible, turned up later. Together with abundant fossil mammals and stone artifacts, they came from lime-impregnated sand on the inner side of a line of dunes. Fluorine and uranium tests suggest that the human remains are contemporary with extinct fauna and hence presumably with the oldest of the three industries, the Fauresmith. This transitional culture between the ESA and the MSA probably dates from about 50,000 BC. The vault of the Hopefield skull is even lower than in the Broken Hill specimen, the forehead is receding, and the supraorbital torus is prominent. The bones are very thick, and the mandible is also extremely robust; if the mandible had been found without the skull, it might well have been assigned to *H. erectus*.

The right side of a child's jaw from the Cave of Hearths in the Makapan Valley, Transvaal, may be slightly earlier; this very thick jaw, with poorly developed chin and large molar teeth, was associated with a final Acheulean–Fauresmith industry.

Perhaps latest in time of all the Rhodesioids is a mandible excavated from ash in the floor of the Porc Epic (Porcupine) Cave near Dire Dawa, Ethiopia; at the time of its excavation (1933) it was labelled as a "Neanderthaloid," but this name was also given to the Broken Hill skull in the past. It may be contemporary with Stillbay artifacts in the cave floor, in which case it could be as recent as about 25,000 BC.

Negro-Bushmanoid remains. An aberrant skull found in 1932 at the mineral spring of Florisbad near Bloemfontein, Orange Free State, may be a representative of the Negro-Bushmanoid stock, of which there were already hints in the Kanjera skulls. It came from a pocket of sand within peaty clay, dated by radiocarbon to about 33,000 BC, and was accompanied by fauna and crude tools. The skull is extremely large, with a rather low vault and an exceptionally broad frontal region; it has been suggested that this might be the result of some congenital or growth deformity. The brow ridges are fairly prominent, the eye sockets are rectangular and set low, and the face is wide and prognathous (characterized by a protruding mouth area).

Examples of Negro-Bushmanoid *Homo sapiens*

Another apparently aberrant skull is from Singa on the Blue Nile in the Sudan. It was first described as "Bushmanoid," mainly because its bulging parietal bones (at the sides of the skull) gave it a pentagonal shape when viewed from above. It has also been suggested that the individual was abnormal, the growth from front to back having been stunted, with a compensatory growth from side to side; nevertheless the skull probably belonged to a member of the Negro-Bushmanoid stock. The fauna includes the extinct giant buffalo already mentioned and a rather indeterminate tool industry that has been described as Proto–Stillbay but is rather more advanced than the industry of Broken Hill man.

Somewhat later, associated with a Stillbay industry, are three individuals from Mumbwa Caves, Zambia, who are certainly members of the Bushmanoid stock. Another is an outstandingly large skull from Boskop in the Transvaal found in 1913 and the first human fossil to be discovered in South Africa. The bossed (protruding) parietals and bulging forehead, both pedomorphic features, are typically Bushmanoid. The dating is uncertain, but a single implement found with the Boskop skull suggests the late MSA. Several other skulls of this large Bushmanoid type are known from the late MSA and the ESA in South Africa.

INFERRED BEHAVIORAL CHARACTERISTICS

Throughout the ESA, hunter-gatherers camped in open savanna and grasslands country, avoiding densely wooded areas. Their camp sites were located near lakes or rivers, where they could ambush game coming down to drink. Evidence from animal bones on these sites shows that the Acheuleans were successful hunters of

large game. The emphasis on large cutting tools—hand axes and cleavers—appropriate for skinning and cutting off joints of meat supports this inference.

There are no depths of deposit to suggest long occupation in any one place; these sites must have been seasonal camping places, though evidently many of them were revisited year after year. Not until about 50,000 years ago, at the end of the ESA in sub-Saharan Africa, is there any evidence of semipermanent occupation of caves and rock shelters.

Changes at the end of the Early Stone Age. The climate in sub-Saharan Africa during this transitional period between the ESA and MSA was relatively dry, but throughout most of the MSA it was wetter and cooler than it had been during much of Acheulean times. This climatic information is derived from traces of higher lake levels and from pollen analysis at such sites as Florisbad and Kalambo Falls at the southern end of Lake Tanganyika in Zambia. These climatic changes were not dramatic, but they must have affected the vegetation pattern and hence the movements of animals and of the hunters who pursued them.

Many Middle Pleistocene mammals became extinct at this time and were replaced by modern forms, although a few, such as a giant buffalo and an antlered giraffe-like animal, survived into the MSA. The cause of these extinctions was not necessarily climatic, but this period was certainly one of faunal change. Fewer remains of large kills are found at MSA hominid sites than are discovered at ESA sites; perhaps large game animals were becoming rarer, or it could be that these prehistoric men began to change their hunting methods.

There was a great technological change after the end of the sub-Saharan ESA. The long tradition of large cutting tools was dying out, to be superseded by heavy-duty chopping tools and a great variety of small flake tools or light-duty equipment. Many specimens of this changing tool industry suggest more emphasis on woodworking, and there is also evidence that man began to move into more wooded country at this time. Ancient wooden artifacts seldom survive, but the waterlogged lake beds at Kalambo Falls have preserved some pointed and obliquely truncated sticks that may have been used for digging, as well as a clublike weapon. The butt end of a throwing stick with a prepared grip was associated with the Florisbad skull, and round perforations found in both this skull and that of Broken Hill could have been inflicted by the point of a wooden spear.

The chopping and scraping of wood is much easier if the material is charred, and the first evidence in Africa of the use of fire is found in this transitional period between the ESA and the MSA. Compared with Asia and Europe, this is surprisingly late; presumably the warmer climate of Africa made the use of fire for warmth less urgent, and the advantages of cooked meat may have been discovered accidentally.

One of the earliest sites with evidence of fire is the Cave of Hearths in the Makapan Valley, Transvaal. Ash more than a metre deep (from bat guano) may have been the result of accidental ignition by a bush fire. Above this basal deposit of ash are hearths at ten levels, ranging in time from a transitional Acheulean–Fauresmith occupation right through the MSA to the Late Stone Age (LSA). At Kalambo Falls, charcoal, charred wood, and ash occur on five of the Late Acheulean floors at the sites of human activity.

Caves and rock shelters occur only in certain parts of Africa; elsewhere, hunters must have lived in the same way as their forebears and very much as the Kalahari Bushmen do today. Modern Bushmen construct temporary shelters; an arc of stones on a Late Acheulean floor at Kalambo was perhaps used to weight some sort of windbreak used by early man. Carbonized grass stems and twigs lying in shallow depressions are reminiscent of contemporary Bushmen's bedding.

Inferences about the diet of MSA people can be made both from their surviving equipment and by analogy with such modern hunter-gatherers as the Bushmen. A typical modern Bushman group kills only about 15–20 large animals in the course of a year, and their mainstay is vegetable food, with the meat of small mammals, reptiles, and birds as a supplement. Fruit, nuts, and seeds preserved in the Kalambo floors are of species still collected and eaten by local Bantu tribes. The primary use of stone balls, found at many MSA sites, must have been for pounding and grinding vegetable foods, though they may also have been used as missiles in hunting. Almost perfectly rounded stones were found at Broken Hill, and the state of decay of ancient Rhodesian man's teeth suggests a starchy and gritty diet. On the other hand, bone gouges were also part of his equipment, and these may have served as marrow scoops. The Kalambo digging sticks were presumably used for grubbing up roots and tubers in the way that a contemporary Bushwoman uses her digging stick.

Food collection methods and hunting

Hunting methods probably varied according to the environment. In open country, animals could be driven into swamps and over cliffs; there is evidence of this in ESA times at Olduvai, but perhaps during the MSA game was also driven by the deliberate lighting of bush fires. During the dry season, game would have concentrated at the few remaining waterholes and would have been easy to ambush. The human remains and enormous numbers of the bones of other animals at Hopefield (Saldanha) lay on the floors of old pans, which once contained water. These are similar to the present Etosha Pan in southwest Africa, where animals from different environments gather to drink. At Florisbad, also, animals would have concentrated around the spring during the very dry climate that prevailed at the time when the owner of the human skull lived.

Along forest edges and in wooded country, fall-traps and pits may have been made. The construction of stakes would have been quite easy with the new woodworking tools and the control of fire. Notched scrapers could also have been used for preparing cord from vegetable fibres to make nooses and trip lines for small game and birds, while ropes could also have helped in climbing trees to obtain honey.

The Fauresmith industry. The Fauresmith and the Sangoan cultural traditions that followed the Acheulean were adapted to different kinds of environment. The Fauresmith, associated with the Saldanha skull, is found mostly in the grasslands and higher country of South Africa, and in the highlands of Kenya and Ethiopia. It is more conservative than the Sangoan, which would be expected since conditions in the open country where the Fauresmith people lived were essentially the same as in Acheulean times. Hand axes and cleavers continued to be made, though they were smaller than Acheulean specimens and often almond shaped. Other Fauresmith equipment included broad side scrapers and end scrapers, trimmed points, long blades and chisels, and also stone balls.

The Sangoan industry. The Sangoan tradition has a far wider distribution than the Fauresmith, spreading into hitherto unoccupied woodland and forest areas, across the Congo Basin into West Africa, in country with more than 30 inches average annual rainfall at the present time. Dates of about 40,000 years ago have been obtained for Sangoan industries at Kalambo and at Mufo in northeast Angola. This culture was thus contemporary with the Mousterian industry of Neanderthal man (*q.v.*) in Europe.

Heavy-duty equipment in the Sangoan includes core axes, picks, and core scrapers. Core axes have a thick, heavy, unworked butt and were probably used in the manner of an adze for woodworking. The roughly circular core scrapers, flat on one face and with a steep angle of retouch on the other, may have been used for working skins and also for planing wood. Among the light-duty tools are small, serrated scrapers that could have served to groove and to gouge wood, while others might have functioned as spokeshaves to scrape rounded sticks.

Many Middle Stone Age sites are littered with cores,

flakes, and tools. Since nomads travel light and it probably was easy to make new equipment, tools would have been discarded when the people moved. In regions where good raw material was scarce, however, suitable rock was carried over great distances. The material affected the technology: in large parts of South Africa, for instance, a hard, homogeneous, indurated shale was used by the Fauresmith people; to produce a strong cutting edge, the tool was retouched by step flaking. They used a hammerstone or tapped the implement on a stone anvil to break the flakes off in a series of steps. When the available material consisted of large boulders or lumps struck from the parent rock, flake tools based on the prepared core technique were often made. These became more common as the MSA developed.

The Stillbay and Lupemban industries. After about 35,000 years ago, when the Rhodesioids were being replaced by men of modern type, specialization of industries in different regions began to be more marked. The Stillbay tradition, with various local forms, is characteristic of the open country of eastern and southern Africa, while in central and western Africa the Sangoan industry was succeeded by the Lupemban. These two traditions continued until the end of the MSA.

Several important technological improvements were introduced during this period. Fine pressure flaking was achieved all over the surfaces of leaf-shaped points in the Stillbay and on beautifully made long lance heads in the Lupemban. Mastery over even the most intractable material is seen, for instance, in fine quartz points found at Bambata Cave in the Matopo Hills, Rhodesia. Some of these weapons were almost certainly hafted, since the butts are often trimmed to make them thin; scrapers and knives were probably also hafted to avoid damaging the hands. Mastic obtained from plants would have held the implement in place, and a binding of vegetable fibres or sinews would have been applied.

Hafting and the arrow invented

The practice of hafting probably was an extremely important step, leading toward the invention of the arrow. During wetter periods in the MSA, movements across parts of the Sahara would have been possible, and the Stillbay people almost certainly came into contact with Aterian peoples of North Africa. The smaller Aterian points with trimmed butts may have been arrowheads, and Stillbay points may have served the same purpose. Hunters able to kill at a distance would have had a tremendous advantage, and the bow and arrow was widespread during the LSA; there are strong hints that this breakthrough was made during the MSA.

There is little evidence of anything other than the material culture of the MSA people. No formal burials have been found, and no rock art has been discovered that is attributable to this period. "Pencils" of the mineral hematite and other colouring materials have occasionally been found with Stillbay tools, but these may have been for decorating the body rather than for painting the walls of shelters.

PHYLOGENETIC AFFINITIES WITH LATER AFRICAN POPULATIONS

The present indigenous inhabitants of sub-Saharan Africa fall into three groups: Negroid, Khoisan (Hottentots and Bushmen), and "Caucasoid" (Eastern Hamites). These groups may be easily distinguished by external features such as skin colour and hair form, but in skeletal features alone there is a good deal of overlap even today, when they have probably become increasingly divergent from their more generalized ancestors. From fragmentary fossil remains, therefore, it is difficult to distinguish among them. Negroids, for instance, typically have a narrow skull and rounded forehead, but Eastern Hamites also tend to have a narrow skull and rounded forehead, and Bushmen also have a rounded forehead. A protruding upper jaw is characteristic of Negroids, but this part of the face is not always preserved in fossil remains.

Relationship of present racial stocks to fossil man

There is no certain evidence of the presence of Caucasoids in sub-Saharan Africa before post-Pleistocene times, when presumably they spread southward into the Horn (the region of modern Somalia and Ethiopia) from either North Africa or western Asia. For present purposes, therefore, only the Negroids and the Khoisan people need be discussed.

Blood-group evidence shows that the Khoisan group must be closely related to the Negroids, although at the present time they are very different in skin colour and appearance. Peoples living in specialized environments, either from choice or as a result of pressure from other groups, become adapted physically in response to intensive selective pressure. Small stature, for instance, is an advantage in forests where there is a shortage of protein. Pygmies may well have had taller Negroid ancestors, and Bushmen, who have been forced to retreat into the Kalahari Desert, are evidently the remnants of a once widespread stock, probably physically much larger, but known only from their larger braincases.

Negroids and Khoisan people may have shared a common ancestor as far back as the ESA, at a stage represented by the Kanjera remains. The problem of the Rhodesioids is essentially the same as that of the Neanderthalers elsewhere: Did they hybridize with more modern forms of *Homo sapiens?* Did they simply become extinct? Indulging in pure speculation, it may be that at least the less extreme Rhodesioids made a genetic contribution to such forms as the Florisbad and Singa individuals. It is known that the makers of Stillbay artifacts, such as the Mumbwa people, were Bushmanoid. Although no Negroids can be recognized with certainty before post-Pleistocene times, it seems reasonable to suggest that the contemporaries of the Stillbay people in the more forested areas of western and central Africa, the Lupembans, were Negroid.

BIBLIOGRAPHY. J.D. CLARK, *The Prehistory of Africa* (1970), a comprehensive book on Stone Age Africa, with excellent illustrations and a valuable appendix on dates and tool compositions of industries; with many illustrations and an extensive bibliography; "The Later Pleistocene Cultures of Africa," *Science*, 150:833–847 (1965), an account of the economy and ecology of the Middle Stone Age; S.M. COLE, *The Prehistory of Africa*, rev. ed. (1963), the only book covering prehistoric archaeology and human remains in eastern Africa as a whole, with many illustrations and an extensive bibliography; K.P. OAKLEY and B.G. CAMPBELL (eds.), *Catalogue of Fossil Hominids, Part I: Africa* (1967), a comprehensive catalog of African fossil human remains with their geological and archaeological contexts, ages, and bibliography.

(S.M.Co.)

Agassiz, Louis

A 19th-century naturalist, geologist, and teacher who made revolutionary contributions to the study of natural science with landmark work on glacier activity and extinct fishes, Louis Agassiz achieved lasting fame through his innovative teaching methods, which were completely to alter the character of natural science education in the United States.

Agassiz was born on May 28, 1807, the son of the Protestant pastor of Motier, a village on the shore of Lake Morat, Switzerland. In boyhood he attended the gymnasium in Bienne and later the academy at Lausanne. He entered the universities of Zürich, Heidelberg, and Munich, and took at Erlangen the degree of doctor of philosophy and at Munich that of doctor of medicine.

As a youth he gave some attention to the ways of the brook fish of western Switzerland, but his permanent interest in ichthyology began with his study of an extensive collection of Brazilian fishes, mostly from the Amazon River, which had been collected in 1819 and 1820 by two eminent naturalists at Munich. The classification of these species was begun by one of the collectors in 1826, and when he died the collection was turned over to Agassiz. The work was completed and published in 1829 as *Selecta Genera et Species Piscium.* The study of fish forms became henceforth the prominent feature of his research. In 1830 he issued a prospectus of a *History of the Fresh Water Fishes of Central Europe*, printed in parts from 1839 to 1842.



Agassiz, between 1855 and 1865.
By courtesy of the Museum of Comparative Zoology, Harvard University

The year 1832 proved the most significant in Agassiz's early career because it took him first to Paris, then the centre of scientific research, and later to Neuchâtel, where he spent many years of fruitful effort. While in Paris he lived the life of an impecunious student in the Latin Quarter, supporting himself and helped at times by the kindly interest of such friends as the German naturalist Alexander von Humboldt—who secured for him a professorship at Neuchâtel—and Baron Cuvier, the most eminent ichthyologist of his time.

Study of extinct fishes

Already Agassiz had become interested in the rich stores of the extinct fishes of Europe, especially those of Glarus in Switzerland and of Monte Bolca near Verona, of which, at that time, only a few had been critically studied. As early as 1829 Agassiz planned a comprehensive and critical study of these fossils and spent much time gathering material wherever possible. His epoch-making work, *Recherches sur les poissons fossiles*, appeared in parts from 1833 to 1843. In it, the number of named fossil fishes was raised to over 1,700, and the ancient seas were made to live again through the descriptions of their inhabitants. The great importance of this fundamental work rests on the impetus it gave to the study of extinct life itself. Turning his attention to other extinct animals found with the fishes, Agassiz published in 1838–42 two volumes on the fossil echinoderms of Switzerland, and later (1841–42) his *Études Critiques sur les Mollusques Fossiles*.

From 1832 to 1846 he served as professor of natural history in the University of Neuchâtel. In Neuchâtel he acted for a time as his own publisher, and his private residence became a hive of activity with numerous young men assisting him. He now began his *Nomenclator Zoologicus*, a catalog with references of all the names applied to genera of animals from the beginning of scientific nomenclature, a date since fixed at January 1, 1758.

In 1836 Agassiz began a new line of studies: the movements and effects of the glaciers of Switzerland. Several writers had expressed the opinion that these rivers of ice once had been much more extensive and that the erratic boulders scattered over the region and up to the summit of the Jura Mountains were carried by moving glaciers. On the ice of the Aar Glacier he built a hut, the "Hôtel des Neuchâtelois," from which he and his associates traced the structure and movements of the ice. In 1840 he published his *Études sur les glaciers*, in some respects his most important work. In it, Agassiz showed that at a geologically recent period Switzerland had been covered by one vast ice sheet. His final conclusion was that "great sheets of ice, resembling those now existing in Greenland, once covered all the countries in which unstratified gravel (boulder drift) is found."

In 1846 Agassiz visited the United States for the general purpose of studying natural history and geology there but more specifically to give a course of lectures at the Lowell Institute in Boston. The lectures were followed by another series in Charleston and, later, by both popular and technical lectures in various cities. In 1847 he accepted a professorship of zoology at Harvard University; and in 1850, after his first wife's death, he married Elizabeth Cabot Cary of Boston, who was well-known as a writer and a promoter of women's education.

Activities in the United States

In the United States his chief volumes of scientific research were the following: *Lake Superior* (1850); *Contributions to the Natural History of the United States* (1857–62), in four quarto volumes, the most notable being on the embryology of turtles; and the *Essay on Classification* (1859), a brilliant publication, which, however, failed to grasp the fact that zoology was moving away from the doctrine of special creation toward the doctrine of evolution. Besides these extensive contributions there appeared a multitude of short papers on natural history and especially on the fishes of the U.S. His two expeditions of most importance were, first, to Brazil in 1865, and second, to California in 1871, the latter trip involving both shores of South America. *A Journey in Brazil* (1868), written by Mrs. Agassiz and himself, gives an account of their experiences. His most important paper on U.S. fishes dealt with the group of viviparous surf fishes of California.

Agassiz was deeply absorbed in his cherished plan of developing at Harvard a comprehensive museum of zoological research. This institution, which was established in 1859 and ultimately grew into the present museum of comparative zoology, enjoyed his fostering care during the rest of his lifetime. In the U.S., Agassiz's industry and devotion to scientific pursuits continued, but two other traits now assumed importance. Quite possibly he was the ablest science teacher, administrator, promoter, and fund raiser in the U.S. in the 19th century. In addition, he was devoted to his students, who were in the highest sense co-workers with him.

Agassiz's method as teacher was to give contact with nature rather than information. He discouraged the use of books except in detailed research. The result of his instruction at Harvard was a complete revolution in the study of natural history in the U.S. The purpose of study was not to acquire a category of facts from others but to be able, through active contact with the natural world, to gather the needed facts. As a result of his activities, every notable teacher of natural history in the U.S. for the second half of the 19th century was either a pupil of Agassiz or of one of his students.

In the interests of better teaching and of scientific enthusiasm, he organized in the summer of 1873 the Anderson School of Natural History at Penikese, an island in Buzzards Bay. This school, which had the greatest influence on science teaching in America, was run solely by Agassiz. After his death, December 14, 1873, it vanished.

Agassiz and Darwin

Because Agassiz was beyond question one of the ablest, wisest, and best informed of the biologists of his day, it may be asked why his attitude toward Darwin's *Origin of Species*, published in 1859, was cold and unsympathetic. It is likely that Agassiz's lifelong view of nature determined his attitude toward the new doctrine of evolution. Although Agassiz was quite familiar with the factual evidence concerning environmental change, variability, and hereditary modification on which Darwin built his arguments, he held that the organic world represented repeated interventions by the Supreme Being. Ordinary physical events on which Darwin relied, such as climatic and geologic change, and even glaciers, could bring about extinctions, but not new species. The sequence in the fossil record from simple animals and plants in the ancient, deeper strata to the more complex, recent forms found near the surface represented a progressive development, Agassiz agreed, but these different animals and plants did not arise because of interactions between populations and external environmental changes, as Darwin argued. Agassiz maintained that since organisms arose by a series of independent and special creations, there could be no hereditary continuity between

different types of organisms. Each species of plant and animal was a "thought of God," and homologies or anatomical similarities were "associations of ideas in the Divine Mind:" Agassiz's view of nature was historically derived from the thought of Plato, for whom the unseen world had more reality than the world of sense experience. Agassiz therefore could not accept Darwin's conceptual view of nature, in which environmental events could evoke organic change.

Agassiz died on December 14, 1873. He was buried in Mt. Auburn at Cambridge, and by his grave stands a boulder from the moraine of the Aar Glacier.

BIBLIOGRAPHY. The definitive biography is EDWARD LURIE, *Louis Agassiz: A Life in Science* (1960; abridged ed., 1966), a substantial, documented analysis of his scientific thought. It contains an extensively annotated bibliography relating to earlier biographical materials and particular aspects of his scientific career.

(D.S.J.)

Agency, Law of

The law of agency governs the legal relationship in which one person (the agent) deals with another (third party) on behalf of still another (the principal). The competent agent is legally capable of acting for this principal vis-à-vis the third party. Thus, the process of concluding a contract through an agent involves a twofold relationship. On the one hand, the law of agency is concerned with the *external* business relations of an economic unit and with the powers of the various representatives to affect the legal position of the principal. On the other hand, it rules the *internal* relationship between principal and agent as well, thereby imposing certain duties on the representative (diligence, accounting, good faith, etc.). The two relationships need not be in full conformity. Thus, an agent's effective powers in dealing with outsiders may extend to transactions that he is under a duty to his principal not to undertake, leading to a situation characterized as "apparent authority."

Agency is recognized in all modern legal systems as an indispensable part of the existing social order. It fulfills the most diverse functions in both public and private law; in particular, it assists in organizing the division of labour in the national and international economy by making it possible for a principal greatly to extend his individual sphere of activity by having one or more persons act for him in his stead. In addition to the individual principal, a principal may be composed of a group of persons carrying on a trade or business by way of a partnership, a registered company, or another kind of "corporate" entity. The need for legal representation in some form has therefore increased as business units have come to involve transactions conducted at a distance (as in the case of factors) or have grown in size (as in the case of the firm, the house, and the corporation). Continental law additionally uses "legal" representatives, such as the father, mother, guardian, or curator (*curateur, tuteur*), to enable minors, insane persons, and other legally incapacitated persons to act. Although a similar category of "authority by law" is not unknown in common law, powers based on family relationships are scarce and appear in only a few cases.

THE DEVELOPMENT OF AGENCY DOCTRINES

The treatment of agency in Roman law. Since concepts grow primarily out of specific situations that have occurred and social needs that have arisen, the doctrine of legal representation developed differently in different times and places, sometimes even within a single legal system. At first it seemed unthinkable that one person, the agent, by making a contract with a third party, could create obligatory rights and duties between a third party and a principal. Even the official law of the Roman Empire never fully recognized the principle of representation. The explanation for this rejection lies mainly in the early Roman conception of a contractual obligation as a personal relationship binding the personalities of the parties in some quasi-mystical way. This type of relationship allowed creditors in some instances to seize the possessions—and in very early times also the person—of the debtor. Usually the formation of such a relationship between two parties took place in a solemn ceremony at which both parties had to be present, certain formal words spoken, and definite acts performed. It appears that in such a situation it was impossible to confer right or duties on a third party. On the other hand, the head of the household could transact business through his slaves or his dependent sons, who were not conceived of as agents but as "long arm" extensions of the contracting master or father. Because of the wide prevalence of slavery, there was no great need for a true agency relationship. As Roman law later developed, the formalities connected with forming legal relationships became less important, and the need for personal representation in commerce increased. In the meantime, however, legal theory and practice had developed so many ways to evade the problem that there was no longer an urgent need for Roman law to overcome its stark conservatism and to develop a legal institution that it had earlier opposed.

Conception of contract

Medieval development under the influence of the canon law and the Germanic law. Labouring under the influence of Roman law, legal development in the Middle Ages strove to overcome disadvantages in daily commercial life caused by the Roman rejection of the principle of agency. Through the efforts of legal scholars (glossators and commentators), Roman law was further developed by means of extensions, emphases, and exceptions—a process already sanctioned by the Romans themselves. Additional impetus for change came from the canon law of the Roman Church. Although manifestly structured after Roman civil law, canon law had its own special development, influenced by Hebraic theological concepts. Certain writers succeeded as early as the year 1200 in constructing a type of agency relationship based on the position of *procurator*, a relationship intended to solve the representation problem in all except legal matters. The issue nevertheless remained in dispute.

About this time, the doctrine of principal and agent developed in England as an outgrowth or expansion of the doctrine of master and servant. Anglo-Norman law created the figures of *ballivus* and *attornatus*. His position in the household of his master empowered the *ballivus* to transact commercial business for his master, reminiscent of the power of the slave to bind his master under Roman law. Later the *ballivus* was given more authority, especially in his frequent role as land administrator, gradually becoming competent to act independently for his master. On the other hand, the *attornatus*, originally just a representative of one of the parties in litigation, soon assumed a position of broader importance. Certain contracts were effective only when made in a judicially prescribed manner. For this reason, the formation of this type of contract always had to be concluded in a court proceeding in which an *attornatus* represented each party. This was the beginning of the role of the *attornatus* as a general agent.

The natural law development: the contribution of Grotius. Recognition of the principle of agency in the field of civil law was finally achieved in continental Europe during the ascendancy of natural law in the 17th century. By this time, however, new objections grounded in state law, feudal law, and the question of the general reasonableness of agency had to be overcome. Hugo Grotius in his best-known work, *De jure belli ac pacis*, explained that on the basis of his mandate a *procurator* could acquire rights directly for his principal. He thereby overcame the Roman rule that allowed slaves and dependent sons, but not free persons, to act directly for the head of the household. Grotius simply maintained that this rule did not contradict natural law. In a later work, *Defensio Fidei Catholicae*, Grotius added, in a theological context, that the principle of agency is based not on essential natural law but on nonessential natural law; that is, agency is not demanded by the nature of things but must only correspond to and be adapted to the nature of things.

18th-century developments. The codifications of the age of rationalism recognized the principle of representa-

tion. The Prussian *Allgemeines Landrecht* (1794), the French *Code Civil* (1804), and the Austrian *Allgemeines Bürgerliches Gesetzbuch* (1811) nevertheless regarded agency as an aspect of "mandate" and the power to act as an agent to be derived solely from that concept.

The master–servant theory

A similar dualism governed the English development of the doctrine of principal and agent, in large part because of the new doctrine's dependence upon that of master and servant. The jurists of the 18th century, employing the phrase *qui facit per alium, est perinde ac si facit per se ipsum* ("whoever acts through another acts as if he were doing it himself"), made much use of the fiction of identity. In contrast, 18th-century commercial law, which at the end of the Anglo-Saxon period had split off from the main body of common law and had been allowed to develop under less stringent controls, modified numerous feudal common law concepts. In particular, it brought a legal flexibility allowing the law of agency, which had grown from three independent common law roots—one relating to actions of debt and assumpsit (informal contract), another connected with the action of account, and the third deriving from deeds—to adapt itself to the peculiar problems and requirements of an individual case. Through this commercial influence, the concept of "undisclosed agency," an especially important mode of representation in commercial transactions in which the agent appears to be the principal, developed alongside open agency representation in English law. Thus, even when a principal with the real economic interest in the transaction remained concealed, claims for and against him began to be recognized, whereas in continental law the necessity for the agent to act openly in the name of the principal retained a more fundamental importance in the general doctrine of contracts. English law attempted to relate agency rules more closely to the everyday needs of the principal–agent relationship, in contrast to the conceptually systematic and more narrowly confining treatment of agency in continental European law. As Alfred Conard, a modern writer, has observed, agency came to be a legal institution in English and North American law, "knitting together the whole subject of the employment of one man by another, whether for one job or for life, and whether on commission or on payroll."

Nineteenth-century features: the effects of Laband's doctrine in the civil law countries. While the first practical treatises on agency in England, by William Paley, and in the United States, by Samuel Livermore and Joseph Story, appeared shortly after the beginning of the 19th century, in continental Europe the doctrines of the theoretical jurists continued to play their traditional leading role in the development of this branch of the law. A particularly important distinction in the European law of agency was made in the second half of the 19th century by the legal scholars Rudolf von Jhering and Paul Laband. Before them, agency was viewed solely in terms of the relationship binding the principal, the person being represented, and the agent, the person representing; that is, agency was equated with the relationship created by the mandate given to the agent. Jhering and, especially, Laband were the first to distinguish sharply between the agent's power to create legal rights and obligations for his principal and the inner contractual relationship governing the personal rights and duties as between principal and agent. This distinction was largely adopted by the continental European legal systems, and was codified in several countries.

Nineteenth-century legal theory also strengthened the so-called disclosure principle (*Offenheitsprinzip*), according to which the acts of an agent have direct legal implications for the principal only when the agent makes it known to the third party by acting in the name of his principal that he acts for him and not for himself. Only in such a case can there be direct representation—where the principal alone is the party to the contract—in contrast to indirect representation. Thus, this distinction became, to varying degrees, a common element in the individual civil law systems.

Anglo-American law, more realistic in this area and standing on different dogmatic grounds, was not influenced to the same degree by such principles, although it was also long plagued by refined doctrinal distinctions, such as the contrast between "general agents," with whom an outsider could deal with moderate safety, and "special agents," whose powers the courts viewed very narrowly. At this stage of development the remaining feudal elements in the English law of agency, whose prototypes had remained the servant and the steward of a lord, were largely overcome. Because of the increasing complexity of most modern commercial transactions, the significance of the various types of representatives grew to the point where agency had become one of the most important of modern legal relationships. The agent had come to be seen as an instrument of the principal without personal responsibility, a view which developed still further the responsibility of the principal for the misconduct of his agent.

Twentieth-century tendencies. Reflecting the generally more realistic attitude of contemporary private law, the modern treatment of agency has retreated from the somewhat artificial conceptualism of the 19th century and has emphasized the actual commercial and social context. Tendencies both to apply general solutions to broad types of cases and to provide specific solutions in the more unique individual cases have produced further refinements in the law of agency just as in the law of contracts. The continuing shift to more complex business units, accompanied by developments in marketing practices, has required solutions more appropriate to modern business necessities. In addition to a reassessment of the relationship between mandate and authority to act for the principal, new questions are being asked concerning the proper role of this kind of authority in relation to the legal and social functions of representation that it serves. Accordingly, the mere fact of one's external position, which without authority cannot properly be used, has been increasingly acknowledged as an independent ground for recognizing an agent's power of representation.

Authority and action

At the same time that modern society's concern for third persons' rights has led to recognition of powers of position, thereby making an agent's power of representation less dependent upon the will of his principal (the latter being bound in these cases whether he wishes to be or not), a better view of private autonomy, as expressed in the act of authorization, sees a fundamental connection between the agent's act and the will of his principal in the legal transaction as a whole. In this view the authorization by the principal and the act of the agent are not two independently complete legal transactions carried out in isolation from each other. Rather, they are two parts of an extended legal event, which would consist of a single act if no agency were involved. Thus, when the principal uses the help of an agent, the offer, which is usually made by a single act, becomes divisible into two parts (the authorization and the acting of the agent). Without such a connection, agency in private law, as it is understood today, would be largely incompatible with the principal's self-autonomy.

DIFFERENCES BETWEEN CONTINENTAL AND ANGLO-AMERICAN TREATMENT OF AGENCY

In order to compare agency in continental and Anglo-American law, the principal types of agency that have developed in practice should be noted.

The continental concept of the "commercial agent" and his functions. The four main types of commercial agent in continental Europe can be identified as follows:

Commission agent (*Germany and Switzerland: "Kommissionär"; France: "commissionaire"; Italy: "commissionario"*). The commission agent accepts or sells goods for the account of his principal, but in his own name. He is independent of his principal, has a claim for his commission, and, except in France, has the right when dealing with certain goods to conduct the transaction as he sees fit. The forwarding agent (Germany and Switzerland: *Spediteur;* France: *commissionaire de transport;*

Italy: *spedizioniere*) ships goods in his own name for the account of his principal and therefore is a type of commission agent.

Commercial agent (*Germany: "Handelsvertreter"; France: "agent commercial," and in part "voyageur," "représentant et placier"; Italy: "agente"; Switzerland: "Agent"*). The commercial agent negotiates and concludes contracts on behalf of his principal. Although the degree of his independence from the principal varies, he is never totally independent. While Italian law limits such an agent's activities to a specific geographical region, German law has such limitations only for the *Agent*, a subclassification of the general commercial agent who remains unrestricted in this regard. Apart from several protective rules, the French commercial agent is subject to the general rules governing the mandate.

Broker (*Germany and Switzerland: "Mäkler"; France: "courtier"; Italy: "mediatore"*). The broker is a business agent who is completely independent of his principal. In the area of employment brokerage or placement services, most European countries have passed special regulatory legislation to protect the interests of those persons using such services to seek employment.

Salesman. This type of agent is a dependent employee of a merchant who concludes contracts for the merchant outside the business establishment. Most European legal systems have no special provisions governing such an agent, but rather treat the position under the general rules governing dependent commercial employees (*e.g.*, Germany: *Handlungsgehilfen*).

The variety of forms of Anglo-American agency. Various kinds of agency relationships are evident in Anglo-American commercial life. The factor and the broker are the most common mercantile agents dealing in transactions involving personal property. The factor is entrusted with possession of the chattels to be sold, or the documents of title thereto, and is empowered to conclude the sale at the best price obtainable. The broker, on the other hand, has no possession of the object of sale but is empowered to make contracts for the purchase or sale of personal property on behalf of his principal. More limited are the powers of the real estate agent, who may show the land and state the asking price to the potential buyer without ordinarily being empowered to make further representations. The store salesman is similarly restricted in his power to represent his principal and can usually do no more than make customary warranties and sell at the price fixed by his employer. In contrast, the itinerant salesman not in possession of the goods normally has authority only to take orders, which are in effect mere offers to buy and are not binding on either party until the principal has accepted. The auctioneer is usually empowered to do no more than sell at the highest price bid.

Managerial and administrative agents

The common function of a second large class of agents is managerial or administrative. The manager of a business has the widest authority of all business agents and normally has complete control of all normal operations of the business. The agent employed to manage investments has a duty to deal only as would a prudent investor with reference to the principal's personal financial situation. The powers of the agent entrusted with his principal's land or chattels is more administrative; he has full power to protect the property and to make it profitable, but may not sell or encumber it. Finally, the attorney at law may do what is necessary to advance his client's interests and may incur the necessary expenses in so doing, but, unless specifically authorized, may not release any substantive rights of his client.

"Apparent authority" and related questions. If the principle of private autonomy were uncompromisingly applied to the law of agency, only an actually authorized agent could create legal rights and obligations for his principal. The will of the principal would then precisely define the boundaries of the agent's competence; however, a third party cannot always make a reliable determination of whether the agent has acted within the scope of this authority. Rather, he must often rely upon the principal's manifestation of the agent's authority, which may go beyond the more restricted authority actually communicated by the principal to the agent. Often he must rely upon the fact that the agent holds a certain position, such as wife, partner, employee, etc., in the belief that the normal incidents of authority implied by such a position are present, even though special restrictions on the agent's authority may have been made. And sometimes a third party must rely upon an earlier declared authorization that has in the meantime been revoked internally by the principal. Obviously, a third party cannot be expected to check all the details of the agent's internal authorization, especially since representation makes sense only when it functions efficiently and since use of an agent is also for the benefit of the principal. Therefore, the risk that the agent has acted without power of representation must be apportioned between the principal and the third party, with individual factors determining who must bear it in a particular case. This compromise is the subject of the doctrine of "apparent authority."

It is obvious that the expression "apparent authority" is inappropriate since in none of the above cases is the competence of the agent "illusory." With respect to the third party, at least, the authority of the agent is no less "real" than it would be if it were in conformity with the will of the principal. The distinction between apparent and real lies rather in the justification of the agent's acts in relation to the principal. In contrast to "apparent authority," "real authority" is more than mere power of an agent, for along with the agent's legal power "looking out" (*posse*) stands his privilege "looking in" to the lawfulness of his conduct (*licere*). In the case of a "real authority," power and privilege to exert the power are coextensive as against the principal, while in the case of "apparent authority," the agent has only the external legal power to perform without the corresponding internal justification vis-à-vis the principal.

Although the expression "apparent authority" spread from continental Europe, where it was used, among others, by the influential French classicist Robert Pothier (1699–1772), into English law, where Lord Ellenborough picked it up in *Pickering* v. *Busk* (1812), its development in the two legal families was independent and went systematically different ways. Nevertheless, the results obtained in concrete situations by the courts of both families often seem rather similar. A similar disparity marked the development of the doctrine within Europe itself, with the formulation of different rules in the French system, which largely identifies authority and mandate, and in the German system, which distinguishes the two ideas. Another general tendency that has arisen in both legal families has been to objectify the more subjective aspects of the doctrine, thus making the legal consequences more certain and predictable. Today both Anglo-American and continental courts permit the legal consequences of an agency relationship to result even when the principal is not willfully responsible for the appearance of authority.

Similar situations can also lead to opposite approaches in some matters. Here also it depends in the first instance upon whether the principal or the third party must bear the risk, since in general an agent acting without authority is not personally responsible. The question arises as to whether and to what degree the authority of the agent to bind his principal is affected by the death of the principal. According to the traditional English view (*Campanari* v. *Woodburn;* 1854), this event automatically terminates the agent's powers, irrespective of whether the agent or the third party knew or should have known of the death. While the rule originally might have rested upon the "fiction of identity," today it is supported more by the technical argument that what a dead man cannot himself do, he cannot do through another, with further support garnered from the so-called doctrine of the meeting of the minds that is necessary for concluding a contract.

The death of the principal

In contrast, the continental legal systems have evolved less doctrinaire solutions to this question based more on considerations of protection of the concerned parties.

Article 2008 of the French *Code Civil* even goes as far as to treat all transactions of an agent who acts in ignorance of the death of his principal as valid. The more balanced solution offered by the courts on the continent, however, is to make the good faith of the third party the determinative factor, since in the usual case the real interest of the agent is simply to avoid any personal obligation.

Disclosed and undisclosed agency. Continental European laws restrict the application of agency rules to cases where the agent acts openly in another's name. Thus, French jurists infer from article 1984 of their *Code Civil*, according to which agency is the act of the agent *pour le mandant et en son nom* ("for and on behalf of the principal"), the negative conclusion that in case an agent does not disclose that he is acting as an agent for a principal, the consequences touch only the "agent" himself. The hidden "principal" is not concerned by the effects of the transaction at all. Section 164 of the German Civil Code expressly provides that "an agent, who acts without disclosing the fact that he is acting as agent, is the only one to acquire any rights and is exclusively personally liable."

In contrast to the continental view, when an agent contracts in his own name without disclosing his principal, the common law allows the "undisclosed principal" under certain conditions to sue or be sued by the third party. Such conditions include that the agent had power to make the contract and that the parties eventually learn their respective identities. This wider concept of agency has no counterpart in continental legal tradition.

The use of this basic doctrine in the common law countries gives rise to questions regarding the identity of the undisclosed principal, the election of remedies that must be made by the third party, the extent of the respective liabilities, the right of the third party to set-off (the amount of its own damages from any sum that might be awarded it), etc. A solution to these conflicts of interests must in final analysis rest upon an evaluation of the extent to which the relationship between the undisclosed principal and the agent should influence the contract made by the agent with a third party.

A COMPARATIVE SURVEY OF THE SUBSTANTIVE LAW OF AGENCY

The external, unilateral act of "authorization." It is still a characteristic feature of the French Civil Code and of other codifications following its model (Spain, Portugal, Romania, Brazil, and other Latin American countries) that agency is not recognized as an isolated institution. This legal family conceives of agency only as a subordinate instance or "external effect" of mandate. The result is that it considers the power to act as an agent as a mere part of mandate and does not have a general concept of authorization as a distinct legal institution. In contrast to this approach, the more modern codifications of Scandinavia and of such countries as Germany, Switzerland, Japan, Poland, Czechoslovakia, Italy, Greece, etc., draw a sharp distinction between the unilateral organizational act on the part of the principal authorizing the agent to act and the internal contractual relations between the principal and the agent. This distinction, one of the major achievements of 19th-century European legal scholars, is also followed by modern English and American legal writers, even though the classical concept of mandate is unknown in the common law.

The insight that authority can exist independent of the underlying contract, and even without it, opens several new practical possibilities. For example, it explains the rule that the authorization of a minor can be valid if he is in fact mentally and physically capable of transacting business, even though the minor is not competent himself to conclude the (internal) contract of employment for himself that establishes fiduciary duties. Thus, the minor is not subject to the liabilities *ex contractu* (from or out of a contract) of an adult agent although he is authorized.

A few countries still prescribe a special form for every authorization. According to the Russian (RSFSR) and Brazilian codes, for example, an authorization must be given in written form. This general rule, which has been mitigated in the meantime to some extent by the courts, is influenced by a distrust of informality and by a concern "to protect the individual and to be better able to examine the lawfulness of a legal transaction." The opposite tendency is followed to the extreme by the German Civil Code. It not only adopts the general principle of "freedom of form"—with the rare exception that a special form is required in those cases where a statutory article expressly prescribes it—but even states that the granting of authority need not be in the form prescribed for the judicial act to which the agency power relates. In spite of this express statutory language, German courts do require compliance with formalities in certain extreme situations under pressure of practical necessity, in order to avoid frustration of the aim of the formal requirement in the principal contract. This attitude thereby approaches the English rule according to which, in most cases, no particular formalities are required, even if the agent is to make an agreement for the sale or lease of land, an agreement which has to be in writing. More attention to the connection between the authorization and the act of the agent appears in provisions like that of the new Greek Code, which says that the authorization must take the form required for the legal transaction for which the authority is issued unless the particular circumstances lead to a different solution. Other modern codes like those of Poland and Italy do not even contain the "unless" clause and prescribe only that an agent's power of attorney has to be granted with the formalities prescribed for the contract that is to be entered into by the agent.

The "internal" agency contract. The internal bilateral relationship between principal and agent—apart from agency of necessity—rests upon what is commonly termed a "contract of agency." This term encompasses all sorts of contracts of employment including gratuitous services (in those cases in which the agent is willing to render a friendly favour).

This internal relationship between principal and agent must begin with an agreement, since otherwise the principal would have no rights against the agent—even the right to beneficial service presupposes the agent's consent to serve. The consent of the principal may be given expressly by a written "power of attorney" or implied by his conduct, such as an established course of dealing. Generally, the formation of the contract of agency requires no formal ritual. Moreover, in many cases the obligation to act as agent is only a part of the greater complex of duties that forms a person's job, profession, or office (for example, those of factors, bailiffs, attorneys, brokers, etc.).

The basic principle of agency is that the agent, in fulfilling his obligation, concludes legal transactions on the part of his principal. When these transactions result from the authorized juristic acts of the agent, the result is that only the principal is bound by them. The competent agent himself remains outside the contractual relations with the third party (but compare "undisclosed agency" above) provided he has not also contracted for himself at the same time—for example, when the third party is not content with the principal's credit and insists upon having the agent's liability in addition to the principal's; in such a case the agent binds both his principal and himself. This general rule that the agent does not become involved may change in the exceptional case of an authorization conferred for the benefit of the agent. Thus, such an agent has been held personally liable by German courts for mistakes made in the course of contract negotiations (*culpa in contrahendo*).

Rights and duties between principal and agent. Continental European codifications generally do not treat the contract of agency as a special, separate type of contract. The standards governing the principal–agent relationship must therefore be derived from the general legal rules governing the mandate, the contracts for perform-

ing work, employment contracts, and partnership contracts, together with the more specific rules (*e.g.*, of the commercial codes) regulating particular agency relations and the individual transaction at hand.

Differing from this general approach, the common law has recognized a number of more specific rights and duties between principal and agent. Regularly the principal must provide the agent with a regular opportunity for service under the contract of employment and has a positive duty to aid, and not inhibit, the agent's performance of such service. He must compensate the agent for his services, provided they are not gratuitously rendered, and must account for those amounts due to the agent, including indemnification for economic loss suffered by the agent on account of the agency relationship (the agent's right of remuneration, right of indemnity, and right of lien). The principal also has the more abstract duty to conduct himself so as not to harm the agent's reputation.

Remuneration, indemnity, and lien: the agent's rights

The primary duties of the agent to the principal are those of care, obedience, and loyalty—similar to those of a trustee. More specifically, the agent must act solely for the interests of his principal and therefore must account for any financial benefit ("secret profit") he might derive from a transaction. It follows that it is a violation of an agent's fiduciary duty to his principal to have unrevealed interests adverse to those of his principal. He must not compete with his principal and may not use or disclose confidential information except for his principal's benefit. If an agent has received money or other property from or for his principal, he must account for it. An agent also may not normally delegate his task to a subagent, since the principal is assumed to have placed his confidence in the person of the agent and not in a subagent. Finally, the agent also has the abstract duty of conducting himself so as not to bring disrepute upon the principal.

Liability of principal for acts of agents. Continental-European law classifies the undertaking of transactions in the place of another as "agency" only when they are legal. It excludes other acts, including unlawful acts, so that when dealing with the law of agency, the rules concerning the liability of a master for the torts of his servant do not come into consideration.

Following the lead of Justice Oliver Wendell Holmes, the opposite view has been taken in English and American literature. The predominant opinion treats the liability of a master for the tortious conduct of his servant as a part of agency law since these cases logically come within the maxim *qui facit per alium, facit per se* ("he who acts through another, acts himself"). The doctrine of *respondeat superior* ("that the master must answer") is therefore treated as a part of agency law, even though the rationale behind the master's liability is that he has assigned to a servant under his control a task that unforeseeably results in damage to a third person. On the other hand, consequences flowing from an agency relationship rest on the idea that an authorized person performs legal acts within his competence not on his own behalf, but for the principal. A growing view in the modern literature recognizes, however, that "the two basic relations are strange bedfellows. . . ."

According to English and American law, the liability of a principal for his agent's torts in the ordinary course of his employment depends upon the existence of a master–servant relationship. If the principal and agent stand in a master–servant relationship to each other, the master is vicariously liable for his servant's tortious conduct committed within the course of his employment irrespective of any personal fault on his own part. This doctrine of *respondeat superior* arose from the belief that since the head of the household or economic enterprise exercises control, he should pay for the harm caused by its members.

Such a relationship presupposes a degree of physical control by the master over his servant, which may exist, for example, over the driver of the principal's delivery truck, but is lacking over the principal's factor or broker. The other major limitation to the doctrine is that the servant's tortious conduct must be committed within the servant's "course of employment." Generally, this includes conduct that is not a serious departure from that authorized, both in manner and space, by the master and that is actuated at least in part by a motive to serve the master. Under this principle, the master's vicarious liability extends in some instances even to claims arising from the servant's intentional torts.

If no master–servant relationship exists between the parties, the principal's liability for the torts of his agent is far more restricted. For example, a principal is not liable for the harm caused by the negligent physical conduct of his nonservant agent. He is liable, however, for any other torts of his agent that either are authorized or the commission of which "normally" are made possible by the agent's position. The chief example of such conduct is tortious misrepresentation by the agent.

Termination of agency. With the termination of actual authority, questions often arise as to whether the authority was usual, was special or general, arose out of necessity or was also apparent. Questions concerning the external and internal agency relationships create special problems, particularly in this area. A principal's termination of the agency relationship does not necessarily have the same effect vis-à-vis the agent as it has with respect to the third party. It may be that the principal can terminate the "real" authority but not the "apparent" authority toward a third party.

In theory, the principal may revoke the real authority of his agent or the agent may terminate his authority by renunciation or revocation at any time. If the revoking party has previously agreed not to do so, the revocation remains valid, but the other party may maintain a breach of contract action against him. The agency relation can also terminate by operation of law in case of death, insanity, bankruptcy, or war, particularly where the principal is concerned. The question then arises as to whether the unknowing agent is protected by the law, since he might be personally liable on the contract.

A major exception to the principal's power of revocation occurs when the agent possesses "authority coupled with an interest." Here the agent who has part ownership in something that is to be disposed of has power also to dispose of his principal's remaining interest in this thing. Such a power for the purpose of effectuating any security or of protecting or securing any interest of the agent cannot be revoked without the agent's consent.

BIBLIOGRAPHY. OLIVER WENDELL HOLMES, "Agency," in *Collected Legal Papers*, pp. 49–116 (1920), is an Anglo-American treatment of the subject, with an historical point of view, written by an outstanding American jurist. W.A. SEAVEY, *Studies in Agency* (1949); M. FERSON, *Principles of Agency* (1954); and S.J. STOLJAR, *The Law of Agency* (1961), are more systematic Anglo-American works on the basic principles of the law of agency. A. CONRAD, "What's Wrong with Agency?" *J. Legal Educ.*, 1:540–562 (1948–49), adds a critical note to the study. W. MULLER-FREIENFELS, *Die Vertretung beim Rechtsgeschäft* (1955); "Law of Agency," *Am. J. Comp. Law*, 6:165–188 (1957); "Legal Relations in the Law of Agency: Power of Agency and Commercial Certainty," *ibid.*, 13:193–215, 341–359 (1964); and "The Law of Agency," in *Civil Law in the Modern World*, ed. by A.N. YIANNOPOULOS, (1965), offer comparisons between Anglo-American and civil law rules. See also A.N. YIANNOPOULOS, "Brokerage, Mandate, and Agency in Louisiana: Civilian Tradition and Modern Practice," *La. Law Rev.*, 19:777–812 (1959); and K.G.W. GRONFORS, *Ställningsfullmakt och Bulvanskap* (1961), in this connection for special questions.

(W.M.–F.)

Aggressive Behaviour

The term aggressive behaviour is used in so many different ways that no single definition can possibly cover all of its meanings. Behaviour that serves to injure an opponent or a prey animal, or to cause an opponent to retreat, is usually considered aggressive. When considering human aggression, some psychiatrists consider any act that has destructive consequences (including suicide) to be aggressive. (For a discussion of aggressive behaviour

in man, see EMOTION.) The role of aggression in behaviour has been—and continues to be—debated by psychologists and ethologists, as does the meaning of the term itself. Frequently, aggressive behaviour encompasses both attack and defense. Other investigators exclude food-gathering behaviour, though it may involve attack on another animal. In order to avoid these ambiguities of definition, a distinction must be made between causation, function, and description of observed behaviour. It is frequently assumed that a single motivational system (aggression) causes all recognizably aggressive behaviour in higher animals. This assumption is certainly invalid for invertebrates and for most higher vertebrates, in which a variety of motivational bases appear to exist. A motivational definition of aggression is thus difficult. The only possible rigorous approach is to list patterns of behaviour, usually held on both functional and causative grounds to be aggressive. Because aggressive behaviour has been most studied in mammals, this article will examine mammalian behaviour first as a basis for comparison with other animals.

GENERAL FEATURES

Most modern theories of aggressive behaviour are based on laboratory experimentation, especially with mice and cats, and on field observations. Mammals show two basic aggressive patterns: attack and defensive threat. Direct evidence from brain-stimulation experiments in marsupials and eutherian (placental) mammals shows that each pattern is a closely coordinated series of events, obtainable in relatively constant form from a wide range of brain sites. Attack behaviour results from stimulating the lateral and anterior regions of the hypothalamus, and defensive threat from some areas near the centre of the hypothalamus and from the central gray region of the mesencephalon (midbrain).

Mammalian attack behaviour

In the opossum, attack involves visual fixation on the prey, seizing it with the jaws, and shaking it from side to side through a figure eight. In defensive threat, the jaws are opened and the animal backs into a corner, hissing. Newborn mice exhibit escape behaviour. Helpless at birth, a mouse will respond defensively to an outside stimulus (*i.e.*, pinching the tail) by squeaking, moving its legs, and moving away. When the mouse is older, its behaviour becomes more offensive in that it will attempt to bite. Still later, at about 12 days of age, it will stand on its hind legs in a posture of defense. Not until it is a month old, however, will it attack. The pattern of fighting in male mice includes kicking, biting, and rolling over until one of the combatants is hurt. The injured mouse might attempt escape, with the attacker in pursuit. If escape is not attempted or is not possible, the injured mouse may show defensive behaviour or feign death, submitting to the attacks of the aggressor. The aggressive patterns in the cat have been studied extensively. There is disagreement about the relation in this species between movements of prey catching and those of attack on other cats, and indeed whether prey catching should be termed aggressive at all. Since the coordination of stalking, biting, seizing with the forepaws, and scratching with the hind legs can be the same in both situations—and the eliciting stimuli also are often similar—it is reasonable to classify them together. The dispute stems from the possibility of the same pattern having two or more rather different causations.

In the cat, it is clear that defensive threat involves a number of component patterns that can occur independently of each other. The patterns include reflexes that protect vulnerable areas, such as narrowing the eyes or flattening the ears and cardiovascular preparations for exertion (*e.g.*, a rise in arterial blood pressure and active dilation of the blood vessels that supply muscles). Hisses and growls, together with erection of back and tail hair, are also typically part of defensive threat, serving to make the animal conspicuous.

Similar patterns can be recognized in birds and can be induced by brain stimulation in roughly the same areas as in the cat. In the collared dove (*Streptopelia decaocto*) attack involves pecks and wing blows, whereas an immobile threat display includes bill opening and feather erection.

Human attack behaviour, defined in this way, might consist of a direct stare, exaggerated by a slight frown, as preparatory behaviour, and then, at full intensity, striking with hand or fist, seizing, and biting. Staccato, loud, or breathy speech, given at such times, may be compared to the violent expiration (hissing, growling) of other primates in threat.

Acquired skilled movements (*e.g.*, boxing, wrestling) could reasonably be included, just as fighting in adults of other species probably also involves skilled movements (*e.g.*, blows with leg spurs in cocks).

CAUSATION

Evocative stimuli. Few stimuli that are highly specific for attack have been described. A scent produced by adult male mice is probably one example. Another is colouring. It has been shown that chaffinches (*Fringilla coelebs*) have a personal territory about them that they protect from intrusion by other chaffinches. This territory may be related to a source of food or water or to space for perching. Invasion of the personal space of one chaffinch by another usually results in either attack or withdrawal. Males may penetrate up to 20 centimetres before either alternative occurs. Females, however, are permitted to penetrate up to 10 centimetres. Experiments have been performed in which the feathers of the female chaffinch's breast are dyed to resemble those of the male. When such a disguised female penetrates the personal territory of a male, attack or withdrawal occurs within 20 centimetres. Experiments with inexperienced birds indicate that the effect is innate.

Members of an animal's own species (conspecifics) that are unfamiliar tend to evoke attack, perhaps partly because they lack the cues that suppress attack in familiar fellows (*e.g.*, group smell in mice; see below *Aggression as a drive*). Another very widely effective stimulus is that of a conspecific approaching within a certain distance, which, in the case of species the members of which ordinarily do not tolerate close approach, may be very uniform; for example, the chaffinches' inviolate personal territory, which has already been mentioned above. This distance may be maintained around objects other than the animal itself. A typical songbird territory includes points, such as song perches or the nest site, that are defended in this way. Approach to the mate often evokes attack; this is particularly clear in species in which paired birds remain in flocks (*e.g.*, some grackles, *Quiscalus*) and approach to the mate is not the same as entry of the territory. Both birds and mammals commonly attack in response to an approach that threatens their offspring. In many ways (*e.g.*, grooming) offspring are treated as an extension of the maternal body, thus, a threat to the offspring is a threat to the parent.

Stimulus of a fleeing animal

A conspecific moving rapidly away tends to evoke chasing and attack in many birds and mammals. In such predators as dogs attacks like this may be partly inhibited, taking the form of aggressive play, but their resemblance to prey catching is so close that any distinction would be artificial. The stimulus is equally effective, however, in such species as baboons, which rarely or never take large prey. For young domestic fowl, inexperienced in attack, repeated rapid retreats and approaches by another individual are the optimal stimuli for sparring and pecking.

Submissive behaviour (*e.g.*, exposing throat or belly) has been described as providing stimuli that specifically inhibit attack. It is more likely that such behaviour is effective either by removing stimuli that tend to evoke attack, or by evoking systems of responses incompatible with attack. Examples of the first category are immobility, as in the resting attitudes assumed by many mammals and birds when persecuted, and the avoidance of behaviour (such as the direct stare) suggestive of intent to attack. The assumption of infantile behaviour (*e.g.*, food begging in many birds) makes attack less likely. Many social mammals, indeed, have evolved special features in their young (*e.g.*, juvenile fur colour of baboons and

guenon monkeys) that are so effective they allow their possessor to behave freely in ways that would certainly produce attack if shown by an adult. The assumption of sexual postures is another example of the second category of submissive behaviour. In baboons and macaque monkeys "presenting" (the type of solicitation used by a sexually receptive female) has become a common response of any subordinate, male or female, when approached by a superior. The Austrian ethologist Konrad Lorenz has argued that predators have unusually effective submissive behaviour because they are so well armed and, in the general spirit of fashionable extrapolation from animals to man, that it is therefore unfortunate that man is not more genuinely a carnivore. In fact, as with most displays, the clearest correlation is that the more social the species, the more elaborate and easily elicited the communicating displays. Man is even better equipped than other primates with infantile displays (weeping, dependent behaviour) and friendly greetings (smiles) that avert attack.

Dominance hierarchies

Dominance hierarchies, first described in terms of peck-right in the domestic fowl, may be regarded as a means of reducing attack within a social group. Once the hierarchy is established, disputes rarely need arise because the inferior immediately gives way when confronted by a superior.

A dominant animal is best defined as one whose actions are not constrained by possible responses of its fellows. Its position need not derive from earlier successful aggressive behaviour, although it usually does, but may instead reflect superior persistence or strength. Although best known in birds and mammals, dominance hierarchies have also been described in crabs and bumblebees. In bumblebees the dominant animal opens its mandibles and butts the inferior one.

Facilitating stimuli. Startling and painful stimuli facilitate attack, which may be directed to objects not associated with the source of pain or startle. The effect has been shown in both rats and squirrel monkeys, following electric shock administered through a floor grid. A clear distinction has usually not been made between attack, defensive threat, and a mixture of the two; but both patterns are clearly facilitated. The human irritability that often accompanies persistent discomfort or pain provides a more familiar example.

The thesis that frustration causes aggression has been very influential since it was first proposed. In a laboratory experiment, a pigeon immediately attacked another whenever a lever that normally delivered food when pressed failed to do so. There is no reason, however, to elaborate such evidence into the proposition that all of the attack is caused by frustration, as has been suggested by some investigators.

Reinforcing patterns. It has been suggested that a number of important patterns of behaviour—most or perhaps all of which have major representation in lateral hypothalamic brain centres—include component responses that produce self-reinforcing stimuli. An animal in which the pattern is activated, either naturally or by hypothalamic stimulation, will learn to run a maze in order to be able to perform the response. This observation has been made of the rat, not only for such obvious patterns as feeding and drinking but also for gnawing on wood and other objects. Following appropriate electrical stimulation, a cat will run a maze in order to be able to attack a rat, even though most cats normally do not respond to rats. This behaviour involves prey catching, rather than attack, if a distinction is to be drawn. There appears to be no direct evidence of this type for attack in animals not artificially stimulated.

Play as aggressive behaviour

It is clear, observationally, that once attack has become likely, an animal will search for the opponent and perform a variety of responses in order to reach him. The study of play behaviour suggests that the movements of prey catching (*e.g.*, kittens pouncing on a ball) or attack (*e.g.*, the play fighting so common in primates and other mammals) in themselves can be reinforcing. In such play, hurting or inducing signs of hurt are avoided, but it is possible that these sometimes are sought and serve as reinforcers, as is certainly suggested by human behaviour.

It is almost certain, however, that attack and defensive threat are also performed under the control of quite other reinforcers, such as removing a social fellow who is an inconvenience or an obstacle. It has been shown that a pigeon can be conditioned to attack a cage mate as an operant response to obtain a food reward from a hopper quite remote from the animal attacked. Full intensity attack was given, but other experimenters have found that attack caused in this way shows the simplification and reduction in intensity typical of operant responses. This reduction is obvious in the light peck used by dominant hens to make inferiors retreat from food or to control them in other ways. There is no reason why attack should be used for such a purpose rather than some other behaviour, if some other type of behaviour was more convenient.

Aggression as a drive. Lorenz is the most recent author to treat aggression as a drive that energizes attack (and some other behaviour), exhausted by performance energized of the behaviour, and then accumulating afresh, and making performance more and more likely. The resemblance to the death wish of Freudian theory is obvious, particularly since both suggest that quite unlikely behaviour may be motivated by the drive. This approach to the causation of behaviour can be traced back through the writings of Charles Darwin to an influential essay by Herbert Spencer on laughter and its role in discharging accumulated nervous energy. No convincing proof for such theories has ever been presented.

There is now direct evidence that threshold (the amount of stimulus required to elicit the response) for attack or threat may be raised or lowered over periods as long as 20 minutes. In the ring dove, following the evocation of either response by a natural stimulus or by brain stimulation, some brain sites that normally give the response show a changed electrical threshold for such a period. In many species, once a stimulus capable of evoking attack has been presented, the individual may remain likely to attack for some time, especially when the attack is evoked by an animal itself too dominant to be attacked. Other conspecifics or, in man at least, inanimate objects may then be attacked instead, a phenomenon called redirection. What evidence there is, however, suggests that such states of readiness, which in man would be termed irritability or anger, will dissipate just as well or better in the absence of actual performance of aggressive behaviour.

There is no reliable evidence that aggressive behaviour becomes progressively more likely over long periods in which it is not evoked, as has been suggested by some workers. Such an increased likelihood is clearly found for feeding behaviour, but the consequences of not eating —such as reduced availability of glucose and free fatty acids to tissues and a reduction of fat stores—provide internal stimuli that make feeding more likely. These effects, of necessity, become more intense the longer a fast is prolonged. In the case of performance of copulation, the act is followed by a reduced likelihood of response to further stimuli. In the male of certain species (domestic cattle, for example) this reduction is almost entirely restricted to copulation with the female just mounted, the bull remaining receptive to stimulation by other cows. In other species there is some contribution from a general rise in the threshold for copulation. There is no evidence for either type of effect in the case of aggressive behaviour, except that isolation promotes fighting. This is well established for male laboratory mice. Attacks on intruders, however, depend largely on odour cues (*e.g.*, male cues increase, female cues decrease, attack). Fighting within a social group of males is much reduced by group cues, and isolation may act to abolish the effectiveness of such cues.

Fighting in laboratory mice

The increased incidence of aggressive behaviour in small groups of isolated men, such as polar explorers or prisoners, has been explained by Lorenz as resulting from a build-up of undischarged aggression. But the comparison is not strictly relevant because the stress and

frustration of such situations might well facilitate aggressive behaviour directly.

Role of hormones in aggression

Physiological effects on aggressive behaviour. It has long been known that the administration of male sex hormones (androgens) makes aggressive behaviour more likely in many male mammals and birds. It has no effect on aggressive behaviour in female rats and mice, apparently because of the general insensitivity of the female central nervous system to androgens. Male individuals castrated at birth are less aggressive than those castrated at the time of weaning, which suggests that masculine aggressiveness is partly determined by early differentiation. Androgens also affect male aggressive behaviour indirectly. In the red deer, testosterone implants cause migration to the rutting (mating) area, which will be defended, and also cause the loss of velvet from the antlers, so that these become effective weapons. The increase in aggressive behaviour caused by the possession of a territory is particularly obvious in passerine birds, the males of which may continue to flock freely with other males on feeding grounds at a time when, if on their territories, they would attack other males fiercely. Hormonal effects on aggressive behaviour also occur in females. The increased aggressivity shown by female hamsters at pregnancy is the result of rising progesterone levels.

Lesions of the amygdala, a structure near the base of the forebrain of great importance in motivation, have been long known to produce taming in a wide variety of mammals and in ducks, and the effective area has been localized to the rostral lateral amygdala in cats. Defensive threat, attack, and fleeing are reduced as a result. The amygdala is not exclusively concerned with aggressiveness; it appears to be especially important in cases in which the novelty, intensity, or past significance of a stimulus determines whether the source is approached and eaten or avoided. The amygdala may facilitate aggressive behaviour as part of a determination that a stimulus is intense or unusual.

Lesions of the septum, a part of the base of the forebrain containing important tracts connecting to the rest of the brain, have the opposite effect: in the laboratory rat, such lesions cause defensive attack to be given to such stimuli as handling or a light (but presumably startling) touch, which were previously ineffective. The septum is also implicated in the inhibition of certain responses (*e.g.*, those that have ceased to be rewarded), and this may be related to its role in suppressing attack.

Threat displays. The main modern work on displays has explained their causation in terms of conflict between such major drives as aggression, fear, and sex. There is considerable evidence, especially from experimental brain stimulation in the cat, that defensive threat is a unitary pattern caused by the same stimuli that evoke attack or fleeing but independent of the latter. Confident threat displays are given by animals unconstrained by the possibility of retaliation, which do not attack because, for example, they do not wish to move. These displays include a wide variety of components. A direct gaze, which may be exaggerated by a frown in mammals with mobile faces, indicates a sustained interest. Protective reflexes indicate that the animal showing them anticipates close and possibly dangerous contact. They are assumed in rough, friendly play; but when they appear during approach—when still at a little distance—they suggest the expectation of attack (*e.g.*, ear flattening in the horse); the greater the distance at which they appear, the less confident the animal showing them. Finally, the animal may perform actual intention movements of attack, such as mouth opening, withdrawing the lips, or striking at the ground.

Many displays, termed threat displays because they are given to strangers at the territorial boundary or during early phases of pair formation when attack is still likely (*e.g.*, long call or choking display of gulls), would be best given a more neutral name.

EVOLUTION

It is instructive to attempt to identify attack in the familiar frogs (*Rana*) and toads (*Bufo*) of the amphibian order Anura. Prey catching in most frogs is highly specialized, involving a lunge of the head or a movement of the tongue only and directed only at small moving prey. In this respect, it is quite unlike attack shown by most land vertebrates. The mounting and clasp of females are so obviously related to copulation that they, too, are not readily confused with attack. Males clasped in error use wiping movements of hind legs that also are evoked by dirt on the body surface, and they struggle as they would if caught in some obstruction. There is thus no obvious attack response. Some of the functions of attack, however, are achieved by displays that are not derived from aggressive behaviour: females that have spawned, and all males, when clasped, give a special call that tends to produce unclasping. In higher tetrapods (land vertebrates), such unwanted copulation attempts would certainly evoke and be terminated by attack. In some anuran species, the calls given by males to attract females elicit avoidance by other males, ensuring optimum spacing of breeding males.

Relation of aggression to mating

Iguanid lizards resemble such frogs in possessing a female rejection display (legs extended, back arched, and tail turned toward the male), which appears to be specific in effect. Males give what appears to be an aggressive display to other males from lookout sites, but there is no evidence that it has evolved from movements of attack; it consists of vertical head bobbing that may well be an exaggeration of movements used in depth perception. The movements precede copulation as well as attack, which in lizards involves bites and sometimes tail blows, and the two may have evolved quite separately. The bites presumably derive originally from prey-catching behaviour, which is clearly as ancient in the vertebrates as the evolution of the jaw. As soon as dangerous bites became possible for feeding in vertebrate evolution, there was strong selection for their use in defense, and the facilitation of attack by pain or startle may be as ancient as this selection. The use of prey-catching behaviour to drive off conspecifics from areas offering important resources (*e.g.*, food, mate, nest) would also be strongly selected, for, once the pattern was available (see below), it would require initially only a slight change in evoking stimuli (*e.g.*, response to rather larger moving objects) to allow this result. The crossopterygian ancestors of the tetrapods and their immediate amphibian descendants were dominant predators, so that prey catching and biting in defense can be assumed to have been present. Comparison with present-day inhabitants of muddy shores, both air-breathing fish (*e.g.*, mud skipper) and crabs, suggests that conspicuous displays given to the approach of a conspecific had probably also already evolved. Causally then, as now, some or all of such displays might have been independent of attack. No doubt they already could be followed either by copulation attempts or by attack. There would be strong selection for appropriate hormones to confine the displays to breeding periods; this selection has happened with frog breeding calls and with bird songs, as well as with bird territorial displays, all of which are now facilitated by testosterone. The fact that this is true also of attack and male copulation has been taken mistakenly in the past to indicate causation of the former categories of bird behaviour by aggression and sex.

Prey-catching movements in displays

Aggressive behaviour of invertebrates. An account of invertebrate aggressive behaviour must be based on an analogy with the functions of that of vertebrates, since there can be no question of homology (common origin). Obviously, even in vertebrates, many of these functions are also served by other behaviour as well.

Aggressive behaviour may allow a vertebrate to maintain, exclusively if possible, its access to some important resource such as food, territory, or mate. In some invertebrates, there is the same use of attack derived from prey-catching responses, as in higher vertebrates. The larger dragonflies (Odonata) maintain a feeding territory by hawking at conspecifics in the same way as at prey. The clearest example is that of nest defense in social wasps and bees. Here a sting and a secretion, originally evolved to paralyze insect prey, have come to be used both

Nest defense by social insects

against conspecifics (*e.g.*, when fighting with "robber" bees attempting to enter the hive) and against vertebrates, which raid bee and wasp nests. Bees are not predatory, so their sting has no food-capturing function. The scent of bee venom strongly facilitates the stinging of large, hairy objects, and a whole hive may pass into a state entirely analogous with that of heightened aggressivity in a vertebrate but with the persistence of lowered threshold (*i.e.*, greater readiness to attack) dependent on the persistence of a chemical signal (pheromone). Both termites and ants have evolved facilitation of colony defense by pheromones released by disturbed individuals, like that described for bees.

In their attack behaviour, ants and bees show an interesting use of responses evolved to keep the nest clean of dead conspecifics and other debris. Drones (males) are seized and thrown out of the hive entrance. Strangers are treated similarly if they remain passive. If the stranger is identified at a distance (*e.g.*, by the side-to-side flight characteristic of robber bees), the guard bees may open their mandibles in a manner reminiscent of vertebrate threat, although there is no clear evidence that it is effective as a signal.

Freely ranging spiders, such as wolf spiders (Lycosidae), respond to conspecifics as they do to prey; the animal attacked commonly gives a low-intensity version of a display shown at full intensity in courtship, which tends to terminate the attack. Perhaps the least advanced group in which behaviour analogous with attack has been described is that of polychaete worms. Predatory ragworms (*Nereis*) defend their burrows against conspecifics with biting attacks. In other instances, biting is the analogue of vertebrate defensive threat, which has evolved from prey-catching behaviour; most crabs, when disturbed, extend and open their chelipeds (large claws), which evolved for feeding.

The function of defending resources against conspecifics is also served by a variety of adaptations that should not be termed aggressive (*e.g.*, Orthoptera stridulation).

Defensive displays. In addition to territorial defense, aggressive behaviour may function to terminate some behaviour of a fellow animal. The crying of human infants effectively manipulates mothers to remove sources of discomfort or fear; it is misleading to term it aggressive, however. In the chimpanzee, screaming tantrums are known to allow infants to win disagreements with the mother and access to food. Tantrums are also used by adults to some extent. In many primates (*e.g.*, baboons), such behaviour shades into the screeches of defensive threat, but it is probably best separated from aggressive behaviour proper. That the distinction is needed is clear for an analogous series of behaviour patterns, the distress calls of young gallinaceous birds (pheasants, quail, and domestic fowl), which are given when there is marked discrepancy between the stimulation expected or sought after and that obtained. Aggressive behaviour can thus be regarded as a source of only one of a range of reinforcers used by higher vertebrates to punish and manipulate their fellows.

Defense against predation can depend on many adaptations other than attack and defensive threat. Those that involve the infliction of damage, pain, or startle by a response aimed at the predator may be considered here as analogous to aggressive behaviour. A number of mustelids, especially skunks, have evolved a new pattern of defensive threat in which the readiness of the scent glands to discharge is conspicuously advertised, in one species by standing on forepaws to present glands. Various petrels (seabirds of the families Hydrobatidae and Procellariidae) discharge the oily and repellent contents of their crops over substantial distances at nest predators; the accuracy of aim may depend on the fact that they face the predator in order to examine him binocularly.

The loud call given by most tetrapods, including frogs and toads, when seized, apparently serves to startle the predator so that a final opportunity for escape may be given. Such a call may well have been involved in the evolution of the screeches given by many mammals in defensive threat. Many invertebrates have specialized startling displays that are comparable: the sudden display of eyespots by many moths when disturbed by a sudden movement or tap, the forewings typically moving forward to expose markings on the hind wing.

BIBLIOGRAPHY. SILVIO GARATTINI and E.B. SIGG (eds.), *Aggressive Behaviour* (1969), up-to-date theory and experiment; KONRAD LORENZ, *Das sogenannte Böse* (1963; Eng. trans., *On Aggression*, 1966), highly readable; R.A. HINDE, *Animal Behaviour*, 2nd ed. (1970), and S.A. BARNETT, *A Study in Behaviour* (1963), general books with particularly useful sections on aggression.

(R.J.A.)

Aging

Gerontology, the study of the aging process, developed from man's awareness of the inevitability of decline and death, an awareness that has long been a major influence in religions and philosophical systems and continues to shape man's intuitive conceptions about the nature and origin of aging. See the articles LIFE-SPAN and DEATH for related information.

GENERAL FEATURES

Gerontology is devoted to the understanding and control of all factors contributing to the finitude of individual life. It is not concerned exclusively with debility, which looms so large in human experience, but deals with a much wider range of phenomena. Every species has a life history in which the individual life-span has an appropriate relationship to the reproductive life-span and to the mechanism of reproduction and the course of development. How these relationships evolved is as germane to gerontology as it is to evolutionary biology. It is also important to distinguish between the purely physicochemical processes of aging and the accidental organismic processes of disease and injury that lead to death.

Gerontology, therefore, can be defined as the science of the finitude of life as expressed in the three aspects of longevity, aging, and death, examined in both evolutionary and individual (ontogenetic) perspective. Longevity is the span of life of an organism. Aging is the sequential or progressive change in an organism that leads to an increased risk of debility, disease, and death; senescence consists of these manifestations of the aging process.

Actuarial functions

The viability (survival ability) of a population is characterized in two actuarial functions: the survivorship curve (A in the Figure) and the age-specific death rate, or Gompertz function (B in the Figure). The relation of such factors as aging characteristics, constitutional vigour, physical factors, diet, and exposure to disease-causing organisms to the actuarial functions is complex; there is, nevertheless, no substitute for them as measures of the aging process and of the effect of environmental or genetic modifiers.

The age-specific mortality rate is the most informative actuarial function for investigations of the aging process. It was first pointed out by an English actuary, Benjamin Gompertz, in 1825 that the mortality rate increases in geometric progression—*i.e.*, by a constant ratio in successive equal age intervals. Hence, a straight line, known as the Gompertz function, results when death rates are plotted on a logarithmic (ratio) scale. The prevalence of many diseases and disabilities rises in the same geometrical manner as does the mortality rate, important exceptions being some infectious diseases and diseases arising from disturbances of the immunological system. Although the life tables of most species are remarkably similar in form, even closely related species can differ markedly in the relative incidence of the major causes of death.

NATURAL HISTORY OF AGING

Reproduction and aging. Reproduction is an all-important function of an organism's life history, and all other vital processes, including senescence and death, are shaped to serve it. The distinction between semelparous and iteroparous modes of reproduction is important for an understanding of biological aging. Semelparous organ-

isms reproduce by a single reproductive act. Annual and biennial plants are semelparous, as are many insects and a few vertebrates, notably salmon and eels. Iteroparous organisms, on the other hand, reproduce recurrently over a reproductive span that usually covers a major part of the total life-span.

In semelparous forms, reproduction takes place near the end of the life-span, after which there ensues a rapid senescence that quickly leads to the death of the organism. In plants the senescent phase is usually an integral part of the reproductive process and essential for its completion. The dispersal of seeds, for example, is accomplished by processes—including ripening and fall (abscission) of fruits and drying of seed pods—that are inseparable from the overall senescence process. Moreover, the onset of plant senescence is invariably initiated by the changing levels of hormones, which are under systemic or environmental control. If, for example, the hormone auxin is prevented, by experimental means, from influencing the plant, the plant lives longer than normal and undergoes an atypical prolonged pattern of senescent change.

The significance of aging patterns in insects

Useful inferences can be drawn from the study of the aging processes of insects that display two distinct kinds of adaptive coloration: the procryptic, in which the patterns and colours afford the insect concealment in its native habitat; and the aposematic, in which the vivid markings serve as a warning that the insect is poisonous or bad tasting. The two adaptation patterns have different optimal species survival strategies: the procryptics die out as quickly as possible after completing reproduction, thus reducing the opportunity for predators to learn how to detect them; the aposematics have longer post-reproductive survival, thus increasing their opportunity to condition predators. Both adaptations are found in the family of saturniid moths, and it has been shown that the duration of their post-reproductive survival is governed by an enzyme system that controls the fraction of time spent in flight: procryptics fly more, exhaust themselves, and die quickly; aposematics fly less, conserve their energies, and live longer.

These examples indicate that in semelparous forms, in which full vigour and function are required until virtually the end of life, senescence has an onset closely coupled with the completion of the reproductive process and is governed by relatively simple enzymatic mechanisms that can be modified by natural selection. Such specific, genetically controlled senescence processes are instances of programmed life termination.

The iteroparous forms include most vertebrates, most of the longer-lived insects, crustaceans and spiders, cephalopod and gastropod mollusks, and perennial plants. In contrast to semelparous forms, iteroparous organisms need not survive to the end of their reproductive phase in order to reproduce successfully, and the average fraction of the reproductive span survived varies widely between groups: small rodents and birds in the wild survive on the average only 10 percent to 20 percent of their potential reproductive lifetimes; whales, elephants, apes, and other large mammals in the wild, on the other hand, live through 50 percent or more of their reproductive spans, and a few survive beyond reproductive age. In iteroparous forms the onset of senescence is gradual, with no evidence of specific systemic or environmental initiating mechanisms; senescence manifests itself early as a decline in reproductive performance. In species that grow to a fixed body size, decline of reproductive capacity begins quite early and accelerates with increasing age. In large egg-laying reptiles, which attain sexual maturity while relatively small in size and continue to grow during a long reproductive span, the number of eggs laid per year increases with age and body size but eventually levels off and declines. The reproductive span in such cases is shorter than the life-span.

These comparisons illustrate the influence exerted by factors of population dynamics on the evolution of reproductive and bodily (somatic) senescence. The proportional contribution of an individual to the rate of increase of the iteroparous population obviously diminishes as the number of his living progeny increases. In addition, his reproductive capacity diminishes with age. These facts imply that there is an optimum number of litters per lifetime. Whether or not these influences of population dynamics lead to the evolution of adaptive senescence patterns has long been debated by gerontologists but has not yet been investigated definitively (see Table).

Life Spans of Selected Mammals, Birds, Reptiles, and Invertebrates

	scientific name	life span (years)
Carnivores		
Coyote	*Canis latrans*	16
Brown bear	*Ursus arctos*	37
Bengal tiger	*Panthera tigris*	26
Bobcat	*Felis rufus*	25
Ungulates		
Asiatic elephant	*Elephas maximus*	70
Syrian wild ass	*Equus hemionus*	36
Hippopotamus	*Hippopotamus amphibius*	51
Duiker	*Cephalophus nigrifrons*	20
Birds		
Carrion crow	*Corvus corone*	17
Pheasant	*Phasianus colchicus*	11
Parrot	*Amazona* sp.	39
Ostrich	*Struthio molybdophanus*	25
House sparrow	*Passer domesticus*	14
Reptiles		
Gecko	*Gekko gecko*	4
Snapping turtle	*Macrochelys temmincki*	59
Galapagos tortoise	*Testudo gigantea*	64+
Alligator	*Alligator mississippiensis*	56+
Gila monster	*Heloderma suspectum*	25
Water moccasin	*Ankistrodon piscivorus*	19
Invertebrates		
Lobster	*Homarus gammarus*	33
Honeybee queen	*Apis mellifera*	5
Fruit fly	*Drosophila melanogaster*	0.25
American cockroach	*Periplaneta americana*	4.5
Primates		
Man (neolithic)	*Homo sapiens*	90
Orang-utan	*Pongo pygmaeus*	50
Capuchin	*Cebus capucinus*	40
Brown lemur	*Lemur fulvus*	31
Montremes and Marsupials		
Echidna*	*Tachyglossus aculeatus*	50
Virginia opossum	*Didelphis marsupialis*	5
Wallaroo	*Macropus robustus*	20
Bats		
Indian fruit bat	*Pteropus giganteus*	17
Vampire bat	*Desmodus rotundus*	13
Little brown bat*	*Myotis lucifugus*	24
Rodents		
Gray squirrel	*Sciuris carolinensis*	15
European marmot*	*Marmota marmota*	14
Pocket mouse*	*Perognathus longimembris*	8
House mouse	*Mus musculus*	3
Roof rat	*Rattus rattus*	5
Whales		
Finback whale	*Balaenoptera physalus*	80
Pilot whale	*Globicephala melaena*	50

*Hibernating mammal.

Species differences in longevity and aging. That there are large differences in life-span between some species of animals has long been known, but only recently have the data become adequate for statistical analysis.

Maximum life-span as an indicator of longevity

Maximum life-span provides an estimate of the potential longevity of mammalian and avian species because of the sharp upper limit of the survival curves in life tables. Also, it is superior to the average life-span because the latter is influenced by environmental factors unrelated to aging (*e.g.*, protection by man).

The taxonomic stratification of longevity can be seen among the mammals. Primates, generally, are the longest lived group, although some small prosimians and New World monkeys have relatively short life-spans. The murid (mouselike) rodents are short-lived; the sciurid (squirrel-like) rodents, however, can reach ages two to three times longer than the murids. Three traits have independent correlations with life-span: brain weight, body weight, and resting metabolic rate. The dependence of life-span on these traits can be expressed in the form of an

equation: $L = 5.5E^{0.54}S^{-0.34}M^{-0.42}$. Mammalian life-span ($L$) in months relates to brain weight (E) and body weight (S) in grams and to metabolic rate (M) in calories per gram per hour. The positive exponent for E (0.54) indicates that longevity of mammals has a strong positive association with brain size, independent of body size or metabolic rate. The negative coefficient for metabolic rate implies that life-span decreases as the rate of living increases, if brain and body weight are held constant. The negative partial coefficient for body weight indicates that the tendency for large animals to be longer lived results not from body size but rather from the high positive correlation of body weight with brain weight and its negative correlation with metabolic rate. The same kind of relation of L to E, S, and M holds for birds, but there is a tendency for birds to be longer lived than mammals of comparable brain and body size despite their higher body temperatures and metabolic rates. The larger reptiles have life-spans exceeding those of mammals of comparable size, but their rates of metabolism are about ten times lower, so that their total lifetime energy expenditures are lower than those for mammals. The more highly cephalized animals (*i.e.*, those with higher brain weight), especially the primates, have greater lifetime energy outputs; the total lifetime energy output per gram of tissue is about 1,200,000 calories for man and 400,000 calories for domestic animals such as cats and dogs.

The above relations hold for the homeothermic mammals, those with nearly constant body temperature. The heterothermic mammals, which are able to enter daily torpor, or seasonal hibernation, thereby reduce their metabolic rates more than tenfold. The insectivorous bats of temperate latitudes are the most dramatic example; although they have life-spans in excess of 20 years, almost 80 percent of that time is spent in deep torpor. As a result, their lifetime energy expenditures are no greater than are those of other small mammals.

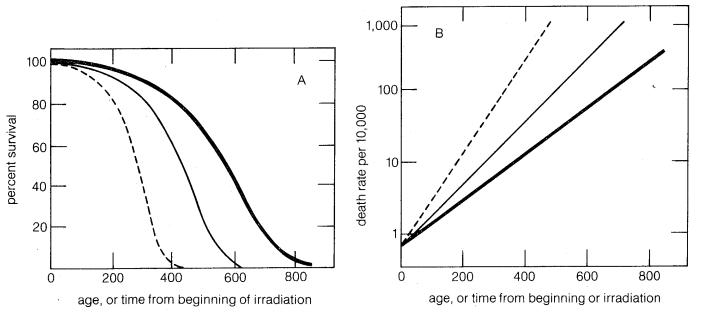
The longevities of arthropod species extend from a few days to several decades. The very short-lived insects have a brief single reproductive phase; the longer lived spiders and crustaceans are iteroparous, with annual reproductive cycles.

The inheritance of longevity. The inheritance of longevity in animal populations such as fruit flies and mice is determined by comparing the life tables of numerous inbred populations and some of their hybrids. The longevity of sample populations has been measured for more than 40 inbred strains of mice. Two experiments concur in finding that about 30 percent of longevity variation in female mice is genetically determined, whereas the heritability in male mice is about 20 percent. These values are comparable to the heritabilities of some physiological performances, such as lifetime egg or milk production, in domestic animals.

Rate of aging

The slope of the Gompertz function line indicates the rate of actuarial aging. The differences in longevity between species are the result primarily of differences in the rate of aging and are therefore expressed in differences in slope of the Gompertz function. Three species that vary in longevity, such as the opossum (7 years), rabbit (12 years), and cat (20 years), would have Gompertz functions like the three lines in B (see the Figure), and they would have survival curves like the corresponding curves in A.

By courtesy of Argonne National Laboratory, Argonne, Illinois



The relation of survivorship curves (A) to age-specific mortality rate curves (B). These curves also exemplify differences in the rate of aging (see text).

Comparison of life tables between mouse strains of a single species indicates that the strain differences result primarily from differences in age-independent hardiness factors. If strains differ in hardiness, the less hardy have death rates higher by a constant multiple at all ages, as shown by the parallel Gompertz functions. It is frequently found that the first-generation (F_1) hybrids of two inbred strains live longer than either parent. There has been no direct comparison of hybrid and inbred mice with regard to the rates of their biochemical aging processes, but life-table comparisons indicate that hybrid vigour is an increase of age-independent vigour and not a decrease in the rate of aging.

Recent research indicates that much of the variation in survival time between mouse strains is attributable to differences in inherited susceptibility to specific diseases. An important task of gerontology is to determine the extent of such genetic influences on aging.

The inheritance of longevity in humans is more difficult to investigate because length of life is influenced by socio-economic and other environmental factors that generate spurious correlations between close relatives. A number of studies have been published, most of them pointing to some degree of heritability with regard to length of life or susceptibility to major diseases, such as cancer and heart disease. Although there is disagreement about the degree of heritability of longevity in man, the evidence for genetic transmission of susceptibility to coronary heart disease and related diseases is strong, as is the evidence that monozygotic (genetically identical) twins tend to have more similar life-spans than do like-sex dizygotic (genetically different, fraternal) twins.

SENESCENCE IN MAMMALS

Changes in body composition, metabolism, and activity. The lean body mass, consisting of the skeletal muscles and all other cellular tissues, decreases steadily after physical maturity until, in extreme old age, it may be reduced to two-thirds its value in young adults. Body weight, however, usually increases with age, because stored fat and body water increases in excess of the loss of lean body mass. The relative amount of extracellular fluid increases with age during adult life, after decreasing steadily throughout fetal and postnatal development. Despite appearances, therefore, all tissues, even the skin, become more laden with water as a consequence of aging. The steady loss of voluntary (striated) muscle tissue mass throughout adult life depends somewhat on the pattern of physical activity. Evidence indicates that a large part of the loss of muscle mass with age is the result of disuse and atrophy rather than loss of muscle fibres.

The decrease of lean body mass is accompanied by a decrease in the level of overall metabolic activity. Basal metabolism is greatest during the period of most rapid mass growth; it then declines rapidly until physical maturity is reached and more slowly thereafter. In the rat the slow phase of decrease amounts to about 20 percent over a three-year period. The interior body temperature is maintained, despite lower heat production, by decreased blood flow through the skin with a consequent decrease of heat loss; the "cooling of the blood" with age, therefore, does not occur in the degree that might be inferred from the decrease in skin temperature. The amount of voluntary physical activity, such as running in an exercise wheel, typically decreases with age but varies considerably between individual animals.

Changes in structural tissues. The structural integrity of the vertebrate organism depends on two kinds of fibrous protein molecules, collagen and elastin. Collagen, which constitutes almost one-third of the body protein, is found in skin, bone, and tendons. When first synthesized by cells called fibroblasts, collagen is in a fragile and soluble form (tropocollagen). In time this soluble collagen changes to a more stable, insoluble form that can persist in tissues for most of an animal's life. The rate of collagen synthesis is high in youth and declines throughout life, so that the ratio of insoluble to soluble

collagen increases with age. Insoluble collagen then builds up with age as a result of synthesis exceeding removal, much like another fibrous tissue, the crystalline lens of the eye. With increasing age, the number of cross-linkages within and between collagen molecules increases, leading to crystallinity and rigidity, which are reflected in a general body stiffness. There is also a decrease in the relative amount of a common mucopolysaccharide (*i.e.*, the combination of a protein and a carbohydrate) ground substance; a measure of this, the hexosamine–collagen ratio, has been investigated as an index of individual differences in the rate of aging. An important consequence of these changes is decreased permeability of the tissues to dissolved nutrients, hormones, and antibody molecules.

The rate of aging of collagen is related to the overall metabolic activity of the animal, for rats kept on low calorie diets have more youthful collagen than fully nourished rats of the same age.

Elastin is the molecule responsible for the elasticity of blood vessel walls. With age, progressive loss of elasticity of vessels occurs, presumably because of fragmentation of the elastin molecule.

Chemical retardation of aging

The cross-linkage of collagen is chemically similar to the cross-linkages that occur in skins when they are tanned to leather. This similarity has stimulated proposals that chemicals that inhibit cross-linkage in tanning will retard aging. Such compounds have been tried on animals, but the problems of toxicity have not yet been solved.

Tissue cell loss and replacement. The tissues of the body fall into two groups, according to whether or not there is continuous renewal of tissue cells. At one extreme are non-renewal tissues such as nerves and voluntary muscles, in which no new cells are formed (at least in mammals) after a certain stage of growth. In renewal tissues such as the intestinal epithelium and the blood, on the other hand, some cell types live only one or a few days and must be replaced hundreds of times in the lifespan of even a short-lived animal such as the rat. Between these limits lie many organs, such as liver, skin, and endocrine organs, that have cells that are replaced over periods ranging from a few weeks to several years in man.

A peripheral nerve is a convenient object to study because the total number of fibres in the nerve trunk can be counted. This has been done for the cervical and thoracic spinal nerve roots of the rat, the cat, and man. In the ventral and dorsal spinal roots of man, the number of nerve fibres decreases about 20 percent from age 30 to age 90. In the cat, the rat, and the mouse, however, the data do not consistently indicate a decrease of number of spinal root fibres with age. In man the number of olfactory nerve fibres, which serve the sense of smell, decreases by age 90 to about 25 percent of the number present at birth, and the number of optic nerve fibres, serving vision, decreases at a nearly comparable rate.

There is a striking decrease in the number of living cells in the cerebral cortex of the brain of humans with age. The cerebellar cortex of the rat and man is about as susceptible to age deterioration as is the cerebral cortex. Other parts of the brain are not so obviously marked by aging.

There is, in short, a tendency for the higher and more recently evolved levels of the nervous system to undergo more severe aging loss than do other regions, such as the brain stem and spinal cord. It is not yet known how much of the loss of brain cells results from conditions within the brain itself and how much results from extrinsic causes, such as deterioration of the blood circulation. The nutrition and maintenance of nerve cells, or neurons, in the central nervous system depends to a considerable extent on neuroglia, small cells that surround the neurons. The absolute number of these cells apparently does not decrease with age, but some of the microscopic changes seen in the neurons of old persons are similar to the changes produced by starvation or physical exhaustion.

It has been shown that after an attack of measles, the virus remains in the host's body for the remainder of life and very infrequently gives rise to a rapidly progressing degeneration of the cerebral cortex. This virus or other inapparent viruses may also be responsible for the individual differences in onset of senility in man.

The renewal tissues are typically made up of a population of proliferative cells, which retain the capability for division, and a population of mature cells, produced by the proliferative cells and with limited life-spans. The production of cells must balance the steady loss and also compensate quickly for unusual losses caused by injury or disease, so each renewal tissue has one or more channels of feedback control to adjust production to demand.

Senescence of renewal tissues

Aging of renewal tissues is expressed in several ways, including decrease in the number of proliferative cells, decrease in the rate of cell division, and decrease in responsiveness to feedback signals. Changes of these factors in the blood-forming tissues of the mouse are small, yet the blood-forming tissues do suffer an aging deficit, for the ability to respond to extreme or repeated demand is significantly reduced in older mice.

The intact skin has a cell turnover time of several weeks, with the capability, shared by all renewal tissues, of temporarily increasing the rate of cell production by a large factor in response to injury. The rate of wound healing decreases with age, rapidly at first and more slowly as age increases.

One of the most regular and striking aging processes is the decrease in the ability to focus on both near and very distant objects. This loss in visual accommodation is the result in part of a weakening of the ciliary muscle of the eye and of a decrease in the flexibility of the lens. A further contributing factor, however, is that the lens continues to grow throughout life at a rate that diminishes with age. This growth is the result of continuous division of epithelial cells near an imaginary midline of the lens, giving rise to fresh cells that differentiate into the precisely aligned lens fibres. Once formed, the fibres remain permanently in place.

An important feature of the renewal mechanism is the stem cell. These cells, which may normally continue to divide at a low rate throughout life, under conditions of increased demand enter a compensatory proliferative phase during which they divide rapidly. Blood-forming tissue has a stem-cell population that responds to injury readily in youth, but its capacity diminishes with age. The increased incidence of anemia in old age and the reduced capacity to respond to blood loss have been attributed to depletion of the blood-forming stem cells. Stem-cell populations have not been identified with certainty in other proliferative tissues. The intestinal mucosa, in particular, has a high cell-division rate without any clear indication of a reserve population of stem cells.

Mammalian cell cultures. Dividing cells from various mammalian tissues can be grown in vitro (outside the body) under careful laboratory control. Various lines of cancer cells have been grown in continuous culture for many decades. In the early period of tissue-culture technology it was claimed that certain chicken cells (fibroblasts) had been maintained in culture for 20 years. This led to the belief that dividing cells were potentially immortal and focussed interest on nondividing cells as the seat of the aging process. This view has lost standing in recent years. It has now been established that a population (clone) of fibroblasts has a finite life history in culture. It has a period of healthy growth, during which it can be transferred, or "split," several dozen times, indicating that the cells have undergone more than that number of generations. The cultures, however, go into a senescent phase and die out, usually before the 50th transfer. Occasionally, the chromosomes in a cell in the culture undergo a mutation (change) that results in a loss of a growth-limiting factor, leading to the establishment of a subclone capable of indefinite growth. This happens fairly often in cultures of mouse cell strains but only rarely in cultures of human cells. Such mutations usually involve chromosomal rearrangements or changes in the number of chromosomes. The present view, therefore, is that dividing mammalian cells with a normal chromosomal complement have a limited growth potential and that the capacity for indefinite growth shown by

cancer cells and transformed cells is the result of the loss of a growth-limiting factor. The number of transfers that cell strains can undergo decreases as the age of the donor increases, in a way reminiscent of the decreased turnover rate of fibroblasts in living chickens and of the decreased rate of wound healing with age.

Wasting of tissues with age

Changes in tissue and cell morphology. There are numerous instances of tissue changes with age. The atrophy of tissues of moderate degree is usual. The shrinkage of the thymus is especially striking and important in view of its role in immunological defense. The diminution of cellular tissue and replacement by fatty or connective tissue is prominent in marrow and skin. In the kidney, entire secretory structures (nephrons) are lost. The secretory cells of the pancreas, thyroid, and similar organs decrease in numbers.

An important age change is the accumulation of pigments and inert—possibly deleterious—materials within and between cells. The pigment lipofuscin accumulates within heart muscle cells; it is not detectable at ten years of age but rises to almost 3 percent of the cell volume by age 90. Amyloid substance, a protein–carbohydrate complex, increases in tissues in middle age; it is presumably a product of autoimmune reactions, immune reactions misdirected against the organism itself. In an extreme case of rare autoimmune disease, amyloid disease, particular organs are virtually choked with amyloid substance. Trace metals also accumulate in various tissues with age, and although the amounts are very small, certain metals can poison enzyme systems, stimulate mutations, or cause cancer.

AGING AS LOSS OF INFORMATION

Aging of genetic information systems. The physical basis of aging is either the cumulative loss and disorganization of important large molecules (*e.g.*, proteins and nucleic acids) of the body or the accumulation of abnormal products in cells or tissues. A major effort in aging research has been focussed on two objectives: to characterize the molecular disruptions of aging and to determine if one particular kind is primarily responsible for the observed rate and course of senescence; and to identify the chemical or physical reactions responsible for the age-related degradation of large molecules that have either informational or structural roles. The working molecules of the body, such as enzymes and contractile proteins, which have short turnover times, are not thought to be sites of primary aging damage. The deoxyribonucleic acid (DNA) molecules of the chromosomes appear to be potential sites of primary damage, because damage to DNA corrupts the genetic message on which the development and function of the organism depend. Damage at a single point in the DNA molecule can be followed by the synthesis of an incorrect protein molecule, which may result in the malfunction or death of the host cell or even of the entire organism. Attention therefore has been given to the somatic mutation hypothesis, which asserts that aging is the result of an accumulation of mutations in the DNA or somatic (body) cells. Aneuploidy, the occurrence of cells with more or less than the correct (euploid) complement of chromosomes, is especially common. The frequency of aneuploid cells in human females increases from 3 percent at age ten to 13 percent at age 70. Each DNA molecule consists of two complementary strands coiled around each other in a double helix configuration. Evidence indicates that breaks of the individual strands occur with a higher frequency than was once suspected and that virtually all such breaks are repaired by an enzymatic mechanism that destroys the damaged region and then resynthesizes the excised portion, using the corresponding segment of the complementary strand as a model. The mutation rate for a species is therefore governed more by the competence of its repair mechanism than by the rate at which breaks occur. This may help to explain why the mutation rates of different species are roughly proportional to their generation times and justifies research to determine whether the enzymatic mechanisms involved are accessible to control. It remains to be seen whether a reduction of mutation rates will retard the onset of generalized aging or of specific disease process.

There are, however, serious objections to the somatic mutation theory. The wasp *Habrobracon* is an insect that reproduces parthenogenetically (*i.e.*, without the need of sperm to fertilize the egg). It is possible to obtain individuals with either a diploid, or paired, set of chromosomes, as in most higher organisms, or a haploid, single, set. Any gene mutation in a haploid cell at an essential position would result in loss of a vital process and impairment or death of the cell; in a diploid cell a serious mutation is often compensated for by the complementary gene and the cell can carry on its vital functions. Experiments have shown that haploid wasps live about as long as diploids, implying either that mutations are not a quantitatively important factor in aging or that parthenogenetic species have compensated for the vulnerability of their haploids by developing an increased effectiveness of DNA repair.

Chromosomes can be separated into DNA and protein molecules, but with increasing difficulty in older cells. The isolated DNA of old animals, however, does not differ from that of the young. Although most of the DNA in a given cell at a given time is repressed (*i.e.*, blocked from functioning), it is more repressed in old animals; it is not yet known whether this is a primary age change or a consequence of reduced cell metabolism arising from other causes.

Aging of the immunological system. Another important molecular information system of the body is the immunological system, part of which, the thymus-dependent subsystem, is specialized for defense against invading micro-organisms and for the as yet poorly understood role of detecting and removing body cells that have changed in such ways that they are no longer recognized by the body as part of its own substance, leading to the autoimmune reactions mentioned above. The immunological system has been implicated in the body's defenses against cancer. Cancerous growths (neoplasms) are thought to arise from single cells that undergo a drastic transformation as a result of either a genetic mutation or the activation of a latent (hidden) virus that may have been transmitted genetically from parent to offspring. The control of cancer susceptibility by genetically governed defense mechanisms has been indicated by the breeding of high and low cancer susceptibility in mice. There is a growing body of evidence that the thymus-dependent immunological system is instrumental in repressing the development of cancer.

Age-related disorders of the immunological system

One piece of evidence is that the immunosuppressive procedures of organ transplantation are often followed by a greatly increased incidence of neoplasms. The thymus-dependent system can itself, however, give rise to age-related autoimmune disease, in which the immunological system perceives normal body tissue as foreign and attacks it with antibodies. The initial step in these diseases is considered to be a somatic mutation in a single cell of the immunological system. Such considerations are the basis of several immunological theories of aging, which seek to explain the phenomena of senescence in terms of mutations in the immunological system.

Aging of neural and endocrine systems. The loss of psychological and neurophysiological capacities with age is undoubtedly the result, in large part, of the loss of neurons, but deficiencies in the metabolic processes of the surviving cells are demonstrably involved. The ability of the eye to dark-adapt (*i.e.*, increase its sensitivity at low light levels) decreases with age, but part of that decrease can be restored by breathing pure oxygen. Various mental processes in old people are also found to be improved by breathing oxygen. The establishment of a memory trace (connections in the brain that are associated with memory) involves the synthesis of protein; any slowed induction of protein synthesis, as from lower oxygen intake, with age could be a factor in the deficits of learning and memory of old people.

A general characteristic of aging of the endocrine system is that the cells that once responded vigorously to hormones become less responsive. A normal chemical in

cells, cyclic adenosine monophosphate (AMP), is thought to be a transmitter of hormonal information across cell membrane; it may be possible to identify the specific sites in the membrane or the cell interior at which communication breaks down.

THE CAUSES OF AGING

External environmental agents. *Ionizing radiations.* The shortening of life caused by ionizing radiations (*e.g.*, X-rays) has been determined for many species, including mice, rats, hamsters, guinea pigs, and dogs. The occurrence of some diseases, such as leukemia, may increase disproportionately after irradiation, with the degree of increase influenced by age and sex.

The permanent nature of radiation damage is shown by the comparison of life-spans of irradiated and control populations. An irradiated population dies out like a chronologically older unirradiated population. Members of a population given a single dose of X-rays or gamma rays in early adult life die of the same diseases that afflict the unirradiated control population, but they die months or even years earlier.

Continuous irradiation throughout life at low dose rates (daily doses from one-thousandth to one-tenth the dose that would kill immediately) speeds the mortality process (see Figure). It is not yet clear if the molecular damage produced by such irradiation is the same as the molecular changes that accompany natural aging. Studies of animals and of cells grown in culture suggest that large doses of radiation kill by producing deleterious rearrangements of chromosomes in the proliferative cell population. Such aberrations also increase with age, but they seem to be less important in the natural aging process. At low radiation doses, chromosome aberrations become relatively less important than other effects, and the primary radiation damage in these conditions may bear a closer relation to the aging lesion. Under conditions of low-dose irradiation, however, the only definite effect is slightly increased cancer incidence; a generalized aging effect has not yet been observed.

The aging effect of natural radiation

Natural radioactivity in the body, arising mostly from radioactive potassium and radium, and natural background irradiation, from the Earth and from cosmic rays, are not major contributors to the aging process, even in the long-lived human species. They are responsible, however, for a small percentage of the cancer incidence in man. Although the dose to the body from medical radiations is a fraction of the background level and the radiation from nuclear weapon test fallout is less than 1 percent of the background, both sources contribute to cancer induction in proportion to their amounts.

Temperature. Flour beetles, fruit flies, fishes, and other poikilothermic (temperature-variable) organisms live longer at the lower range of environmental temperature. These observations led to the rate-of-living hypothesis, which, simply stated, holds that an organism's life-span is dependent on some critical substance that is exhausted more rapidly at higher temperature. Careful analysis of the data on temperature–longevity relations shows, however, that the rate-of-living hypothesis is inadequate in its original form. The most telling evidence comes from experiments in which fruit flies were kept at one temperature for part of their lives and at another temperature for the remainder. The results are not consistent with the rate-of-living hypothesis, but no satisfactory theory has appeared as yet to take its place. An important factor that has not yet been adequately taken into account is the relation of metabolic efficiency to temperature. The energy cost of the biosynthetic processes studied has been discovered to be minimal at an intermediate temperature in the range to which the species is adapted and to increase at higher or lower temperatures. A related phenomenon holds for longevity; the number of calories expended by fruit flies per lifetime is maximal at an intermediate temperature, so the rate of aging per calorie is minimal at that temperature.

There is a question of the degree to which aging occurs as a result of heat destruction (thermal denaturation) of proteins. Thermal denaturation is predominately a disruption of the folding of molecules, which requires the breaking of numbers of low-energy bonds. It seems not to be a strong contributing factor to aging. There still exists the possibility that rare events, such as mutations, may arise to a significant degree from thermal denaturation.

Physical wear of non-renewable structures. One of an animal's most important assets is his chewing apparatus, including jaws and teeth. Adaptation to tooth rate of wear is especially important for animals that consume large quantities of grass and herbage. Such adaptations include higher tooth crowns (hypsodonty), larger grinding area, and longer tooth growth period. Tooth wear may be limiting for survival in adverse environments, but, on the whole, it is not an important life-limiting characteristic. The same can be said for other external organs subject to physical wear.

Infectious disease and nutrition. The populations in poor environments, characterized by high rates of infectious disease and poor nutrition, have higher death rates than populations in good environments at all ages, yet there is no positive evidence that disadvantaged populations experience a higher rate of aging.

Rats kept on diets restricted in calories live longer and have lower cancer incidence than do rats that are allowed to eat at will; however, maximum longevity is achieved at a nutritional level that keeps the animal sexually immature and much below normal weight.

Internal environment: consequences of metabolism. The metabolic activities of organisms produce highly reactive chemicals, including strong oxidizing agents. The internal structure of the cell, however, minimizes the harmful effects of such agents; the critical reactions take place within enclosed structures such as ribosomes, membranes, or mitochondria, and counteractive enzymes such as peroxidases are present in abundance. It is nevertheless likely that low concentrations of these reactive substances can reach vital molecules and contribute to the characteristic rate of aging injury. Experiments in which mice are fed low levels of antioxidants such as butylated hydroxytoluene (BHT) have been encouraging but are still somewhat equivocal.

Aging and membrane integrity

Membranes are the site of much of the metabolic activity of cells; they provide the barriers that keep incompatible reactions separated. Certain membrane structures, called lysosomes, contain enzymes capable of digesting the cell if released; the stability of cells and organisms is therefore very much bound up with the stability of membranes. A number of drugs, including corticosteroids, salicylates, and antihistamines, act by stabilizing cell membranes against inflammatory stimuli. Some of them are found to prolong life in fruit flies and to prolong survival of cells in vitro. The mode of action of these drugs is connected to substances called prostaglandins, which can alter specific membrane characteristics.

BIBLIOGRAPHY. A.M. BRUES and G.A. SACHER (eds.), *Aging and Levels of Biological Organization* (1965), the report of an interdisciplinary conference, with much discussion of immunological, genetic, cellular, environmental, and evolutionary aspects of aging; ALEX COMFORT, *Ageing: The Biology of Senescence*, rev. ed. (1964), the best general treatise, except for biochemical mechanisms; PETER L. KROHN (ed.), *Topics in the Biology of Ageing* (1966), with special emphasis on cellular mechanisms of aging; P.J. LINDOP and G.A. SACHER (eds.), *Radiation and Ageing* (1966), pathological changes with aging, life table characteristics of laboratory animal populations, and the influence of ionizing radiations; NATHAN W. SHOCK (ed.), *Perspectives in Experimental Gerontology* (1966), a volume of collected brief essays on aging; B.L. STREHLER, *Time, Cells, and Aging* (1962), an examination of cellular and molecular mechanisms and theories of aging; G.E.W. WOLSTENHOLME and MAEVE O'CONNOR (eds.), *Ciba Foundation Colloquia on Ageing*, 5 vol. (1955–59), the best source for molecular and cellular theories and data: vol. 1 covers *General Aspects;* 2, *Aging in Transient Tissues;* 3, *Methodology;* 4, *Water and Electrolyte Metabolism;* and 5, *Lifespan of Animals*, with some consideration of ecological and evolutionary aspects; *United Nations Demographic Yearbook*, includes mortality statistics for all countries, showing the influence of economic, social, and climatic factors on mortality.

(G.A.Sa.)

Aging, Human

Aging consists of the progressive changes that take place in a cell, an organ, or the total person with the passage of time. It is a process that goes on over the entire adult life-span and is as much a part of living as are infancy, childhood, and adolescence. The science of gerontology (the study of aging) is concerned primarily with the changes that occur between the attainment of maturity and the death of the individual. The goal of research in gerontology is to identify the factors that influence these changes. Application of this knowledge is expected to reduce the disabilities now associated with aging.

Aging has many aspects, but the primary ones fall into three major categories; namely, biological-physiological, psychological-behavioral, and social-economic.

The biological-physiological aspects of aging include both the basic biological factors that underlie aging and the general health status. Since the probability of death increases rapidly with advancing age, it is clear that changes must occur in the individual which make him more and more vulnerable to disease. For example, a young adult may rapidly recover from pneumonia, whereas an elderly person may die.

Physiologists have found that the performance of many organ systems such as the heart, kidneys, or lungs shows a gradual decline over the life-span. Part of this decline is due to a loss of cells from these organs, with resultant reduction in the reserve capacities of the individual. Furthermore, the cells remaining in the elderly individual may not perform as well as those in the young. Certain cellular enzymes may be less active, and thus more time may be required to carry out chemical reactions. Ultimately the cell may die.

BIOLOGICAL THEORIES OF AGING

Genetic theories. One theory of aging assumes that the life-span of a cell or organism is genetically determined —that the genes of an animal contain a "program" that determines its life-span just as eye colour is determined genetically. Although long life is recognized often as a familial characteristic, and short-lived strains of fruit flies, rats, and mice can be produced by selective breeding, other factors clearly can significantly alter the basic genetic program of aging.

Another genetic theory of aging assumes that cell death is the result of "errors" introduced in the formation of key proteins, such as enzymes. Slight differences induced in the transmission of information from the deoxyribonucleic acid (DNA) molecules of the chromosomes through ribonucleic (RNA) molecules (the "messenger" substance) to the proper assembly of the large and complex enzyme molecules could result in a molecule of the enzyme that would not "work" properly. These so-called error theories have not yet been firmly established, but studies are in progress.

As cells grow and divide, a small proportion of them undergo mutation; that is, they become "different" with a change in their chromosome structure that is then reproduced when they again divide. The "somatic mutation" theory of aging assumes that aging is due to the gradual accumulation of mutated cells that do not perform normally.

Non-genetic theories. Other theories of aging focus attention on factors that can influence the expression of a genetically determined "program." One of these is the "wear-and-tear" theory, which assumes that animals and cells, like machines, simply wear out. Animals, however, unlike machines, have some ability to repair themselves, so that this theory does not fit the facts of a biological system. A corollary to the wear-and-tear theory is the presumption that waste products accumulate within cells and interfere with function. The accumulation of highly insoluble particles, known as "age pigments," has been observed in muscle cells in the heart and nerve cells of both human beings and other animals.

With increasing age, tendons, skin, and even blood vessels lose elasticity. This is due to the formation of cross-links between or within the molecules of collagen (a fibrous protein) that give elasticity to these tissues. The "cross-linking" theory of aging assumes that similar cross-links form in other biologically important molecules, such as enzymes. These cross-links could alter the structure and shape of the enzyme molecules so that they are unable to carry out their functions in the cell.

Auto-immunity; control theories

Another theory of aging assumes that immune reactions, normally directed against disease-producing organisms as well as foreign proteins or tissue, begin to attack cells of the individual's own body. In other words, the system that produces antibodies loses its ability to distinguish between "self" and foreign proteins. This "auto-immune" theory of aging is based on clinical rather than on experimental evidence.

These theories all attempt to explain aging in terms of cellular and molecular changes. Actually, age changes are much more marked in the overall performance of an individual than in cellular processes that can be measured. The age decrement in the ability to perform muscular work is much greater than any changes that can be detected in the enzyme activities of the muscles that perform the work. It is possible that aging in an individual is actually due to a breakdown in the control mechanisms that are required in a complex performance.

Aging has many facets. Hence there are a number of theories, each of which may explain one or more aspects of aging; as yet there is no single theory that explains all of the phenomena of aging.

PSYCHOLOGICAL ASPECTS OF AGING

Intellectual functions in the aged

The most outstanding psychological features of aging are the impairment in short-term memory and the lengthening of response time. Both of these factors contribute to lower scores of the elderly on standard tests of "intelligence." When the aged are given all the time that they wish on tests that are not heavily dependent on school skills, their performance is only slightly poorer than that of young adults. Age decrements are negligible on tests that depend on vocabulary, general information, and well-practiced activities. The age decrement is largest on tests in which associations formed early in life must be displaced by new ones.

Experimental studies on learning show that, although the elderly learn more slowly than the young, they can acquire new material and can remember it as well as the young. Age differences in learning increase with the difficulty of the material to be learned. Learning can be improved in the elderly, however, by a change in teaching methods. For example, when more time is given the subject to respond, age differences in learning decrease.

Age itself is no bar to creativity, and many major achievements in the arts and sciences have been reported for persons of advanced age, but systematic surveys indicate that outstanding achievements, especially in mathematics, physics, and biology, are more likely to be made by individuals in their 30s than in their 60s. On the other hand, achievements in politics, administration, and religion, areas in which experience is a dominant factor, are most likely to occur later in life.

Aged people tend to become more cautious and rigid in their behaviour and to withdraw from social contacts. These behaviour patterns may be the result of social institutions and expectancies rather than an intrinsic phenomenon of aging. Many persons who "age successfully" make conscious efforts to maintain mental alertness by continued learning and by expansion of social contacts with individuals in a younger age group. The maintenance of a broad spectrum of activities and interests is essential for good adjustment in old age. Two other factors that are helpful in this respect are good health and an adequate income. Older people who live independently in the community are more likely to be well adjusted than those living in institutions for the aged.

SOCIAL AND ECONOMIC ASPECTS OF AGING

Social and economic problems of the aged in advanced nations arise primarily from the increasing number of the elderly in the population and the shift from a rural to an

urban society. In 1900, in the United States, there were only 3,100,000 people over the age of 65 in a total population of about 76,000,000 (4 percent). By 1965 the number over age 65 had reached 18,200,000 (9.1 percent of the total population) with an estimate of 20,000,000 by 1970 (10 percent of the population) and 25,000,000 by 1985. In Japan in 1967 about 6,700,000 persons (about 6.7 percent) were age 65 or over in a population of about 100,000,000; it is estimated that in France, in 1966, 6,200,000 persons (12.5 percent) in a population of 49,400,000 were age 65 or over. In a rural society, aged parents and grandparents often live in the homes of their children or grandchildren. In an urban society, two- and three-generation families are rare because of limitation of space and changing cultural patterns. Consequently, the need for housing, recreational facilities, and medical services for elderly people has grown at an overwhelming rate.

Life expectancy

In 1900, in the United States, the average life expectancy at birth was 47.9 years for males and 51.1 years for females. Estimates for 1975 are 69.5 for males and 75.8 for females. (Figures for Japan and France for 1966 were comparable.) This marked increase in life expectancy was due primarily to the substantial reduction in deaths of infants and children. The average years of life remaining for males after age 65 increased from 11.5 years in 1900 to 13.0 years in 1967, with an estimate of 14.0 years by 1975. For women, the value increased from 12.2 years in 1900 to 16.5 years in 1967, with an estimate of 16.9 years by 1975.

The leading causes of death in the United States are cardiovascular diseases (50 percent of all deaths in 1960–62), cancer (15.8 percent of all deaths in males and 18.2 percent of female deaths), and accidents (8.4 percent of male deaths and 4.6 percent of female deaths). It has been estimated that, if cardiovascular disease were eliminated, life expectancy at birth would increase to 79 years in males and 87 years in females. Life expectancy after age 65 would increase to 24 years in males and 29 years in females. Although continued research will undoubtedly result in a reduction in death rates from cardiovascular diseases, it is doubtful whether they can be completely eliminated in the foreseeable future, so that this potential gain will be only partially realized. A 30 percent reduction in death rates would add only 3–4 years to average life-span after the age of 65. (Data published elsewhere are based on these United States statistics.)

Income

The advent of Old Age and Survivors Insurance and Medicare has greatly improved the economic situation of the aged in the United States. The average income of individuals over the age of 65, however, is still low in comparison with the total population and living costs. As a group, 60 percent of the population over age 65 received incomes below the poverty level in 1968.

With increasing age the need for medical and hospital services increases. For example, in 1967 the average number of visits to a physician was 6.7 times for individuals over age 65 compared with 4.7 visits per year for those aged 45–65. Persons over the age of 65 spent an average of 3.8 days per year in bed as a result of acute illnesses, compared with 2.8 days by 45–64-year-olds. Chronic diseases produce some degree of disability in about 30 percent of the population over age 65, but only 5 percent are confined to hospitals or nursing homes. Medicare and Medicaid programs now offer partial coverage of costs, but additional physicians, nurses, hospitals, and nursing homes are needed to meet the needs.

EFFECT OF AGING OF THE BODY SYSTEMS

Cardiovascular system. Diseases of the heart account for over 50 percent of deaths after age 65. Thus, with increasing age the heart becomes more vulnerable to disease. Even in the absence of detectable disease, the heart undergoes deleterious changes with advancing age. Structural changes include a gradual loss of muscle fibres with an infiltration of fat and connective tissue. There is a gradual accumulation of insoluble granular material (lipofuscin, or "age pigment") in cardiac muscle fibres. These granules, composed of protein and lipid (fat), make their first appearance by the age of 20 and increase gradually, so that by the age of 80 they may occupy as much as 5–10 percent of the volume of a muscle fibre.

The heart also shows a gradual reduction in performance with advancing age. The amount of blood pumped by the heart diminishes by about 50 percent between the ages of 20 and 90 years. There are marked individual differences in the effects of age. For example, some 80-year-old individuals may have cardiac function that is as good as that of the average 40-year-old individual.

Under resting conditions, the heart rate does not change significantly with age. During each beat, however, the muscle fibres of the heart do not contract as rapidly in the old as in the young. This reduction in power, or rate of work, is due to the age-associated reduction in the activities of certain cellular enzymes that produce the energy required for muscular contraction.

In spite of these changes, the heart, in the absence of disease, is able to meet the demands placed upon it. In response to physical exercise it can increase its rate to double or triple the amount of blood pumped each minute, although the maximum possible output falls, and the reserve capacity of the heart diminishes with age.

Arteriosclerosis, or hardening of the arteries, increases markedly in incidence with age, and is often regarded as part of aging. This is not necessarily true. Arteriosclerosis may appear even in adolescents. It is a progressive disorder and is present to some extent in practically all individuals by middle life. It is, therefore, impossible to make a clear distinction between the effects of aging and the effects of disease in blood vessels in human beings. In some animal species, as, for example, the rat, that do not develop arteriosclerosis, age changes in the heart and blood vessels can be identified.

Blood vessels and blood pressure

In general, blood vessels become less elastic with advancing age. There is a progressive thickening of the walls of larger blood vessels with an increase in connective tissue. The connective tissue itself becomes stiffer with increasing age. This occurs because of the formation of cross-links both within the molecules of collagen, a primary constituent of connective tissue, and between adjacent collagen fibres. These changes in blood vessels occur even in the absence of the deposits on the arterial wall characteristic of atherosclerosis, which interfere with blood flow through the arteries. The gradual loss of elasticity increases with resistance to the flow of blood so that blood pressure may increase. This in turn increases the work that the heart must do in order to maintain the flow of blood.

While both systolic and diastolic blood pressures (blood pressures at contraction and dilation of the heart, respectively) increase with age, the rate of systolic increase exceeds that of diastolic so that the pulse pressure widens (see the Table). The increase in pressure stops in the eighth decade of life, and there may even be a slight decline in pressure in extreme old age.

Mean Blood Pressure in Adults by Age and Sex, 1960–62

(millimetres of mercury)

age (years)	men		women	
	systolic	diastolic	systolic	diastolic
18–24	121.7	71.6	111.8	69.4
25–34	124.7	76.4	115.6	72.9
35–44	128.6	80.7	122.8	78.0
45–54	133.8	83.2	133.8	82.0
55–64	140.3	83.1	146.6	84.9
65–74	148.0	81.0	160.2	83.7
75–79	154.3	79.4	156.6	79.3

Source: U.S. Public Health Service.

On the average, obese people have higher blood pressures than those with normal body weights. Since the incidence of obesity increases with age at least up to the age of 55–60, this factor may contribute in part to the increase in blood pressure with age.

Digestive system. Loss of teeth, which is often seen in elderly people, is more apt to be the result of long-term neglect than a result of aging itself. The loss of teeth and incidence of oral disease increase with age, but, as programs of water fluoridation are expanded and the incidence of tooth decay in children is reduced, subsequent generations of the elderly will undoubtedly have better teeth than the present generation.

While it is true that the secretion by the stomach of hydrochloric acid, as well as other digestive enzymes, decreases with age, the overall process of digestion is not significantly impaired in the elderly. Sugar, proteins, vitamins, and minerals are absorbed from the stomach and intestine as well in the elderly as in the young. Some investigations indicate a slight impairment in fat absorption, but the reduction is probably of little practical significance.

These findings have important implications for nutrition of the elderly. There is no evidence that the intake of any nutrient, such as vitamins and minerals, need be increased in the elderly because of impaired absorption. Nutritional deficiencies can be avoided as long as the diet is varied to assure adequate intake of all nutritional elements. Deficiencies are most likely to develop from poor eating habits, such as excessive intake of carbohydrate with a reduction in protein. In the elderly these deficiencies are most apt to be in the intake of protein, calcium, iron, vitamin A, and thiamine (also called vitamin B_1).

Nervous system. Changes in the structures of the brain due to normal aging are not striking. It is true that with advancing age there is a slight loss of neurons (nerve cells) in the brain. This is because, in the adult, neurons have lost the capacity to form new neurons by division. The basic number of neurons in the brain appears to be fixed by about the age of ten. The total number of neurons is extremely large, however, so that any losses probably have only a minor effect on behaviour. Since the physiological basis of memory is still unknown, it cannot be assumed that the loss of memory observed in elderly people is caused by the loss of neurons in the brain.

Neurons are extremely sensitive to oxygen deficiency. Consequently, it is probable that neuron loss, as well as other abnormalities observed in aging brains, results not from aging itself, but from disease, such as arteriosclerosis, that reduces the oxygen available to areas of the brain by reducing the blood supply.

Changes in the brain

There are probably functional changes in the brain that account for the slowing of responses and for the memory defects that are often seen in the elderly; and even small changes in the connections between cells of the brain could serve as the basis for marked behavioral changes, but, until more is known about how the brain works, behavioral changes cannot be related to physiological or structural changes. It is known that, because of the slow course of aging, the nervous system can compensate and maintain adequate function even in centenarians.

Human behaviour is highly dependent on the reception and integration of information derived from sensory organs, such as the eye and ear, as well as from nerve endings in skin, muscle, joints, and internal organs. There is, however, no direct relation between the sensitivity of receptors and the adequacy of behaviour, because the usual level of stimulation is considerably greater than the minimum required for stimulation of the sense organs. In addition, an individual adapts to gradual impairments in one sensory organ by using information available from other sense organs. Modern technology has also provided glasses and hearing aids to compensate for reduced acuity in the sense organs.

The incidence of gross sensory impairments, of which many are the result of disease processes, increases with age. One survey conducted in the United States classified 25.9 per 1,000 persons aged 65–74 as blind, in contrast to 1.3 per 1,000 aged 20–44 years. In the age group 65–74, 54.7 per 1,000 persons were classified as functionally deaf, compared with 5.0 per 1,000 in the age range 25–34 years.

Vision. Visual acuity (ability to discriminate fine detail) is relatively poor in young children and improves up to young adulthood. From about the middle 20s to the 50s there is a slight decline in visual acuity, and there is a somewhat accelerated decline thereafter. This decline is readily compensated for by the use of eyeglasses. There is also reduction in the size of the pupil with age. Consequently, vision in older people can be significantly improved by an increase in the level of illumination.

Aging also brings about a reduction in the ability to change the focus of the eye for viewing near and far objects (presbyopia), so that distant objects can ordinarily be seen more clearly than those close at hand. This change in vision is related to a gradual increase in rigidity of the lens of the eye that takes place primarily between the ages of 10 and 55 years. After age 55 there is little further change. Many people in their 50s adopt bifocal glasses to compensate for this physiological change.

The sensitivity of the eye under conditions of low illumination is less in the old than in the young; that is, "night vision" is reduced. Sensitivity to glare is also greater in the old than in the young.

The incidence of diseases of the eye, such as glaucoma and cataracts (characterized, respectively, by increased intra-ocular pressure and opaque lenses), increases with age, but recent advances in surgery and the development of contact lenses have made it possible to remove cataracts and restore vision to many individuals.

Hearing. Hearing does not change much with age for tones of frequencies usually encountered in daily life. Above the age of 50, however, there is a gradual reduction in the ability to perceive tones at higher frequencies. Few persons over the age of 65 can hear tones with a frequency of 10,000 cycles per second. This loss of perception of high frequencies interferes with identifying individuals by their voices and with understanding conversation in a group, but does not ordinarily represent a serious limitation to the individual in daily life. Listening habits and intellectual level play an important role in determining the ability to understand speech, so that there is often a disparity between measurements of pure tone thresholds and ability to perceive speech.

Other sensory impairment. After the age of 70 other sense organs may show a reduction in sensitivity. Reduced taste sensitivity is associated with atrophy and loss of taste buds from the tongue in the elderly. The effect of aging on the sense of smell has not been precisely determined because this sense is extremely difficult to assess quantitatively; in addition, smoking and exposure to occupational odours and noxious substances in the air influence sensitivity to smells.

Sensitivity to pain is difficult to evaluate quantitatively under controlled laboratory conditions. There is some evidence that it diminishes slightly after the age of 70.

There is a general slowing of responses in the elderly. Reflexes become slightly more sluggish and the speed of conduction of impulses in nerves is slightly slowed. Old people require more time to respond to the appearance of a light than do young. The slowing with age is greater in situations where a decision must be made. For example, more time is required to initiate a response in experiments in which the instructions are "Press the button with your right hand when the green light comes on, but with your left hand when the red light comes on" than if the instructions are, "Push the button if either light comes on." From these and other experiments it is concluded that the primary site of slowing of responses is within the brain rather than in the end organ (eye) itself.

Skin. The primary age change in the skin is a gradual loss of elasticity. Although this basic change plays a role, other factors, such as exposure to the weather and familial traits, also contribute to the development of wrinkles and the pigmentation associated with senescence. The ability of the skin to take up slack and remain closely adherent to the underlying structures is due to the presence of fibres of the proteins elastin and collagen. Studies of the minute structures of the skin show a gradual reduction in elastin. In addition, the collagen fibres show

an increase in cross-links, which greatly restricts the elastic properties of the collagen network.

The effectiveness of facial massage in retarding the development of wrinkles has not been evaluated under carefully controlled conditions. The application of creams containing female sex hormone stimulates regeneration of skin and improves its elastic properties. Other effects, which may be undesirable or even hazardous, may follow repeated administration of these hormones.

Endocrine system. Because of the importance of hormones in the regulation of many physiological systems, impairments in endocrine (ductless) glands have traditionally been cited as important determinants in aging.

Thyroid and adrenal hormones

Thyroxine, the hormone secreted by the thyroid gland, regulates the level of activity of all the cells of the body. When thyroxine secretion is reduced, all metabolic processes proceed at a reduced rate and basal metabolism falls. (Metabolism consists of the chemical changes taking place within the cells of an organism during the processes of growth and restoration of tissues and the production of energy necessary for bodily processes; basal metabolism is the metabolism, as measured by the rate at which heat is given off, when an organism is in a resting and fasting state.) Since basal metabolism decreases with age, it seemed reasonable to ascribe aging to a loss of thyroid function, but this assumption has proved to be incorrect. Experimental studies have shown that the ability of the thyroid gland to produce thyroxine is not reduced in the elderly, and that there is a reduction in the utilization of thyroxine in various tissues of the body. Further studies of cellular metabolism are needed to find out why this is so.

Since aging is associated with reduced ability to adjust to stresses, and since the adrenal cortex (the outer part of the adrenal gland) plays a role in many of these adjustments, numerous attempts have been made to assess senescent changes in the function of the adrenal cortex. Although after the age of 50 there is a reduction in blood levels of the hormones secreted by the adrenal cortex, the ability of the gland to produce hormones when stimulated by the experimental administration of adrenocorticotrophic hormone (ACTH), the pituitary hormone that regulates the activity of the adrenal cortex, has been shown to be as good in the old as in the young.

Hormones of pituitary and of pancreas

The pituitary gland is often referred to as the master gland of the body, since it produces hormones that stimulate the activities of other endocrine glands, such as the adrenal, the thyroid, and the ovary. It was therefore once assumed that reduction in the function of these glands associated with aging is due to lack of proper stimulation from the pituitary gland. New methods for determination of the very small amounts of these regulating hormones present in the blood have been developed only recently, and no systematic studies of age differences in blood levels of these hormones have been reported.

The pancreas secretes insulin, the hormone that regulates the utilization of sugar and other nutrients in the body. When the pancreas fails to produce adequate amounts of insulin, diabetes occurs. One test for diabetes involves measuring the rate of removal of sugar from the blood, that is, the glucose-tolerance test. One characteristic of aging is a reduction in the rate of removal of excess sugar from the blood. At present it is not known whether this represents the early stages of diabetes or whether it is a normal age change. It does appear in aged individuals who do not show any of the other symptoms of diabetes. Furthermore, it has been shown that, unlike the diabetic, elderly subjects can, with additional stimulation, produce more insulin. In normal young persons the pancreas releases more insulin in response to even a slight rise in blood sugar levels. In the elderly, the sensitivity of the pancreas is reduced so that a higher level of blood sugar is required to stimulate it to action. With maximum stimulation the pancreas in the aged can produce as much insulin as the pancreas in the young.

Sexual activity

It has long been known that the excretion of both male and female sex hormones diminishes with age. In the female, the excretion of estrogens (female sex hormones) falls markedly at the menopause (see MENOPAUSE). In the male, the excretion of androgens (male sex hormones and their degradation products) falls gradually over the age span 50–90, so that the existence of a male "climacteric" is highly improbable.

Sexual activity, as reported in interview studies, diminishes progressively between the ages of 20 and 60 in both males and females. In males the frequency of marital intercourse falls from an average of four per week in 20-year-olds to one per week in 60-year-olds. Practically all males aged 20–45 reported some level of sexual activity. Between the ages of 45 and 60 only about 5 percent of males reported loss of sexual activity.

Few systematic studies have been made of sexual behaviour in individuals over the age of 60, but clinical reports indicate that at least some males remain sexually active at 90.

There are wide individual differences in the level of sexual activity in both males and females. In human beings, sexual behaviour is influenced more by psychological and social factors than by the levels of sex hormones circulating in the blood. Nevertheless, the use of male sex hormones has had a long, and stormy, history as a rejuvenating agent for males. Attempts to rejuvenate elderly males by injecting crude extracts from testicles of animals, as well as various androgenic compounds, were made, but the effects, if any, were only transitory. In the early 1900s, sex glands from other animals were transplanted into human beings, but the results were questionable and the side effects were often disastrous. At about the same time, an operation was devised in which the spermatic ducts were tied off. It was assumed that preventing the loss of sperm would stimulate the sex glands to produce androgenic hormones which would rejuvenate the individual. None of these assumptions proved correct, so that the operation was soon abandoned as a rejuvenating procedure.

Since tissue loss does occur with aging, the administration of anabolic hormones (hormones that promote the buildup of tissues) may represent an important future development. The compounds that are currently available have a number of undesirable side effects and cannot be used routinely. Chemists and pharmacologists continue research to produce new steroids that will have anabolic effects without the undesirable side effects.

Skeletal system. With aging, the bones gradually lose calcium. As a result they become more fragile and are more likely to break, even with minor falls. Healing of fractures is also slower in the old than in the young. Recent advances in orthopedic surgery, with the replacement of parts of a broken bone or joint with new structures or the introduction of metallic pegs to hold broken parts together, have been of great value to elderly people.

The incidence of osteoporosis, a disease characterized by a loss of calcium and minerals from bone, also increases with age. It occurs more frequently in women after menopause than in men and is especially evident in the spinal column. Back pain is a primary symptom of the disease. It can be treated by increasing calcium intake in association with the administration of anabolic hormones.

The mobility of joints diminishes with age and the incidence of arthritis increases (see JOINT DISEASES AND INJURIES).

Respiratory system. Vital capacity, or the total amount of air that can be expelled from the lung after a maximum inspiration, diminishes with age, as does the total volume of air that can be contained in the lungs. In contrast, the amount of air that cannot be expelled from the lung increases. These changes in respiratory mechanisms are primarily a reflection of the increased stiffness of the bony cage of the chest and decreased strength of the muscles that move the chest during respiration.

The lung also contains elastin and collagen to give it elastic properties. As indicated previously, the formation of cross-links in elastin and collagen that takes place with aging reduces the elastic properties of the lung.

The transfer of oxygen and carbon dioxide from the air

in the lungs to the blood is influenced by the amount of blood flowing through the lungs as well as by the amount of air moved in and out. The characteristics of the membranes that separate blood and air in the lungs are also important in maintaining an adequate supply of oxygen to the body. Although with age there is a slight reduction in the amount of oxygen that can be moved from the air to the blood in the lungs, the reduction becomes apparent only when large amounts of oxygen are required, as during strenuous exercise. It is believed that a primary factor in the impairment of oxygen transfer in the lungs of elderly subjects is the lack of appropriate adjustment of the blood flow to the air sacs in the lung.

Incidence of emphysema

Emphysema, abnormal distension of the lungs with air, is a lung disease reaching its highest incidence between the ages of 45 and 65. In the United States the death rate from emphysema increased by almost 400 percent between 1950 and 1960. Although the exact causes of the disease are still unknown, the presence of noxious or toxic agents in the air may be a contributing factor. Many studies have shown a relationship between the incidence of emphysema and bronchitis (inflammation of the bronchi) and smoking. Among British physicians death rates from bronchitis were six times higher in those smoking 25 cigarettes a day than in nonsmokers.

Measurements of lung function are significantly lower in cigarette smokers than in nonsmokers of the same age. Values for cigarette smokers are, on the average, about equal to those of nonsmokers who are 10–15 years older. There is evidence, however, that when cigarette smokers quit smoking, measurements of pulmonary function closely approach those of nonsmokers within one to two years, even in the case of heavy smokers 50–60 years old.

Kidney. The kidney removes wastes from the body by separating them from the blood and forming urine. In this process many substances are accumulated in the urine at a higher concentration than in the blood. With advancing age the concentrating ability of the kidney falls, so that a greater volume of water is required to excrete the same amount of waste material. This loss in concentrating ability is probably partially offset by a decrease in the excretory load because of reduced activity, alterations in food intake, and the reduction in muscle mass of the elderly. The changes in kidney function that have been described may not be reflected in urine volume, since volumes fluctuate widely at all ages and are determined primarily by fluid intake.

The reduction in renal (kidney) function is due in part to a gradual reduction in blood flow to the kidney. Since the kidney receives a great excess of blood (about 25 percent of the blood pumped by the heart each minute), the reduction with age does not normally result in an accumulation of waste products in the blood. Any such accumulation is the result of disease that damages the kidney.

The reduced concentrating ability of the kidney results from a loss of some of the nephrons, the functional elements of the kidney, and the reduced activity of cellular enzymes.

Regulatory mechanisms. Some physiological characteristics, such as the mechanisms that regulate the acidity of the blood or its sugar level, are adequate to maintain normal levels under resting conditions even in very old people; however, the aged require more time than the young to re-establish normal levels when changes from the normal occur.

In order to test the effectiveness of control mechanisms of the body, physiologists produce changes experimentally and determine the rate of recovery. When the acidity of the blood is increased to the same extent in old and young subjects, it is returned to normal within 6–8 hours in the young; in the elderly 18–24 hours are required.

Similarly, the rate of return to fasting levels after sugar has been administered intravenously or orally is slower in the old than in the young. The response to insulin, which accelerates the removal of sugar from the blood, is also diminished in the elderly.

Adjustments to changes in environmental temperature are less adequate in the old than in the young. Consequently older people may prefer more uniform and slightly higher temperatures for comfort than the young. High temperatures are also more hazardous to the elderly. The incidence of heat prostration in hot weather increases with age.

Exercise and work

Exercise is one of the physiological stresses of daily living. In reasonable amounts it is a valuable stimulus to maintain physiological vigour. A number of studies have indicated a lower incidence of cardiovascular disease among adults who indulge in physical activity than in those who do not.

The capacity to perform muscular work diminishes progressively in the elderly. Muscle strength diminishes; however, the reduction in strength is less in muscles that continue to be used throughout adult life than in those that are not. Thus a part of the reduction in muscle strength may be an atrophy of disuse.

Maximum work capacity is reduced in the elderly, largely because of the inability to deliver enough oxygen to the working muscles. In the young, the need for oxygen is met for the most part by increasing the heart rate. Under conditions of maximum work, young adults can increase their heart rate to over 200 beats per minute; the elderly to only about 150 per minute. In addition, the transfer of oxygen from the lungs to the blood is reduced in the elderly under conditions of strenuous exercise.

With less than maximum exercise, there is a greater increase in blood pressure, heart rate, and respiration in the old than in the young; that is, a given work load induces a greater physiological stress in the old than in the young. Furthermore, recovery of blood pressure, heart rate, and respiration to resting values takes longer in the old.

Premature aging. Progeria is an extremely rare disease of early childhood characterized by many of the superficial aspects of aging, such as baldness, thinning of the skin, prominence of blood vessels of the scalp, and vascular disease. These children have the general appearance of "little old men." They rarely live beyond the age of 15–18. Death is usually caused by cardiovascular disease.

The disease is extremely rare. In fact, only about 50 cases have been identified for study. In spite of the appearance of premature aging, these patients fail to show an acceleration of other age changes, except for the early development of cardiovascular disease. Most tests of physiological and psychological functions give values which are normal for their chronological age. It is doubtful whether the child with progeria is suffering only from accelerated aging; rather, progeria should be regarded as a rare disease with a superficial resemblance to senescence.

SUMMARY

Aging is a normal part of the life-span. It is associated with a gradual reduction in the reserve capacities of most of the organ systems of the body. There are, however, wide individual differences in aging. Aging affects different organ systems, even in the same individual, at different rates. For example, some features, such as fasting blood sugar levels, do not change with age. Speed of conduction of a nerve impulse diminishes only about 10 percent over the age span 30–90, whereas the amount of blood pumped by the heart diminishes by 40 percent. Performances such as severe exercise, which require the integrated activity of a number of organ systems, show the greatest decrements with age. Many of these decrements have only a minimal influence on daily life. They become important when the individual is called upon to adjust to stresses. Consequently, diseases, especially those that develop slowly, such as arteriosclerosis (hardening of the arteries), heart disease, and cancer, take an increasing toll with advancing age.

Aging in the human being results from the complex interplay of physiological, psychological, and social factors. The adaptability of the human organism makes it possible to compensate for many of the physiological im-

pairments of aging. Other limitations of the elderly can be remedied by changes in social customs, so that much can be done to improve the status of the aged.

The basic biological causes of aging are still unknown, but it is highly probable that changes at the cellular level play an important role.

BIBLIOGRAPHY. Additional information on this subject may be found in the following texts: J.E. BIRREN (ed.), *Handbook of Aging and the Individual* (1960); A. COMFORT, *Ageing: The Biology of Senescence*, rev. ed. (1964); H.J. CURTIS, *Biological Mechanisms of Aging* (1966); A.M. HOFFMAN (ed.), *The Daily Needs and Interests of Older People* (1970); E.V. COWDRY, *Problems of Ageing: Biological and Medical Aspects*, 3rd ed. by A.I. LANSING (1952); R.W. PREHODA, *Extended Youth: The Promise of Gerontology* (1968); M.W. RILEY and A. FONER, *Aging and Society*, vol. 1, *An Inventory of Research Findings* (1968); N.W. SHOCK, *Trends in Gerontology*, 2nd ed. (1957); B.L. STREHLER, *Time, Cells, and Aging* (1962); and C. TIBBITTS and W. DONAHUE (eds.), *Aging in Today's Society* (1960).

(N.W.S.)

Agnatha

The class Agnatha (phylum Chordata) comprises the jawless, fishlike hagfishes and lampreys and several fossil groups. Hagfishes are minor pests of commercial food fisheries of the North Atlantic, and lampreys have been a serious pest of food fisheries in the Great Lakes in North America. Agnathans are otherwise of little economic importance. The group is of great evolutionary interest, however, because it includes the oldest known craniate fossils and because the living agnathans have many primitive characteristics.

General features. The body of the hagfish is soft-skinned and nearly cylindrical, with a single nostril at the anterior end, overlying the mouth, and a low caudal fin around the tail. The eyes are vestigial and covered by skin. All of the 20 known species are restricted to cold, marine bottom waters at depths ranging from 10 metres (about 33 feet) in high latitudes to 1,300 metres in equatorial oceans. Adults are 40 to 80 centimetres (15 to 30 inches) long. All species are superficially similar except in the number and position of the gill apertures.

Lampreys

Lampreys, which number about 22 species, occur in cool, fresh, and coastal waters of all continents except Africa. All species are rather similar. The body is smooth and eel-shaped, with well-developed dorsal and caudal fins; the mouth is surrounded by a suctorial oral disk bearing horny teeth. The eyes are well developed, and the nostril is on the top of the head. Adults range from 15 to 100 centimetres long.

Although the gonad of a hagfish usually includes both ovary and testis, there is no evidence either of hermaphroditism (the reproductive organs of both sexes functioning in the same individual) or of self-fertilization. A female produces a small number of tough-skinned, yolk-filled eggs about two centimetres long that hatch into miniature adults.

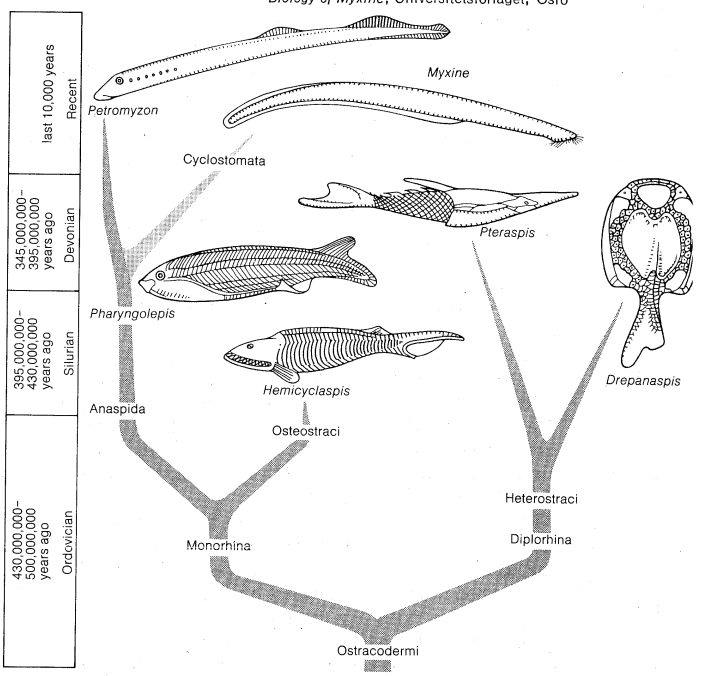
Hagfishes locate their food by scent. Although some are known to eat fishes immobilized in nets, the best studied species, *Myxine glutinosa*, normally feeds on soft-bodied invertebrates and larger dead animals. *Myxine* burrows into soft marine sediments and rests with only the tip of the head protruding. During respiration, water enters through the nostril and passes by a nasopharyngeal duct to the pharynx and gills. When stimulated by the scent of a dead fish, *Myxine* leaves its burrow and swims against the current. On making contact with the fish, it coils around it and bites into it by protruding and retracting the comblike horny tooth plates on the floor of the mouth. Along each side of the body is a row of prominent glands that produce a distasteful gelatinous slime when the animal is disturbed.

Lampreys mate in a nestlike depression excavated by the male in the gravel bed of a stream; the numerous eggs, about one millimetre (.04 inch) in diameter, lodge in the gravel around the nest. The egg hatches into an ammocoete larva—a blind, wormlike animal that burrows in silt. The larva's mouth is overhung by a hoodlike upper lip that protrudes above the surface of the silt. A continuous stream of water passes in through the mouth and out through the seven pairs of gills. Microscopic plants, the food of the ammocoete, are filtered from the respiratory current by strands of mucus produced by the endostyle, a gland in the floor of the pharynx.

After about three years, when the ammocoete has grown to about ten centimetres (four inches), it undergoes a radical metamorphosis. The eyes complete their development; the upper lip becomes transformed into a suctorial oral disk; the endostyle changes into a thyroid gland; and the fins along the back increase in height. On completion of metamorphosis, a typical lamprey such as *Petromyzon marinus* migrates to the sea, where it feeds by attaching itself with its sucker to bony fishes. It rasps into the flesh with a toothed, tonguelike structure on the floor of the mouth. Saliva containing an anticoagulant facilitates the ingestion of blood and muscle tissue. On attaining full adult size, the lamprey ceases to feed, migrates upstream to a spawning ground, mates, and dies.

Members of several lamprey species do not migrate to sea but feed in freshwater. *Petromyzon marinus dorsatus* once seriously affected commercial fishing in the Great Lakes. The brook lampreys do not feed after metamorphosis but mature sexually and reproduce.

From A. Brodal and R. Fange (ed.), *The Biology of Myxine*; Universitetsforlaget, Oslo



Systematic division of fossil and recent Agnatha.

Form and function. The elongated bodies of hagfishes and lampreys are adaptations to a burrowing habit—throughout life in the hagfish, during the larval period in the lamprey. Considerable variation in body form occurs among extinct agnathans. Some Osteostraci, for example, were flattened and, although possessing a powerful swimming tail, appear to have been bottom-dwellers. The size and shape of the mouth suggest that they were filter feeders. The laterally compressed, fishlike form of the anaspids (*e.g.*, *Pharyngolepis*) indicates a free-swimming habit; the absence of fossilized impressions of the mouth region suggests the presence of large suctorial or grasping lips. The extinct heterostracans include obvious bottom-dwellers (*e.g.*, *Drepanaspis*) and others (*e.g.*, *Pteraspis*) that were apparently adapted to mid-water, or nektonic, life. For example, some heterostracans had movable enamel plates inside the lower lip, probably to provide a biting or grazing mechanism. Osteostracans, anaspids, hagfishes, and lampreys (Monorhina) have one median nostril, but heterostracans (Diplorhina) appear to have had two, one at each corner of the mouth.

Variation in body form

Evolution. If evidence from fossil and living forms is combined, the Agnatha are distinguishable from the oth-

er craniates (Gnathostomata) by what they lack: jaws, lateral fins supported by fin rays, vertebrae, a horizontal semicircular canal in the ear, and genital ducts. The nervous, sensory, endocrine, circulatory, excretory, and muscular systems have the same basic structure as those of gnathostomes but are generally simpler. The presence in the ammocoete larva of an endostyle, a gland that otherwise is found only in protochordates (*e.g.*, amphioxus, tunicates), suggests that the Agnatha represent an evolutionary level intermediate between the protochordates and gnathostomes; but the degree of specialization of known agnathans (particularly the single nostril of most forms) rules out the possibility that they are ancestral to the gnathostomes.

Annotated classification. In the classification below, the groups indicated by a dagger (†) are extinct and known only from fossils.

CLASS AGNATHA

Craniate chordates, lacking jaws; Silurian Period (395,000,000–430,000,000 years ago) to Recent.

Subclass Monorhina

With 1 nostril.

†*Order Osteostraci*

Late Silurian to Late Devonian (345,000,000–395,000,000 years ago); head and gills enclosed in heavy bony shield; nostril between eyes; gill openings ventral; about 7 families.

†*Order Anaspida*

Late Silurian to Late Devonian; active swimmers with tail bent downward; nostril between eyes; gill openings lateral; about 5 families.

Order Cyclostomata (lampreys and hagfishes)

Scales absent, skeleton cartilaginous, horny teeth in mouth, body eel-shaped.

Suborder Myxini (slime eels and hagfishes). Recent; nostril at anterior tip; 5–15 pairs of gills; 2 rows of horny teeth; eyes vestigial.

Family Eptatretidae. Five to 15 gills opening separately to exterior; all oceans except North Atlantic; 2 genera, 13 species.

Family Myxinidae. Five to 7 gills opening by common apertures on each side; all oceans; 4 genera, 8 species.

Suborder Petromyzones (lampreys). Pennsylvanian (280,000,000–325,000,000 years ago) to Recent; dorsal nostril, 7 pairs of gills, horny teeth on oral disk; eyes large, 1 or 2 dorsal fins.

Family Petromyzonidae. Single horny toothplate above the mouth bearing pointed or rounded teeth; 5 genera, 17 species; Eurasia and North America.

Family Mordaciidae. Two tricuspid teeth above mouth; 1 genus, 4 species; eastern Australia and western South America.

Family Geotriidae. Single horny toothplate above the mouth, bearing 4 spatulate teeth; 1 genus, 1 species; southern Australia, New Zealand, South America.

Subclass Diplorhina

With 2 nostrils.

†*Order Heterostraci*

Ordovician to Upper Devonian (345,000,000 to 500,000,000 years ago); usually heavily armoured; gills with common aperture on each side; length 5–30 cm; about 12 families.

†*Order Coelolepida*

Upper Silurian and Lower Devonian (395,000,000 years ago); a little-known group of unknown affinities; length to about 10 cm; 1 family.

Critical appraisal. Although lampreys have many features in common with the Osteostraci and Anaspida, general disagreement prevails on the classification of the hagfishes. Their affinity with lampreys is questionable, and they should perhaps be removed from the Cyclostomata as a separate order. It has been suggested that the hagfishes are closely related to the heterostracans, but this view is not widely accepted.

BIBLIOGRAPHY. V.C. APPLEGATE, *Natural History of the Sea Lamprey, Petromyzon marinus, in Michigan*, Spec. Scient. Rep. U.S. Fish Wildl. Serv. 55 (1950), an account of lamprey life history and the pest problem; A. BRODAL and R. FANGE (eds.), *The Biology of Myxine* (1963), an authoritative work by 24 authors; M. FONTAINE, "Formes actuelles de cyclostomes," in P.P. GRASSE (ed.), *Traité de zoologie*, vol. 13, pp. 1–172 (1958), the best general account of cyclostomes (well illustrated); M.W. HARDISTY and I.C. POTTER (eds.), *The Biology of Lampreys* (1972), a comprehensive reference; A.J. MARSHALL, "Agnatha," in T.J. PARKER and W.A. HASWELL, *Textbook of Zoology*, 7th ed., vol. 1 (1962), a good account of comparative anatomy, written for university students; E. STENSIO, "Les Cyclostomes fossiles ou ostracodermes," in J. PIVETEAU (ed.), *Traité de paléontologie*, vol. 4, pp. 96–382 (1964); R. STRAHAN, "The Behaviour of Myxinoids," *Acta Zool. Stockh.*, 44:73–102 (1963), the most recent accounts of the subject.

(Ro.S.)

Agnosticism

The word agnosticism was first publicly coined in 1869 at a meeting of the Metaphysical Society in London by T.H. Huxley, a British biologist and champion of the Darwinian theory of evolution. He coined it as a suitable label for his own position. "It came into my head as suggestively antithetical to the 'Gnostic' of Church history who professed to know so much about the very things of which I was ignorant."

Nature and kinds of agnosticism. Huxley's statement brings out both the fact that agnosticism has something to do with not knowing, and that this not knowing refers particularly to the sphere of religious doctrine. Etymology, however, and now common usage, do permit less limited uses of the term. The Soviet leader Lenin, for instance, in his *Materialism and Empirio-Criticism* (1908), distinguished the extremes of true Materialism on the one hand and the bold Idealism of George Berkeley, an 18th-century Idealist, on the other. He recognized as attempted halfway houses between them the "agnosticisms" of the Scottish Skeptic David Hume and the great German critical philosopher Immanuel Kant—agnosticisms that here consisted in their contentions about the unknowability of the nature, or even the existence, of "things-in-themselves" (realities beyond appearances).

Huxley's nonreligious agnosticism. The essence of Huxley's agnosticism—and his statement, as the inventor of the term, must be peculiarly authoritative—was not a profession of total ignorance, nor even of total ignorance within one special but very large sphere; rather, he insisted, it was "not a creed but a method, the essence of which lies in the rigorous application of a single principle," viz., to follow reason "as far as it can take you"; but then, when you have established as much as you can, frankly and honestly to recognize the limits of your knowledge. It is the same principle as that later proclaimed in an essay on "The Ethics of Belief" (1876) by the British mathematician and philosopher of science W.K. Clifford: "It is wrong always, everywhere and for everyone to believe anything upon insufficient evidence." Applied by Huxley to fundamental Christian claims, this principle yields characteristically skeptical conclusions: speaking, for example, of the Apocrypha (ancient scriptural writings excluded from the biblical canon), he wrote: "One may suspect that a little more critical discrimination would have enlarged the Apocrypha not inconsiderably." In the same spirit, Sir Leslie Stephen, 19th-century literary critic and historian of thought, in *An Agnostic's Apology, and Other Essays* (1893), reproached those who pretended to delineate "the nature of God Almighty with an accuracy from which modest naturalists would shrink in describing the genesis of a black beetle."

Contrasts with atheism, Skepticism, Positivism

Agnosticism in its primary reference is commonly contrasted with atheism thus: "The Atheist asserts that there is no God, whereas the Agnostic maintains only that he does not know." This distinction, however, is in two respects misleading: first, Huxley himself certainly rejected as outright false—rather than as not known to be true or false—many widely popular views about God, his providence, and man's posthumous destiny; and second, if this were the crucial distinction, agnosticism would for almost all practical purposes be the same as atheism. It was indeed on this misunderstanding that Huxley and his associates were attacked both by enthusiastic Christian polemicists and by Friedrich Engels, the co-worker of Karl Marx, as "shame-faced atheists," a description that is perfectly applicable to many of those who nowadays adopt the more comfortable label.

Agnosticism, moreover, is not the same as Skepticism (*q.v.*), which, in the comprehensive and classical form

epitomized by the ancient Greek Skeptic Sextus Empiricus (2nd and 3rd centuries AD), confidently challenges not merely religious or metaphysical knowledge but all knowledge claims that venture beyond immediate experience. Agnosticism is, as Skepticism surely could not be, compatible with the approach of Positivism, which emphasizes the achievements and possibilities of natural and social science—though most agnostics, including Huxley, have nonetheless harboured reserves about the more authoritarian and eccentric features of the system of Auguste Comte, the 19th-century founder of Positivism (see POSITIVISM AND LOGICAL EMPIRICISM).

Religious agnosticism. It is also possible to speak of a religious agnosticism. But if this expression is not to be contradictory, it has to be taken to refer to an acceptance of the agnostic principle combined either with a conviction that at least some minimum of affirmative doctrine can be established on adequate grounds, or else with the sort of religion or religiousness that makes no very substantial or disputatious doctrinal demands. If these two varieties of agnosticism be admitted, then Huxley's original agnosticism may be marked off from the latter as (not religious but) secular and from the former as (not religious but) atheist—construing "atheist" here as a word as wholly negative and neutral as "atypical" or "asymmetrical." These, without pejorative insinuations, mean merely "not typical" or "not symmetrical" (the atheist is thus one who is simply without a belief in God).

Huxley himself allowed for the possibility of an agnosticism that was in these senses religious—even Christian—as opposed to atheist. Thus, in another 1889 essay "Agnosticism and Christianity," he contrasted "scientific theology," with which "agnosticism has no quarrel," with "Ecclesiasticism, or, as our neighbours across the Channel call it, Clericalism"; and his complaint against the latter's proponents was not that they reach substantive conclusions different from his own but that they maintain "that it is morally wrong not to believe certain propositions, whatever the results of strict scientific investigation of the evidence of these propositions." The second possibility, that of an agnosticism that is religious as opposed to secular, was realized perhaps most strikingly in the Buddha (Gautama). Typically and traditionally, the ecclesiastical Christian has insisted that absolute certainty about some minimum approved list of propositions concerning God and the general divine scheme of things was wholly necessary to salvation. Equally typically, according to the tradition, the Buddha sidestepped all such speculative questions. At best they could only distract attention from the urgent business of salvation—salvation, of course, in his own very different interpretation.

Historical antecedents of modern agnosticism. It is convenient to distinguish the antecedents of secular agnosticism from those of religious agnosticism.

Antecedents of secular agnosticism. The ancestry of modern secular and atheist agnosticism may be traced back to the Sophists and to Socrates in the 5th century BC; not, of course, the "Socrates" of Plato's *Republic*—the would-be founding father of an ideal totalitarian state—but the shadowy historical Socrates supposedly hailed by the oracle of Apollo's Delphi as the wisest of men—who knew what, and how much, he did not know. But the most important and immediate source of such agnostic ideas was surely Hume, while Hume's successor Kant—the second of the two great philosophers of the Age of Enlightenment—may well be seen as the prime philosophical inspirer of religious reactions against them.

Huxley, as noted above, demanded that a thinker recognize and accept the limits of his knowledge. In taking it that these limits do not include either the findings of a general positive natural theology or the contents of a particular special divine revelation, Huxley was accepting a Humean critique. (It is significant that Huxley's study of Hume was the most sympathetic appraisal to be published in the 19th century.) Hume's critique is found in his *Enquiry Concerning Human Understanding* (first published in 1748 under another title), which attempts, in the manner of Locke and later Kant, to determine the limits of man's possible knowledge, and in his posthumous *Dialogues Concerning Natural Religion* (1779).

The work of David Hume

Two sections of the *Enquiry* refer directly to these limits: "Of a Particular Providence and of a Future State" and "Of Miracles." In the first, Hume starts from his basic Empiricist claims: that, generally, "matters of fact and real existence" cannot be known a priori (prior to and apart from experience); and that, particularly, one cannot know a priori that any thing or kind of thing either must be or cannot be the cause of any other thing or kind of thing. These considerations dispose of all the classical arguments for the existence of God other than the argument from design—that the structure and order of the universe and its constituents implies a design and a designer. But here, Hume urges, argument from experience can find no purchase because both the supposed effect, the universe as a whole, and the putative cause, God, are essentially unique and incomparable. Later, in his *Dialogues*, he develops the suggestion—which he acknowledges as stemming from the 3rd-century-BC philosopher Straton of Lampsacus, next but one after Aristotle as head of his Lyceum—that whatever order man discerns in the universe should be attributed to the universe itself and not to any postulated outside cause.

In the section "Of Miracles," Hume takes his stand on the agnostic principle: "A wise man . . . proportions his belief to the evidence." He then argues that no attempt to appeal to the alleged occurrence of miracles—conceived as authoritative endorsements by a power beyond and greater than nature—can succeed in establishing the truth of a claim to constitute special divine revelation. Hume's distinctive contribution here is methodological: the contention that the principles and presuppositions upon which the critical historian must rely, in first interpreting the remains of the past as historical evidence and in then building up from this evidence his account of what actually happened, are such as to make it impossible for him "to prove a miracle and make it a just foundation for any such system of religion."

In this two-phase attack, Hume challenged what was in his day, and long remained, the standard framework for systematic Christian apologetics. Indeed, the contrary contentions—of the possibilities, both of developing a positive natural theology and of establishing the authenticity of a supposed revelation by discovering endorsing miracles—were defined as essential and constitutive dogmas of Roman Catholicism by decrees of the first Vatican Council of 1869–70.

In view of the future history of Western thought, it must be emphasized that Hume's position, like Kant's, was (officially) that knowledge in this area is practically impossible. This thesis is stronger than that of those who simply confess that they just do not know:

> The God-men say when die go sky
> Through pearly gates where river flow,
> The God-men say when die we fly
> Just like eagle, hawk and crow—
> Might be, might be; I don't know.
> (Aboriginal song from the Northern Territory, Australia.)

Yet Hume's thesis was, on the other hand, weaker than that of his 20th-century neo-Humean successors, the logical positivists of the Vienna Circle, who held that any talk about a transcendent God must be "without literal significance." This view was presented brilliantly, and in an uncompromisingly drastic form, by A.J. Ayer in his *Language, Truth and Logic* (2nd ed., 1946). Similar conclusions were reached less high-handedly by several contributors to *New Essays in Philosophical Theology* (ed. by A. Flew and A. MacIntyre, 1955).

Antecedents of religious agnosticism. Looking backward, it is possible now to see what Hume himself did not know—that his attack on the possibility of a positive natural theology had to a considerable extent been anticipated by 14th-century Christian Scholastics: generally, by William of Ockham; and, with particular reference to the lack of a priori knowledge of causal relations, by Nicholas of Autrecourt.

The claims of Hume and Kant—and, indeed, those of the logical positivists and their successors—about the

Negative theology

practical, or theoretical, impossibility of such knowledge should also be compared with the long traditions of "negative theology." Such a theology maintains that the nature of God passes so far beyond the comprehension of any creature that God must be characterized largely or entirely by indirection—as Infinite, as Incomparable, and so on. Thus Thomas Aquinas, the foremost Scholastic of the 13th century—who contrived on other occasions to tell his readers as much as his most practical church could wish about the deeds, plans, and demands of the Ineffable—nevertheless had his agnostic moments as well. But he did elaborate a doctrine of so-called analogical predication designed to show how it is possible for finite creatures to say and to understand something positive about God by means of comparisons with known entities or qualities. By contrast, the 12th-century philosopher Moses Maimonides, often dubbed anachronistically "the Jewish Aquinas," had been much more drastic than his successor, "the Christian Maimonides," in his insistence that everything that can be truly said about the Creator—not excluding the proposition that he exists—has to be construed as purely negative.

Tension between the religious and agnostic

Although it is clearly possible to speak of a religious agnosticism without self-contradiction, the foregoing considerations suggest the difficulty of intermingling religious and agnostic concerns. The easiest case is that in which the religion is altogether without metaphysical content: thus, one of Huxley's biographers reports that the 19th-century Scottish sage Thomas Carlyle "taught him that a deep sense of religion was compatible with an entire absence of theology." The next simplest case is that in which worship is combined with a total noncommitment about the attributes of the object of worship:

> He is not a male: He is not a female: He is not a neuter.
> He is not to be seen: He neither is nor is not.
> When He is sought He will take the form in which He is sought.
> It is indeed difficult to describe the name of the Lord.
> (Poem from the Telugu, inscribed on a cult object in the Royal Ontario Museum.)

In its original setting this expression of a Hindu piety has power and charm. Yet its intellectual inadequacy becomes manifest when the doctrine of the Unknowable in the broad synthetic system of Herbert Spencer, a late-19th-century evolutionary philosopher, is recalled. For to affirm, as Spencer did, the existence of a being about whom absolutely nothing else can be said is a rather comical hypostatization (taking of an abstraction as real), which is surely indiscernible from affirming no being at all. Nor, perhaps, is it any great improvement to aver that much else can indeed be said about him, but only in words that here must bear an extraordinary meaning—unless, of course, those meanings can be specified. It was the suggestion that the goodness of God might thus be goodness in a quite unusual sense—what would elsewhere be called badness—that provoked the ire of John Stuart Mill, a mid-19th-century Empiricist, against certain developments from Sir William Hamilton's "Philosophy of the Unconditioned." Mill wrote: "I will call no being good, who is not what I mean when I apply that epithet to my fellow creatures."

The third, and surely the most promising, way in which the reconciliation may be attempted is by essaying some distinction between the essence or the internal nature of God and his external relations with the creation. It may then be suggested that, whereas man's knowledge of the former must be at least exiguous and at worst simply lacking, he can nevertheless know as much as he needs to know about the latter. As to the rest, he should be reverently agnostic.

Incompatibility with fideism. What cannot, however, by any means be squared with agnosticism in Huxley's sense are attempts to transmute the very limitations of human knowledge into grounds for accepting some wholly unevidenced faith. Such transmutations have been made in the interests of many mutually irreconcilable systems, and they apparently remain perennially attractive to thinkers with a different understanding of the ethics of belief.

St. Augustine of Hippo, near the end of the 5th century, felt the challenge of classical Skepticism in Cicero's *Academica* and *De natura deorum* ("On the Nature of the Gods") and gave his response in *Contra academicos* ("Against the Academics"). Skepticism, he thought, can be overcome only by revelation. The orthodox Muslim philosopher and mystic al-Ghazālī (late 11th century) deployed Skeptical arguments similarly, as a propaedeutic, or study preparatory to the acceptance of his rival revelation. With the rediscovery in the 16th century of the works of Sextus Empiricus, a course of Skepticism became commonly a preliminary to fideist commitment. Fideism is the thesis that truth in religion is accessible only to faith. The course persuaded the inquirer that reason cannot attain truth; yet certainty in true religious belief was still thought absolutely necessary for salvation. Martin Luther was speaking for his times (first half of the 16th century) when he thundered against the extremely cautious and restricted agnosticism of Desiderius Erasmus, foremost figure of the northern Renaissance: "Spiritus sanctus non est Skepticus" ("The Holy Spirit is not a Skeptic").

The only resort was, it seemed, faith: whether the easy-going Roman Catholic faith of the 16th-century Skeptic Michel de Montaigne; the polemical Counter-Reformation fervour of his contemporary Gentian Hervet, veteran of the Council of Trent and Latin translator of the *Adversus mathematicos* (1569; "Against the Pundits") of Sextus Empiricus; or, one century later, the vestigial Huguenot loyalty of Pierre Bayle—stocker of a great arsenal of secular argument, the *Dictionnaire historique et critique* (1695–97).

The decisive objection to any and every such rationally unfounded flight into faith was posed by John Locke, the 17th-century British Empiricist, who set a tone of coolly unfervent Anglicanism for the following century:

> We may as well doubt of our being, as we can whether any revelation from God be true. So that faith is a settled and sure principle of assent and assurance, and leaves no room for doubt or hesitation. Only we must be sure that it be a divine revelation, and that we understand it right: else we shall expose ourselves to all the extravagancy of enthusiasm, and all the error of wrong principles . . . (*An Essay Concerning Human Understanding*, Book IV, ch. xvi, 14).

Many thinkers have agreed that it is all very well to depreciate the potentialities of unaided natural reason and to insist that if man is to have any knowledge of God this must depend largely or wholly upon whatever special steps God may have taken to reveal himself; and they have also agreed that, if man's commitment of faith is not to be arbitrary and frivolous, then he clearly must have some good reason for believing, first, that there is a God who has so revealed himself, and, second, that his preferred candidate—and not one of its innumerable rivals—truly is that revelation.

These points are crucial—both for the appreciation of the history of ideas and for a reasonable contemporary understanding. Clearly, they were upheld by Aquinas, who in the *Summa contra gentiles*—before proceeding to present his own reasons for accepting Christianity, rather than Islām, as the authentic revelation—applied that same word frivolous to any such unsupportable commitment. Again, Judah ha-Levi, an early-12th-century Jewish poet and philosopher, has been authoritatively described as "concerned to bring men to a mystical and non-rational appreciation of religious truths" by his Skeptical attacks on the established Aristotelian natural theology. Yet ha-Levi's main work, entitled *Kuzari: The Book of Proof and Argument in Defence of the Despised Faith*, does in fact offer rational evidences of the truth of Judaism.

Skeptical propaedeutics to faith are now out of fashion. But the same challenge applies to all of the various responses to Kant's famous invitation: "I have found it necessary to deny *knowledge* in order to make room for *faith*" (Preface to the *Critique of Pure Reason*). Natural theology may, indeed, for Hume's reasons as reinforced by Kant, be impossible. The way of religious discovery may indeed be mystical experience, personal encounter

with the divine Thou, or whatever else. But there is, and can be, no substitute for a man's having some sound grounds for identifying his experience not only as really mystical but also as experience of the real God; for holding his faith in some putative revelation not only to be real religious faith but also to be faith in a genuine revelation of the Real; and so on.

Rejections of the agnostic principle. Anyone who insists on the foregoing touchstones may still be agnostic as well as religious. What cannot consist with agnosticism is a calculated commitment to faith seen as altogether without evidential warrant. The classic example of such commitment was provided in the 17th century by the Wager Argument of the French mathematician Blaise Pascal, who assumed, for the sake of the argument, that "reason can decide nothing here" and then urged that the only sane bet is Roman Catholicism; for we have nothing but this one short life to lose, and all eternity to win.

Blaise Pascal and William James

Pascal's Wager Argument is unsound because, on its own stated assumption of total and inescapable ignorance, the gambler is not entitled to limit the betting options to two—and to one particular two, at that. A similarly parochial inattention to the variety of candidacies for belief has characterized most fideists. Thus Søren Kierkegaard, an influential mid-19th-century Danish lay theologian, happily glorified the essential irrationality of religious faith, while taking it always that faith will, of course, be Protestant. Elsewhere, Pascal himself did notice, and tried to meet, some of the competition; his neglect here is the more remarkable because his wager was originally imported into Christendom from Islām (see Miguel Palacios, *Los precedentes de Pari de Pascal*). What makes it a landmark is that it constituted a direct, reasoned rejection of the agnostic principle—a rejection in which the reason proposed for believing was explicitly a motive for self-persuasion rather than some evidence of truth. Thus, when William James, a pre-World War I American psychologist and philosopher, in *The Will to Believe*, developed the best known systematic attack on that principle it was, rightly, Pascal whom he hailed as his first inspiration. James distinguished those hypotheses that, for any individual, represent psychologically "live options" from those that do not, and he urged that, when evidential grounds are lacking, the choice may properly be determined by one's passional nature. For men often have to act on some unproved hypothesis, and sometimes such firm commitments may help to make the belief come true. Consider, for example, some belief that a man is trustworthy. The objections are that belief in the existence of God is clearly not of this case, and generally that to act decisively on some hypothesis does not require the agent to believe it as a known truth.

BIBLIOGRAPHY

Statements of agnosticism: The fundamental primary sources are T.H. HUXLEY, "Agnosticism" and "Agnosticism and Christianity" in his *Collected Essays*, vol. 5 (1894). Other basic sources are W.K. CLIFFORD, *The Ethics of Belief, and Other Essays* (1876, reprinted 1947); and LESLIE STEPHEN, *An Agnostic's Apology, and Other Essays* (1893), which first appeared as an essay in 1876.

Classical antecedents: DAVID HUME, *An Enquiry Concerning Human Understanding* (1748), of which the best modern edition is that of C.W. HENDEL (1955). For a study of this, which was Hume's first *Enquiry*, see ANTONY FLEW, *Hume's Philosophy of Belief* (1961), which devotes special attention to its implications for religion.

Religious agnosticism: HENRY MANSEL, *The Limits of Religious Thought Examined in Eight Lectures* (1858).

Secondary sources: LESLIE STEPHEN, *History of English Thought in the Eighteenth Century*, 2 vol. (1876), the best analytical study done at the time of Huxley and one that still deserves this high rating; R.F. FLINT, *Agnosticism* (1903); R.A. ARMSTRONG, *Agnosticism and Theism in the Nineteenth Century* (1905); for a recent sympathetic biography, see CYRIL BIBBY, *T.H. Huxley: Scientist, Humanist, and Educator* (1959).

Rejectors of agnosticism: BLAISE PASCAL, *Pensées* (1960), posthumous fragments given their definitive arrangement by L. LAFUMA; WILLIAM JAMES, "The Will to Believe," in *The Will to Believe, and Other Essays in Popular Philosophy* (1897).

(A.G.N.F.)

Agricola, Georgius

A 16th-century German scholar and writer on mining and metallurgical practice who based his work to a degree unprecedented in his time on observation and inquiry in the field, Agricola has been called the father of mineralogy. While a highly educated classicist and humanist, well regarded by scholars of his own and later times, he was yet singularly independent of the theories of ancient authorities. He was indeed among the first to found a natural science upon observation, as opposed to speculation.

By courtesy of the Museum National d'Histoire Naturelle, Paris



Agricola, portrait by an unknown artist, from Joannes Sambucus' *Icones veterum aliquot ac recentium medicorum philosophorumque . . .*, Antwerp, 1574.

Agricola was born Georg Bauer on March 24, 1494, in Glauchau, Saxony, of obscure parentage. From 1514 to 1518 he studied classics, philosophy, and philology at the University of Leipzig, which had recently been exposed to the Humanist revival. Following the custom of the times, he latinized his name to Georgius Agricola. After teaching Latin and Greek from 1518 to 1522 in a school in Zwickau, he returned to Leipzig to begin the study of medicine but found the university in disarray because of theological quarrels. A lifelong Catholic, he left in 1523 for more congenial surroundings in Italy. He studied medicine, natural science, and philosophy in Bologna and Padua, finishing with clinical studies in Venice.

Student years

For two years he worked at the Aldine Press in Venice, principally in preparing an edition of Galen's works on medicine (published in 1525). In this task he collaborated with John Clement, who had been Thomas More's secretary during the writing of *Utopia*. More's book may well have influenced Agricola to concern himself later with the laws and social customs of the Saxon mining district. In Italy he also met and won the friendship of the great scholar Erasmus, who encouraged him to write and later published several of his books. (Erasmus wrote an introduction to Agricola's first book, the mineralogical treatise *Bermannus*. Agricola shared the honour with More and only three other scholars.)

In 1526 he returned to Saxony, and from 1527 to 1533 he was town physician in Joachimsthal, a burgeoning mining town in the richest metal-mining district of Europe. Partly in the hope of finding new drugs among the ores and minerals of his adopted district (a hope eventually to be disappointed), he spent all his spare time visiting mines and smelting plants, talking to the better educated miners, and reading classical authors on mining. These years shaped the rest of his life and provided the subject matter for most of his books, beginning with *Bermannus; sive, de re metallica* (1530), a treatise on the Erzgebirge mining district. There are indications that he owned a share in a silver mine.

He appears not to have been particularly distinguished

as a physician, though in this pursuit he made use of direct observation rather than of received authority. He introduced the practice of quarantine into Germany, and his books make many references to miners' occupational diseases. In 1533 he became town physician in Chemnitz, where he remained to the end of his life.

Chief works

His magnum opus, for which the little book *Bermannus* was a curtain raiser, was *De re metallica*, published posthumously in 1556. In it, among other things, Agricola surveys historical and classical allusions to metals, and assesses the content and distribution of metal mines in antiquity. He treats the pattern of ownership and the system of law governing Saxon mines, together with the details of their day-to-day labour management. He was mainly concerned, however, with mining and metallurgy, and he discussed the geology of ore bodies, surveying, mine construction, pumping, and ventilation. There is much on the application of waterpower. He describes the assaying of ores, the methods used for enriching ores before smelting, and procedures for smelting and refining a number of metals, and he concludes with a discussion of the production of glass and of a variety of chemicals used in smelting operations.

In *De natura fossilium* (the book on which rests his right to be regarded as the father of mineralogy), Agricola offers a classification of minerals (called "fossils" at that time) in terms of geometrical form (spheres, cones, plates). He was probably the first to distinguish between "simple" substances and "compounds." In Agricola's day, chemical knowledge was almost nonexistent, and there was no proper chemical analysis (other than analysis of ores by the use of fire), so the classification of ores was necessarily crude.

In several other books, notably *De natura eorum quae effluunt ex terra* (1546) and *De ortu et causis subterraneorum* (1546), Agricola describes his ideas on the origin of ore deposits in veins and correctly attributes them to deposition from aqueous solution. He also describes in considerable detail the erosive action of rivers and its effect in the shaping of mountains. His readiness to throw overboard received authority, especially that of classical authors, such as Aristotle and Pliny, is impressive.

Career as mayor and diplomat

In 1546 Duke Maurice, the elector of Saxony, appointed Agricola burgomaster (mayor) of Chemnitz. He also served as an emissary in the Protestant ruler Maurice's ambiguous negotiations with Charles V, the Holy Roman Emperor. The religious wars of the period rapidly eroded the degree of tolerance that had hitherto prevailed in the Protestant German states, a tolerance from which Agricola had benefitted. He died on November 21, 1555.

Apart from his diplomatic role, Agricola took only limited interest in politics. His youthful "Turkish Speech" of 1529, a vigorous call to the Holy Roman emperor Ferdinand I to undertake a war against the Turks, was a patriotic hymn to Germany and a call to political and religious unity. It made a great impression on the public and was often reprinted.

Agricola's scholarly contemporaries regarded him highly. Erasmus prophesied in 1531 that he would "shortly stand at the head of the princes of scholarship." Later Goethe was to liken him to Francis Bacon. Melanchthon praised his "grace of presentation and unprecedented clarity." The mining engineer Herbert Hoover (later U.S. president), who translated *De re metallica* into English in 1912, regarded Agricola as the originator of the experimental approach to science, "the first to found any of the natural sciences upon research and observation, as opposed to previous fruitless speculation."

BIBLIOGRAPHY. No detailed life of Agricola has been published in English. A German biography is H. HARTMANN, *Georg Agricola, 1494–1555* (1953). The most important source is the work published by the East German Academy of sciences on the occasion of the 400th anniversary of his death, *Georgius Agricola, 1494–1555, zu Seinem 400. Todestag 21. Nov.* (1955), which contains a number of essays on Agricola's life and work, including his diplomatic activity. F.D. ADAMS, *The Birth and Development of the Geological Sciences* (1938), is informative on Agricola's contribution to geology. Of his books, only *De re metallica* and *De natura fossilium* have as yet been translated into English.

(R.W.C.)

Agricultural Economics

Agriculture is the source of livelihood for more than half of the world's population. In some countries more than 70 percent of the inhabitants support themselves by farming, while in the more industrialized countries the proportion ranges much lower—to less than 7 percent in the United States and less than 4 percent in Great Britain. In general one can say that when a large fraction of a nation's population is dependent upon agriculture for its livelihood, average incomes are low. This does not mean that a nation is poor because most of its population is engaged in agriculture; it is closer to the truth to say that because a country is poor, most of its people must rely upon agriculture for a living.

AGRICULTURE AND ECONOMIC DEVELOPMENT

The rural surplus

As a country develops economically, the relative importance of agriculture declines. The primary reason for this was shown by the 19th-century German statistician Ernst Engel, who discovered that as incomes increase the proportion of income spent on food declines. For example, if a family's income were to increase by 100 percent, the amount it would spend on food might increase by 60 percent; if formerly its expenditures on food had been 50 percent of its budget, after the increase they would amount to only 40 percent of its budget. It follows from this that, as incomes increase, a smaller fraction of the total resources of society is required to produce the amount of food demanded by the population.

Progress in farming. This fact would have surprised most economists of the early 19th century, who feared that the limited supply of land in the populated areas of Europe would determine that continent's ability to feed its growing population. Their fear was based on the so-called law of diminishing returns: that in a given state of the arts an increase in the amount of labour and capital applied to a fixed amount of land would result in a less than proportional increase in the output of food. The principle of diminishing returns is a valid one. What the classical economists could not foresee was the extent to which the state of the arts and the methods of production would change in the decades ahead. Some of the changes occurred in agriculture; others, such as the reduction in the costs of shipping, occurred in other sectors of the economy but had a major effect on the supply of food.

Throughout the last three or four centuries, man's increased understanding of nature has been applied to agriculture at least as effectively as it has to other fields of endeavour. Advances in basic science have made clear the factors affecting the growth of plants and animals. Chemistry, physics, biology, and genetics have been applied to the production process in agriculture. Fertilizers, insecticides, crop rotation, plant and animal breeding, and methods of plowing and cultivating are largely the offspring of science. The fact that science has not been applied equally in all countries is one of the chief reasons for the uneven development of the world's agriculture.

In looking back upon the history of the more developed countries, one can see that agriculture has played an important part in the process of their enrichment. For one thing, if development is to occur, agriculture must be able to produce a surplus of food to maintain the growing nonagricultural labour force. Since food is more essential for life than are the services provided by merchants or bankers or factories, an economy cannot shift to such activities unless food is available for barter or sale in sufficient quantities to support those engaged in them. Unless food can be obtained through international trade, a country does not normally develop industrially until its farm areas can supply its towns with food in exchange for the products of their factories.

Economic development also requires a growing labour force. In an agricultural country, most of the workers needed must come from the rural population. Thus agri-

culture must not only supply a surplus of food for the towns but it must also be able to produce the increased amount of food with a relatively smaller labour force. It may do so by substituting animal power for human power or by gradually introducing labour-saving machinery.

Agriculture may also be a source of the capital needed for industrial development to the extent that it provides a surplus that may be converted into the funds needed to purchase industrial equipment or to build roads and provide public services.

For these reasons a country seeking to develop its economy may be well advised to give a significant priority to agriculture. Experience in the developing countries has shown that agriculture can be made much more productive with the proper investment in research, fertilizers, insecticides, and herbicides. Improvement of irrigation systems may also be essential.

Fortunately, many of the recent advances in applied science do not require massive amounts of capital, although it may be necessary to expand marketing and transportation facilities so that farm output can be brought to the entire population.

One difficulty in giving priority to agriculture is that most of the increase in farm output and most of the income gains will be concentrated in some regions rather than extending through a whole country. The remaining farmers will not be able to produce more and will actually suffer a disadvantage as farm prices decline. There is no easy answer to this problem, but governments of the developing countries need to be aware of it; economic progress is consistent with lingering backwardness, as can be seen in parts of southern Italy or in the Appalachian Mountain area of the United States.

Peasant agriculture. One characteristic of undeveloped peasant agriculture is its self-sufficiency. Farm families in those circumstances consume a substantial part of what they produce. While some of their output may be sold in the market, their total product is generally not much larger than what is needed for the maintenance of the family. The typical farm worker will produce enough for himself and about three other persons, of whom two will be members of his own household. This was true of the United States in 1840 and is true of India in 1970. Not only is productivity per worker low under these conditions but yields per unit of land are also low. Even where the land was originally fertile, the fertility is likely to have been depleted by decades of continuous cropping. The available manures are not sufficient, and the farmers cannot afford to purchase them from elsewhere.

Peasant agriculture is often said to be characterized by inertia. The peasant farmer is likely to be illiterate, suspicious of outsiders, and reluctant to try new methods; food patterns remain unchanged for decades, even centuries. Recent evidence, however, suggests that the apparent inertia may be simply the result of a lack of alternatives. If there is nothing better to change to, there is little point in changing. Moreover, the self-sufficient farmer is bound to want to minimize his risks; since a crop failure could mean starvation, farmers have been reluctant to adopt new methods if doing so exposed them to greater risks of failure.

The spread during the 1960s of high-yielding varieties of rice and wheat showed that farmers were willing and able to adopt new methods when their superiority was demonstrated. In India the new wheat varieties were introduced in 1966–67 and 500,000 hectares were seeded. By 1968–69 the area seeded to the new varieties had increased to 4,000,000 hectares and increased further to 5,000,000 hectares, or 30 percent of the total wheat area in 1969–70. The area planted to high-yielding varieties of rice increased threefold between 1967–68 and 1969–70. The introduction of the new rice varieties in Pakistan did not occur in any significant degree until 1968–69, yet in West Pakistan a fourth of the total rice area was planted to the new varieties one year later. In 1969–70 2,500,000 out of 4,200,000 hectares of irrigated wheat in West Pakistan were seeded to high-yielding varieties.

The labour force. As economic development proceeds, a large proportion of the farm labour force must shift from agriculture into other pursuits. Trends in various parts of the world are illustrated in Table 1. In the United States, employment in agriculture declined from about 72 percent of all employment in 1820 to about 5 percent in 1968. In Africa, about three-quarters of the population are engaged in agriculture, and the absolute number so employed grew from 128,000,000 in 1937 to 231,000,000 in 1965—reflecting the growth in total population. Central and South America were still, in 1965, passing through the phase of economic development in which the absolute number employed in agriculture increases, although in both regions the percentage of the total population in agriculture declined significantly over the three decades.

This fundamental shift in the labour force is made possible, of course, by an enormous increase in output per worker as agriculture becomes modernized. This increase in output stems from various factors, and the relative importance of them can be estimated by comparing the productivity of agriculture in different countries. Where land is plentiful the output per worker is likely to be higher because it is possible to employ more fertilizer and machinery per worker. Table 2, taken from a study by Yujiro Hayami, compares the agricultures of the United States, Japan, and India. Japan has a modern agriculture, but the ratio of land per worker is very low and this adversely affects the amount of fertilizer and machinery that can be employed. Statistical analysis indicates that the 43:1 ratio in farm output per worker between the United States and India is attributable in varying degrees to the factors shown in Table 2: the use of fertilizer per worker, the availability of land per worker, the use of

Table 1: Total Population and Population Engaged in Agriculture, World and Regions: 1937, 1950, 1965

	1937			1950			1965		
	total population (000,000)	population engaged in agriculture (000,000)	percentage of total engaged in agriculture	total population (000,000)	population engaged in agriculture (000,000)	percentage of total engaged in agriculture	total population (000,000)	population engaged in agriculture (000,000)	percentage of total engaged in agriculture
Europe	372	133	36	392	129	33	445	103	23
U.S.S.R.	190	108	57	181	90	50	231	73	32
North and Central America*	178	56	31	218	52	24	294	55	19
North America	140	32	23	166	23	14	214	13	6
Central America	38	24	63	52	29	56	80	42	53
South America	84	52	62	110	65	59	165	76	46
Asia†	691	504	73	818	518	64	1,132	732	65
Africa	168	128	76	218	166	76	313	231	74
Oceania	11	3	26	12	3	22	17	3	19
World†	1,694	985	58	1,949	1,023	53	2,597	1,273	49
Mainland China	447	40	76	547	377	69	764	481	63
World total	2,141	1,325	62	2,496	1,403	56	3,361	1,754	52

*Combined figures; figures for individual regions given below. †Excluding mainland China.
Source: UN Food and Agriculture Organization, *Production Yearbook*.

Table 2: Farm Output and Inputs: United States, Japan, and India

	United States	Japan	India	United States* / Japan	United States† / India
Farm output (wheat units per worker)	96.2	10.7	2.22	9.0	43.3
Fertilizer (metric tons per worker)	2.13	0.32	0.0038	6.7	560.5
Machinery (horsepower per worker)	45.9	1.07	0.0077	42.9	5,961.0
Agricultural land (hectares per worker)	129.7	1.43	1.99	90.7	65.2
Education (school enrollment ratios)	100	90	26	1.1	3.8
Research and extension (college graduates per 10,000 farm workers)	22.8	14.2	0.40	1.6	57.0

*Column one divided by column two. †Column one divided by column three.
Source: Yujiro Hayami, *American Journal of Agricultural Economics*, vol. 51, no. 3, August 1969.

machinery per worker, the amount of education, and the availability of farm research and extension services. A significant comment on the unevenness of the march of science is that in the late 1960s North America and Europe (excluding the U.S.S.R.) had three-fourths of the world's tractors and used more than three-fifths of the total tonnage of commercial fertilizer.

The modern farm is a complex enterprise, dependent upon other sectors of the economy for much of its input. The level of management required is correspondingly high. While the manager-operators of modern farms have often been trained in agricultural colleges, most of them have obtained their skills through experience and from various specialized sources of information such as government extension services, farm publications, and farm suppliers. Considering that farm operators must make most of their decisions subject to the uncertainties imposed by nature, the level of managerial ability required is probably comparable to that in many lines of business. Much has been written about the decline of the family farm. Yet experience in the United States, western Europe, and Japan indicates that family farming is not incompatible with modern techniques. In the United States, family farms were responsible for about two-thirds of all farm sales in 1929 and in 1964. The average size of the family farm in the U.S. is over 300 acres. In Japan, where family farms average about 3 acres, agriculture is nevertheless highly complex and mechanized.

EFFORTS TO CONTROL PRICES AND PRODUCTION

The need for control

Governments in recent decades have undertaken to control both prices and output in the agricultural sector, largely in response to the pressures of the farmers themselves. In the absence of such control, farm prices tend to fluctuate more than do most other prices; and the incomes of farmers fluctuate to an even greater degree. Not only are incomes in agriculture unstable but they tend to be lower than incomes in other economic sectors.

The instability of farm prices results from several factors. One is the relative slowness with which farmers are able to respond to changes in the demand for their product. Farmers generally must produce on the basis of expectations, and if their expectations turn out to be wrong, the resulting surplus or shortage cannot be corrected until the beginning of the next production cycle. Once a crop is planted, very little can be done to increase or decrease production in response to market prices. As long as prices cover current operating costs, such as the cost of harvesting, it will pay farmers to carry through their production plans even if prices fall to a very low level. It is not unusual for the prices of particular farm products to vary by a third or a half from year to year. This extreme variability results from the relatively low responsiveness of demand to changes in price—*i.e.*, from the fact that in order to increase sales by 5 percent, it may be necessary to reduce the price by 15 percent.

Along with the instability of farm prices goes instability of farm income. While gross income from agriculture generally does not vary as much as do individual farm prices, net income may vary more than prices. In modern agriculture, costs tend to be relatively stable; the farmer is unable to compensate for a drop in prices by reducing his payments for machinery, fertilizer, labour, etc.

The incomes of farm workers are generally far below those of other workers. There are two major reasons for this inequity. One is that in most economies the need for farm labour is declining, and each year large numbers of farm people, especially young ones, must leave their homes to seek jobs elsewhere. Part of the difference in returns to labour is required to bring about this transfer of workers out of farming; if the transfer did not occur, the income differences would be even larger. The second main reason for the income differences is that farm people generally have less education than do nonfarm people and are able to earn less at nonfarm jobs. The difference in education is of long standing and is found in all countries, developed and undeveloped; it also exists whether the national education system is highly decentralized, as in the United States, or highly centralized, as in France or the U.S.S.R.

Governments have employed various measures to maintain farm prices and incomes above what the market would otherwise have yielded. These have included tariffs or import levies, import quotas, export subsidies, direct payments to farmers, and limitations on production. Tariffs and import quotas can be effective only if a nation normally imports some of its supply. Export subsidies result in higher prices to domestic consumers than to foreign purchasers; their use requires control over imports to prevent foreign supplies from entering the domestic market and bringing prices down. Direct payments to farmers have been used in order to maintain prices to consumers at reasonable levels while assuring farmers a return that would be above world-market levels. Limitations on production, intended to reduce supply and thus increase prices, have been used mainly in Brazil (for coffee) and in the United States (for major crops).

The British approach. Efforts to control agricultural prices go far back in English history, although the early objectives were quite different from those of more recent times. The Corn Laws of the 15th century were designed to prevent prices from becoming too high; restrictions were imposed on the right to export corn (wheat) when the domestic price exceeded a specified level. In 1663 the laws were revised to prevent prices from falling too low, by including import duties when the home price did not exceed a specified level. The general trend, until the Corn Laws were abandoned in 1846, was increasingly toward ensuring higher prices for home producers through the payment of export bounties and by preventing imports until prices reached specified levels. After 1846 the British followed a free-trade policy for agricultural products, but moved to the protection of agriculture and the establishment of minimum prices for certain farm products during the depression of the 1930s. Protection was expanded after World War II by the Agricultural

Acts of 1947 and 1957. These sought to support farm prices primarily through deficiency payments to farmers, covering about 95 percent of total output. In most cases, the domestic price was free to vary with changing demand and supply conditions; local products competed with imported supplies that were generally subject to relatively low tariffs. The farmer was reimbursed for the difference between his average realized price and a guaranteed price. Under the Agricultural Act of 1957, the government had the right to limit the amount of output on which deficiency payments were made as a means of reducing the cost of the program and of restricting the incentives to increase domestic production.

A feature of the system was that domestic market prices fluctuated in response to changing conditions in world markets. If market prices fell because of increased imports, an increase in deficiency payments from the treasury was required. A rise in world prices resulted in higher consumer prices but lower payments from the treasury. Bilateral market-sharing agreements for butter, bacon, and grains have been negotiated with other countries in an attempt to stabilize prices. In 1964 a minimum import price agreement for grain and grain products was negotiated with all major suppliers, including the United States; if the import price falls below the minimum, a levy is imposed on buyers to bring the cost up to the minimum, the purpose being to restrict imports. As a part of this agreement, deficiency payments to grain producers were limited to a standard quantity. Similar standard quantities were established for other products, such as pigs and milk.

Returns to producers were also supported through producers' marketing boards, which were set up for milk, eggs, wool, hops, and potatoes. These boards were given a monopoly on the marketing of these products; they attempted to influence prices by various devices, including grading, distribution of supplies among the different markets, and, in the cases of potatoes and hops, assigning production quotas to individual producers.

The U.S. approach. Since the enactment of the Agricultural Adjustment Act of 1933, the United States has had programs designed to limit the production of major farm crops through restrictions on acreage. Since that date it has also offered price supports for major crops such as wheat, feed grains, rice, tobacco, and cotton as well as for manufactured dairy products. It has not had price-support programs for perishable crops or for major livestock products except for a few years during and after World War II.

The Commodity Credit Corporation

The price support method most widely used has been the non-recourse loan made by the Commodity Credit Corporation (CCC): the farmer may repay the loan by delivering his produce at the support price or "redeem" it in cash if the market price is higher. Though the amount of particular crops offered for price-support loans varied greatly from year to year, as did redemptions, surplus stocks in CCC hands increased during the late 1950s, were reduced significantly during the early and mid-1960s, and then increased again as the 1970s began.

Most of the farm products given price supports were crops normally exported by the United States. Until the mid-1960s, the price supports were above the export prices. Unless export subsidies were paid to make up the difference between domestic prices and the prices foreign buyers were willing to pay, exports thus became impossible. Export subsidies were accordingly paid on several farm commodities such as cotton, wheat, and feed grains. In the 1960s the support prices for the major export commodities, except tobacco, were established at levels near or slightly below world prices to permit market forces to manage the distribution of supplies between domestic and foreign markets. The lowering of support prices was accompanied by a substantial increase in the size of direct payments to farmers. By the end of the 1960s such payments had come to constitute a high percentage of the cash receipts from farm marketings: in cotton, 60 percent; in wheat, 40 percent; and in feed grains, 30 percent.

In order to receive payments, farmers had to agree to limit the acreage devoted to specified crops. At the beginning of the 1970s, the various programs had resulted in the diversion of approximately 50,000,000 acres (20,000,000 hectares) of land from the production of major farm crops.

Price objectives for sugar are met by limiting the amount that can be imported; the domestic price has been maintained at two to three times the import price. Payments to domestic producers, financed by a processing tax on sugar refining, are also made. While acreage has been limited, an increasing fraction of U.S. sugar supplies has been provided from domestic production.

Acreage limitations have not been used for soybeans, and in most years the price support for soybeans has been set near or slightly below market prices. In a few years in the late 1960s, the price support for soybeans was established at a level that resulted in the accumulation of substantial stocks by the CCC.

A major component of U.S. farm price policy since World War II has been the disposal of surplus produce abroad through the economic aid program. This began as an outgrowth of wartime Lend Lease; American food made a major contribution to the postwar recovery of western Europe. The Agricultural Trade and Development Act of 1954 provided a base for continuing such activities, and gradually the emphasis shifted from western Europe to the developing countries. By 1970 approximately $20,000,000,000 worth of food and other farm products had been shipped abroad under this legislation. One of the important effects was to dispose of farm products that could not be sold either domestically or in regular commercial foreign trade. Without this the farm income and price objectives could not have been achieved except by more stringent output limitations, lower farm prices, and larger direct government payments.

Policies of the EEC. The founders of the European Economic Community (EEC) provided for the establishment of a common agricultural policy (CAP) for the six Common Market countries (France, Germany, Italy, Belgium, Luxembourg, and the Netherlands). The CAP, to be worked out for each major farm commodity, was to create free trade for that commodity within the community. Special subsidies by the individual countries, and other national farm programs, were to be eliminated to prevent competitive advantages. The first of the regulations implementing the CAP were enacted in 1962 and applied to grain (except rice), poultry, eggs, live hogs and whole hog carcasses, fruit and vegetables, and wine. Similar programs were developed later for beef, dairy products, sugar, rice, and fats and oils.

Features of the common agricultural policy

The most important features of the CAP mechanism are the target prices, the threshold prices, the support or intervention prices, the variable levies on imports to make up the difference between landed prices and threshold prices, and export subsidies or refunds equal to the difference between market prices in the EEC and in the importing country. The support or intervention price, when applicable, is generally 90–95 percent of the target price. For most CAP commodities, the primary device for achieving target prices is the variable import levy. This levy, which fluctuates with the import cost of a commodity, keeps the domestic price at or near the target price if the commodity is imported. When EEC production of a commodity exceeds EEC consumption, the authorities may purchase the commodity for storage, pay to have it processed for another use (*e.g.*, wheat may be denatured and sold as a feed grain), or subsidize its export to countries outside the EEC. With these techniques, the EEC has been able to maintain farm prices at levels substantially higher than those prevailing in the United States, Canada, and the United Kingdom. In the early 1970s, the EEC prices for grains were 60 to 90 percent higher than the import costs of the same grains; similar differences existed for beef and veal, and substantial but smaller ones for poultry and pork.

Throughout the 1960s the EEC did nothing to limit or control the production of agricultural products. When

large stocks of butter and dry skim milk accumulated, and as the costs of maintaining dairy product prices and subsidizing wheat exports mounted, consideration was given to reducing production. A payments program to induce shifts from dairy to beef production was inaugurated, and there was talk of reducing the cultivated area for grain. Output limitation has been made difficult, however, by the significant differences in circumstances among the farmers in the various EEC countries.

Soviet policies. The farm policies of the Soviet Union were established during the First Five Year Plan (1928–32), when agriculture was collectivized. For all practical purposes, regular markets for farm products were abolished at that time, and each collective farm was required to deliver an assigned quota of produce to the state at very low prices. If a farm had anything left after meeting the obligatory quotas, it could sell the surplus to the state at higher prices or to the local free markets. Until 1958 the collective farms also had to make payments in kind to the machine tractor stations in return for work done.

After Stalin's death in 1953, farm prices were increased significantly; by 1956 the average of procurement prices for food products had increased almost fourfold compared to 1950. In 1958 the multiple-price system was abandoned, and the prices paid to collective farmers became almost seven times the average paid in 1950. The machine tractor stations were abolished in 1958 and the machinery transferred to individual farms. Another major revision of prices was made by the post-Khrushchev government in 1964–65. A two-price system replaced the single prices; prices for deliveries of grain up to the planned amount were about 10 percent higher than the previous single price, and deliveries above the plan level received a premium of 50 percent. Significant regional price differentials were established to cover the higher costs of production in some regions. Prices of livestock products had already been increased by about 35 percent in 1962, and in 1965 further increases of perhaps a third were made. Another important measure was a commitment that planned purchases were to be fixed at specified levels during the Eighth Five Year Plan (1966–70), both in the aggregate and for individual farms. Prior to that time, if a farm significantly increased its production or if other farms in the same region failed to meet their deliveries, the delivery quota might be arbitrarily increased for the farm that had some output available for delivery.

The Soviet two-price system

Before 1953 Soviet price policy was clearly designed to obtain farm products as cheaply as possible. The low prices were generally not passed on to consumers; a significant fraction of total governmental revenue was derived from high taxes on farm products. The changes made after 1953 were intended to provide farmers with an incentive to raise production and to make more efficient use of resources. Only a part of the increase in prices paid to farms was passed on to consumers; much of the increase was at the expense of government revenue.

The problem of standards. None of the governments engaged in regulating farm prices and incomes has been able to apply a meaningful standard as to what a fair price or reasonable income is. The actual measures adopted, such as specific price supports or intervention prices, have been determined through the political process with little reference to formal principles or standards.

In the United States the Agricultural Adjustment Act of 1933 stated that the goal should be to establish prices having the same purchasing power as those of the period 1910–14. By the end of World War II, it became clear that the "parity price" relationships of 1910–14 were no longer relevant to existing conditions. The Agricultural Act of 1948 retained the 1910–14 average as a parity for all farm products but stated that the parity for individual products was to be the average of prices over the most recent ten-year period. Since the application of this formula would have resulted in a significant reduction in the parity prices for some politically important farm products, especially cotton and wheat, legislation passed in 1949 declared that the parity price for an individual commodity was to be determined by either the old or the new formula, whichever was higher. Not until 1955 were the "modernized parity" prices put into effect on a gradual basis. At the end of the 1960s, U.S. farm price and income policies followed no identifiable objective. Parity prices were no longer used as a guide except for wheat.

A similar lack of objective or measurable standards is evident in Great Britain, where the Agriculture Act of 1947 declares its intention to be that of

> promoting and maintaining a stable and efficient agricultural industry capable of producing such part of the nation's food and other agricultural produce as in the national interest it is desirable to produce in the United Kingdom, and of producing it at minimum prices consistently with proper remuneration and living conditions for farmers and workers in agriculture and an adequate return on capital invested in the industry.

Several other countries have legislation that aims, without specifying operationally what is meant, to obtain for farm people the same level of income as that of other groups in the economy or that states that farm people should share in the rise in real per capita incomes. Finland, Japan, France, Sweden, and Norway have such policy objectives. West German legislation declares that agriculture should share in the progressive development of society and requires that the government each year prepare a report showing the extent to which the return to farm labour, or properly managed holdings under average conditions, is in line with that of wage earners in comparable nonagricultural occupations in rural areas.

The agricultural price objectives of the Treaty of Rome establishing the EEC also lack operational significance:

> To ensure . . . a fair standard of living for the agricultural population, particularly by the increasing of the individual earnings of persons engaged in agriculture; to stabilize markets; to guarantee regular supplies; and to ensure reasonable prices in supplies to consumers.

Accomplishments. The effects of price and income policies are difficult to assess. The policies have unquestionably worked to raise agricultural production in the countries where they have been applied, but their usefulness as a means of enhancing the economic well-being of farm people is debatable. The governments of the industrial countries were able, in the 1960s and earlier, to raise the returns from agriculture above the levels that would have prevailed in the absence of such intervention. In addition to maintaining prices, they provided subsidies for some of the agricultural inputs such as tractor fuel and chemical fertilizers; they also gave assistance in consolidating small farms into larger ones and in improving farm buildings and structures. They did not, except for the United States and Great Britain, attempt to moderate the effects of these policies on production. In 1969 agricultural production in the United States was 21 percent greater than in 1957–59; production in the European Economic Community had increased by 30 percent, in Australia by 46 percent, in Canada by 41 percent, and in the U.K. by 46 percent.

The UN Food and Agriculture Organization recommended in 1969 that the rich countries change their policies to place less emphasis on self-sufficiency in farm output and more emphasis on international trade. It pointed out that if the developing countries were enabled to sell more of their produce to the industrial countries, all parties would benefit (FAO, *Provisional Indicative World Plan for Agricultural Development*).

The level of income and the economic well-being of farm people in general are determined by many factors, including not only the prices they receive for their output but also the rate at which the economy in general is growing, the ease with which people can move from farm to nonfarm jobs, the prices they must pay for their productive inputs, and their level of education. With respect to average income per person, as distinguished from total income, the prices received and paid are probably less important than the other factors mentioned. This becomes obvious when one compares farm incomes in the United States or the United Kingdom with those in Argentina or

India; the differences in real income have to do mainly with the levels of economic development and not with farm prices or subsidies. Government efforts to increase farm prices are likely to be offset, in the long run, by an increase in the number of persons engaged in farming, and this will tend to keep the returns to farm labour from rising much faster than they would in the absence of such policies.

There are two other reasons for believing that the income effects of higher farm prices or subsidies are relatively insignificant in the long run compared to other factors affecting incomes of farm workers (including farm operators and the members of their families). One is that an increase in farm prices induces farmers to use more fertilizer, machinery, fuel and oil, etc. If a significant part of any increase in gross income is used for such things, the absolute increase in net farm income is much smaller than the increase in gross farm income. The other reason is that a given increase in government-supported farm prices, say 25 percent, generally occurs only once. After the increase in returns has been realized, the higher farm prices will contribute nothing further to incomes in the future. In contrast, general economic growth along with the continued reduction of the farm labour force has cumulative effects on the return to farm labour; in the United States the returns to farm labour grew at an average annual rate of about 3 percent during the 1960s. Farm prices would have to increase at least 3 percent annually (assuming other prices did not change) to have the same effect on returns to farm resources.

Costs. The costs of the agricultural price and income policies of the industrial countries are substantial; they include not only direct governmental outlays but the increased costs to consumers in those countries, as well as the losses to developing countries of potential export markets. An estimate of the total costs to taxpayers and consumers in the Common Market countries has been made for the late 1960s. According to the U.S. Department of Agriculture, the total excess costs were $14,400,000,000, of which $6,400,000,000 represented excess consumer costs and the remainder were costs borne by taxpayers. The total costs amounted to $78 per capita and may be compared to per capita food expenditures in the EEC of $356 and national income per capita of $1,412. They amounted to $1,300 for each farm worker in the EEC. In the United States in the late 1960s the federal budgetary expenditures for agriculture averaged about $1,200 per farm worker, as compared to a similar figure for the EEC of about $700. Excess consumer costs in the United States, however, were much smaller than in the EEC, either on a per farm worker basis or as a percentage of total consumer expenditures on food.

LAND, OUTPUT, AND YIELDS

Land and population

Only a small fraction of the world's land area—about 11 percent—may be considered arable, if arable land is defined as land planted to crops, including temporary fallow and temporary meadows, and under-tree crops. Less than a fifth of the world's land area is in permanent meadows and pastures. The remainder is either in forests or is not being used for productive purposes.

There are great differences in the amount of arable land per person in the various regions of the world, as can be seen in Table 3. The greatest amount of arable land per capita is in Oceania; the least is in Asia. No direct relationship exists between the amount of arable land per capita and the level of income; Europe has almost as little arable land per capita as Asia, and far less than Africa; Japan and The Netherlands have very limited amounts of arable land per capita, 0.06 and 0.07 hectares respectively.

The relationship between land, population, and farm production is a complex one. In traditional agriculture, where methods of production have changed little over a long period of time, production is largely determined by the quality and quantity of land available and the number of people working on the land. Until the early years of the 20th century, most of the world's increase in crop production came either from an increase in land under cultivation or from an increase in the amount of labour used per unit of land. This generally involved a shift to crops that would yield more per unit of land and required more labour for their cultivation. Wheat, rye, and millet require less labour per unit of land and per unit of food output than do rice, potatoes, or corn (maize); but generally the latter yield more food per unit of land. Thus as population density increased, the latter groups of crops tended to be substituted for the first. This did not hold true in Europe, where wheat, rye, and millet expanded at the expense of pasture land; but these crops yielded more food per acre than did the livestock that they displaced.

Table 3: Arable Land

	arable land (000,000 hectares)	arable land per person (hectares)
Europe*	149	0.33
U.S.S.R.	224	0.94
North and Central America	253	0.82
South America	89	0.39
Asia†	335	0.27
Mainland China	109	0.13
Africa	204	0.60
Oceania	43	2.32
World total	1,447	0.39

*Excluding U.S.S.R. †Excluding mainland China.
Source: UN Food and Agriculture Organization, *Production Yearbook*, Vol. 23, 1969. Population data are for 1968; arable land data are generally for 1968, but precise dates vary from country to country.

As agriculture becomes modernized, its dependence upon land as well as upon human labour decreases. Animal power and machinery are substituted for human labour; then mechanical power replaces animal power. The substitution of mechanical power for animal power also reduces the need for land. For example, in the United States in 1910–14 the supply of feed for horses and mules required 91,000,000 out of 330,000,000 acres of crops harvested; in addition, a significant part of the total pasture was used by horses and mules, and almost a tenth of all labour was employed in caring for the horses and mules. The increased use of fertilizer as modernization occurs also acts as a substitute for both land and labour; the same is true of herbicides and insecticides. By making it possible to produce more per unit of land and per hour of work, less land and labour are required for a given amount of output.

Recent trends. Reliable data on crop yields and area cultivated are now available for most countries in the world. In the period from the 1930s to the late 1960s, a number of interesting developments occurred. Over this period of approximately three decades, the world crop area increased by 46 percent; total agricultural production increased by 94 percent. Consequently the increase in average yield per unit of land under crops was 33 percent; the increase in crop area was accordingly more important than the increase in yields in influencing total agricultural production.

In Europe (excluding the U.S.S.R.) and North and Central America, the total area under crops declined; in South America and Africa it increased by more than half and in Asia by a third. The large increase in Oceania was due to immigration (see Table 4).

Grain yields in the developed regions of the world increased consistently over the three decades (Table 5). In the rest of the world the prewar yields were not achieved until the mid-50s. The area sown to grain in the developed regions declined by 4 percent, while the grain area elsewhere increased by 49 percent. The increases in total grain production were identical (66 percent).

If yields of individual crops in particular countries are compared, they differ much more than do the averages for the two world regions. Among important producing countries rice yields vary from 1.1 tons per hectare in Cambodia to 5.3 tons per hectare in Japan; wheat yields vary from 0.5 ton in Algeria to 4.0 tons in the United Kingdom, and corn yields from 1.3 tons in the Philippines to 4.7 tons in the United States.

Table 4: Changes in Areas Under Major Crops and Changes in Population, 1937–65 (1965 figures; given as percentages of 1937; 1937 = 100)

crops	world	Europe*	North and Central America	South America	Asia†	Africa	Oceania
Grain	143	89	89	143	139	147	162
Other§	165	101	151	170	122	210	‡
Cotton	111	‡	58	178	101	156	‡
Total‖	146	92	95	153	134	158	162
Population increase	157	120	164	196	156	187	155

*Excludes U.S.S.R. †Excludes mainland China. ‡Not given when less than 1 percent of crop area devoted to the broad category. §Potatoes, sweet potatoes and yams, pulses, and oilseed. ‖World totals include U.S.S.R. and mainland China.
Source: UN Food and Agriculture Organization, *Production Yearbook*, 1965, 1968.
Data for crop areas are for 1934–38 and 1965.

Food production and total agricultural production exhibit nearly identical trends, and therefore changes in food production can be taken as indicative of changes in total agricultural production. Since 1948–52, food production has increased at approximately the same rate in the developed regions as in the rest of the world, but the trends in population growth have been widely different. Table 6 shows that in the developed regions population increased by 21 percent from 1952–56 to 1970, while in the rest of the world it increased by 50 percent. This meant a much more modest increase in food supplies per capita in the less-developed than in the developed regions: about 6 percent as compared to 24 percent.

THE ORGANIZATION OF FARMING

Except in nations with Communist governments, most farm land is privately owned. This does not mean, however, that the land is owned by those who farm it. In most countries a major aspiration of farm people has been to achieve the ownership of the land they work. After World War II, Japan and Taiwan underwent land reforms that were intended to broaden ownership; these are generally considered to have been highly successful. Similar reforms are being advocated in other countries.

The variety of forms of ownership and operation

On a cooperative farm the land is owned jointly by the members of the group who farm it. The cooperative generally also owns all of the major means of production and the members supply all or most of the labour. While there are examples of cooperative farms in many countries, they loom large only in Israel, where the kibbutzim control about a fifth of the total agricultural land.

In a collective farm, at least as organized in the Soviet Union, the land is owned by the state but is permanently leased to the *kolkhoz* (collective farm). The *kolkhoz* owns its own equipment and livestock and is required to meet certain commitments to the state in the form of deliveries of farm products. In theory, the members of the *kolkhoz* elect the officers of the farm and establish the procedures by which the net product is divided among the members for services performed. In practice, their autonomy has been severely limited by the economic plans. In most cases these plans have been incredibly detailed, specifying the crops to be grown, the times of plowing, planting, harvesting, etc.; the quantities of fertilizer and manures to be used; and the kinds of livestock to be maintained.

On state farms the land and all other means of production are owned by the state. The workers are paid in wages, and management decisions are made by individuals directly responsible to the state.

Kinds of farm operation. If a family farm is defined as one where the farm operator and members of his family supply at least half of the labour, the majority of farms in the world are family farms. Family farming is carried on under a wide range of conditions, from the small farms of Asia to the highly mechanized farms of Canada, the United States, and the United Kingdom.

The family farm may be owned by the farmer or rented. The most rapidly expanding type of tenure in the United States is that in which the farmer owns part of the land and rents the remainder; in 1964 almost half of all farmland in the United States was cultivated by part-owner farms. This arrangement enables the farmer to increase the size of his farm through renting and invest his capital in machinery and livestock.

Family farms may be large in terms of total assets or sales. In the United States in 1964, a seventh of all farms

Table 5: Grain Yields
(metric tons per hectare)

	1934–38	1952–56	1965–67	1969–70
Developed countries*	1.14	1.37	1.95	2.14
Rest of the world	1.16	1.12	1.29	1.41
Total world	1.15	1.23	1.56	1.70

*Includes North America, Europe (including the U.S.S.R.), and Oceania.
Source: UN Food and Agriculture Organization, *Production Yearbook*, 1956, 1968; *Monthly Bulletin of Agricultural Economics and Statistics* (January, 1972).

Table 6: Index Numbers of Food Production, Population, and Per Capita Food Production, 1948–52 to 1970, Selected Years
(1952–56 = 100)

	1948–52	1956	1960	1967	1968	1970
Food production						
Western Europe	84	103	118	142	145	146
Eastern Europe and U.S.S.R.	83	114	133	168	176	176
North America	92	104	111	132	133	129
Oceania	92	101	121	145	175	164
Israel, Japan, and South Africa	81	110	123	172	173	173
Developed regions	87	107	120	147	151	149
Latin America	87	109	117	150	151	165
Near East (excluding Israel)	82	110	122	151	157	161
Far East (excluding Japan and mainland China)	87	107	122	141	149	161
Africa (excluding South Africa)	88	106	117	131	135	140
Developing regions	87	108	120	143	148	158
World (excluding mainland China)	87	107	120	145	150	152
Population						
Western Europe	97	102	105	112	113	115
Eastern Europe and U.S.S.R.	95	103	109	118	119	122
North America	93	104	111	123	124	127
Oceania	91	105	115	133	136	140
Israel, Japan, and South Africa	94	103	108	117	119	123
Developed regions	95	103	108	117	118	121
Latin America	90	106	118	144	148	157
Near East (excluding Israel)	90	105	117	140	144	153
Far East (excluding Japan and mainland China)	92	104	114	137	140	148
Africa (excluding South Africa)	91	105	115	137	140	148
Developing regions	92	105	115	138	142	150
World (excluding mainland China)	93	104	112	129	132	137
Food production per capita						
Western Europe	86	101	113	127	129	127
Eastern Europe and U.S.S.R.	87	111	122	142	148	144
North America	99	101	100	107	107	101
Oceania	102	96	106	109	129	117
Israel, Japan, and South Africa	87	107	114	147	145	141
Developed regions	91	104	111	125	128	124
Latin America	97	103	99	104	101	105
Near East (excluding Israel)	90	104	104	108	109	105
Far East (excluding Japan and mainland China)	94	103	107	103	106	109
Africa (excluding South Africa)	96	101	102	95	96	94
Developing regions	94	103	104	103	104	106
World (excluding mainland China)	93	103	107	112	114	111

Source: UN Food and Agriculture Organization, *Monthly Bulletin of Agricultural Economics and Statistics* (January 1971).

with sales in excess of $100,000 were family farms, as were three-fourths of all farms with sales between $20,000 and $99,999. The relative importance of family farms among the largest farms in the United States has increased over the past two decades.

A system of tenant farming known as sharecropping developed in the South of the United States following the freeing of the slaves in the 19th century. It was an adjustment of the plantation system to permit the owners to maintain a large measure of control over farm operations. The sharecropper usually supplied only the labour, while the owner provided animal power, machinery, and most of the other inputs in the form of an advance. The cropper received what was left after he had paid back the owner—generally about half of what had been produced.

For various reasons, including the exodus of Negroes from American agriculture, the introduction of farm machinery, and the reduction in the acreage of cotton, the number of sharecroppers in the South diminished by 80 percent or more between 1935 and 1970.

In recent decades there has been a growth of large-scale farming run as a business enterprise. These "industrial farms" have not yet become of great significance in world agriculture. Farms covering extensive areas of land exist in Africa, South America, and Australia, but most of them do not rely heavily upon machinery or other purchased inputs. In the United States in 1964 there were 919 farms with total sales of $1,000,000 or more, accounting for about 7 percent of all farm sales. More than a third of such farms were livestock farms, primarily feed lots; vegetable and fruit and nut farms, livestock ranches, and poultry farms each accounted for approximately a tenth of the very large farms.

Strengths and weaknesses. If they were free to choose, most farm families would want to own the land they farm. Wherever collectivization of private farmers has been carried out, it has required the use of force or the threat of force. But if family farming is to be viable it must function efficiently, which means that farmers must have access to adequate sources of credit; must be able to obtain fertilizers, machinery, and other equipment; and must be able to market their produce easily. Laws and institutions must be sufficiently flexible to permit the average size of farms to increase as economic growth occurs.

Experience has shown that family farming can flourish under widely varying conditions: family farms in Japan are very small, while many in the United States dispose of $250,000 in capital or more.

Collective farming has not fulfilled the hopes of its early advocates. In the Soviet Union the collective farm was used by Stalin as a means of exploiting the rural population in order to finance industrialization. Since that time the incomes of collective farm members have increased, and it is believed that many remaining difficulties could be eliminated if the farms were given greater freedom in running their affairs. Nothing in the concept of the collective farm requires the imposition of delivery quotas, centralized control of farm investment, or a particular organization of farm labour. Another weakness of collective farms has been the failure to allow adequate incentives for individual members. The difficulties involved in rewarding members for their work on the common land have had the consequence that the household plots of the members tend to flourish at the expense of the collective.

There is no ideal form of organization that will fit all farming. Under some circumstances the ownership of land may absorb so much capital that other investments —machinery and livestock—are neglected. Land rental may be a better alternative for many families, especially those with limited capital. The Israeli kibbutz made it possible for many persons with little or no agricultural experience to learn farming techniques quickly and efficiently. The most important consideration is whether the other institutions—economic, political, and social—are adequate to provide farmers with a wide range of alternatives.

BIBLIOGRAPHY. T.W. SCHULTZ, *Transforming Traditional Agriculture* (1964), a work that has already become a classic on the complex problems involved in the modernization of agriculture; H.M. SOUTHWORTH and B.F. JOHNSTON (eds.), *Agricultural Development and Economic Growth* (1967), a collection of articles dealing with agricultural development in relation to social structure, education, farm size, land reform, food and nutrition, marketing, international trade, taxation, and price policy; ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT, *Agriculture and Economic Growth: A Report by a Group of Experts* (1965), a good summary of the interrelationships between agriculture and the rest of the economy in the course of economic growth, and *Agricultural Policies in 1966* (1967), a descriptive and analytical review of developments during the previous decade in Europe, North America, and Japan, with summaries of major agricultural policies of the OECD countries; L.R. BROWN, *Man, Land and Food: Looking Ahead at World Food Needs* (1963), a discussion of world land use, differences in the productivity of land among major regions, trends in grain production and food supply, and long-term trends in imports and exports of grain; V.W. RUTTAN, A.D. WALDO, and J.P. HOUCK (eds.), *Agricultural Policy in an Affluent Society* (1969), a collection of articles on all major aspects of farm policy in the United States.

(D.G.J.)

Agricultural Sciences

The agricultural sciences, dealing with the aims and problems of farm production, include soil cultivation, crop growing and harvesting, animal husbandry, and, to some extent, the processing of plant and animal products to a marketable stage.

Wherever scientific methods have been applied, they have revolutionized agriculture. In prescientific agriculture, six people could produce barely enough food for themselves and four others. In the frequent bad harvest years they could produce only enough for themselves. The application of scientific research and advanced technology has made it possible for five people to produce enough for 95 others. The farmer has been enabled to increase his yield per acre and per animal, reduce his losses from pests and spoilage, and maximize his net production by improving his processing. New techniques of preserving food products made it possible to transport them over greater distances, facilitating worldwide adjustment between production and consumption, with further beneficial effects on production efficiency.

Agricultural science deals primarily with two major types of crops: (1) those that represent direct human food, such as cereals, vegetables, fruits, and nuts; and (2) those that serve as forage and feed for cattle, sheep, horses, pigs and poultry. Special branches have developed to deal with such areas as citrus cultivation, ornamental plants, and viticulture. Certain branches deal with the cultivation of raw materials for industry; *e.g.*, cotton, hemp, sisal, and silk, some of which are losing importance in the face of the development of synthetics. Branches that deal with such tropical crops as coffee, tea, cocoa, bananas, coconuts, sugarcane, and pineapples, to the contrary, promise to retain their importance.

History. Early knowledge of agriculture was a collection of experiences verbally transmitted from farmer to farmer. Some of this ancient lore had been preserved in religious commandments. The traditional sciences, such as philosophy, theology, mathematics, and philology, rarely dealt with so commonplace a subject. Though Pliny and some other Classical writers discussed agricultural science, and much was written in the Middle Ages, it never gained a place in the academic structure; even the natural philosophers of the 18th century thought a general education sufficient for farmers. But at about the turn of the 19th century a movement began in central Europe to train farmers in special academies. The earliest were established at Keszthely, Hungary, in 1796, and at Hohenheim, Germany, in 1818. These efforts were empirically grounded; farmers were taught the experiences of other farmers. A truly scientific approach was inaugurated in 1840 by Justus von Liebig of Darmstadt, whose classic work, *Die organische Chemie in ihrer Anwendung auf Agrikulturchemie und Physiologie*, launched the sys-

Ancient agricultural lore

tematic development of agricultural sciences. The old empirical-training centres disappeared and were replaced by agricultural academies, colleges, and schools in all European countries and in the U.S. In Europe, a system of training soon developed that comprised four steps: (1) compulsory agricultural continuation schools providing basic knowledge in agriculture, attended at 15 to 18 years of age; (2) voluntarily attended day-release schools (usually one day per week) to train farm managers; (3) schools of engineering, normally attended for three years; and (4) college training for four to five years at colleges or special universities. Under Liebig's influence, the academic training concentrated for many years on natural sciences, but more recently in the countries with a highly developed agriculture, education has broadened to include more practical agricultural training. In some countries agricultural day-release schools are combined with farms; most agricultural colleges operate experimental farms.

Major divisions of agricultural science. The agricultural sciences are normally subdivided into five groups:

1. Soil science, which deals with the geological generation of soil, with soil physics, and soil chemistry, as they pertain to agriculture, and with all other factors relevant to soil fertility.
2. Plant production, which comprises applied plant physiology, plant nutrition, plant breeding (including genetics), weed control, plant diseases, and pest control.
3. Animal production, comprising animal breeding and genetics, animal nutrition, and animal husbandry. (In most countries the science of animal diseases is not counted among the agricultural sciences.)
4. Economics and management science, including behavioral science, agricultural policy, market research, and rural sociology.
5. Agricultural engineering, including appropriate areas of mechanical engineering, construction, hydraulics, and soil mechanics.

In addition to these five, several other studies are included in schools in various countries.

Soil science. Soil science began with the formulation of the theory of humus in 1809. A generation later, Liebig introduced experimental science, including a theory of the supply of soil with mineral nutrients. In the 20th century, the general theory of soil fertility developed, embracing soil cultivation, enriching soil with humus, enriching soil with nutrients (chemical fertilizers), and preparation of soil in accordance with crop demands. Water regulation, principally drainage, is also included. The study of soil science requires experimental demonstrations to show visually the different stages of soil evolution.

Plant production. Although plant production became a scientific discipline at the end of the 19th century, its beginnings are very old. Instructions on the dates of sowing are reported in Egypt by 2000 BC. Greek and Roman classics, medieval herbals, and 18th-century Rationalist treatises gave instructions on how to achieve higher yields. The impulse toward development of plant production into a science did not come from botany but from agricultural chemistry, whose application led to the development of plant physiology. Field experiments were started in Rothamsted, England, in 1834, and soon after, elsewhere in Europe. Improved methods of computation helped make possible the comparative study of plants, together with their varieties and cultivation systems. The science of plant nutrition matured between the two world wars. Cultivation of plant varieties, originally crossbred only between different areas of cultivation, led in the 18th century to the systematic selection of cereal varieties according to predicted yield. The discovery of Mendel's laws of heredity and of ways to cause mutations led to modern plant breeding, with momentous results including the breeding of crops for regions with climatic extremes (the development of short-stemmed wheat won a Nobel Prize in 1970). Major advances in the study of plant disease were recorded in the 19th century. Serious calamities, resulting from the introduction of plant diseases into districts without immunity against them and the invasion by such insects as the vine louse and fungi such as *Peronospora* in viticulture and the *Phytophthora* in potato cultivation, stimulated research efforts. Later, in the 20th century, all plant diseases became objects of systematic research.

Early field experiments

In pest control, two approaches can be observed: first, the search for chemical poisons effective against bacteria or insects, and, second, the adjustment of the biological balance to reduce losses from disease and pests. Chemical pest control often merely substituted new pests for old through the disturbance of the natural biological balance. Chemical residues also created problems, leading to a new field of scientific activity.

Animal production. The horse was the first animal used for crossbreeding. Instructions were written by the Hittites as early as 1600 BC. Long before a science of genetics developed, draft horses in medieval times and racehorses in the 18th century were subjected to selection and crossbreeding. Unlike plant breeding, animal production depends almost exclusively on crossbreeding; there has been little success with artificially caused mutations. Animal nutrition, which developed into a science in the 20th century, proved highly successful with livestock and poultry.

Animal diseases have been effectively combatted by culling and through the development of preventive drugs.

Economics and management science. Agricultural economics and scientific management have their roots in the German-speaking countries where, early in the 19th century, the individual farm enterprise was studied. In the U.K. and France, these studies tended to remain part of the general science of economics. The idea that a farm enterprise forms a unit, affected by location, production techniques, and market factors, originated in the 19th century. In the 20th century, it was supplemented by the theory of optimum utilization of production factors by the selection of production lines. Further refinement came about through applications of modern accounting methods. In the U.S., research into management problems was pursued. After World War II, mathematical planning systems, originally developed for military purposes, and statistical computation of farm enterprise data improved understanding of farm managerial problems. In recent years, interest has been drawn to behaviour studies of farm managers; it has been shown that even where, theoretically, the same rules of management were applied, results varied because farm managers made the application in different ways.

Mathematical planning systems

Agricultural policy is concerned with the relations among agriculture, economics, and society. Originally combined with agricultural economics and farm management, as in Germany, or regarded as part of general economics, as in the U.K., it became a separate study during the 20th century. Land ownership and the structure of farm enterprises were regarded as primarily social problems. The growth of agricultural production in the 20th century, accompanied by a decline in numbers of the rural population, gave a fresh impetus to research in agricultural policy, concentrating in the capitalist countries on the influence of prices and market mechanisms, and in the Communist countries on the artificially created market structure.

Market research originally had been limited to the problem of supply and demand, but the agricultural crises of the interwar years brought new analytic studies. In Europe, the growth of the cooperative movement, begun in Germany in the 19th century as a response to capital shortage and farm indebtedness, brought a satisfactory solution to the problem of distribution of products from farm to processor. Consequently, little interest developed in Europe in market research until after World War II. Today, the main areas of interest in this branch of agricultural science centre around statistical computations of past market trends to supply data for forecasting.

Rural sociology is a young discipline, though some of its problems were observed earlier by students of agricultural policy, but only after World War II did it become a scientific discipline. Intensively fostered in the U.S., it is now represented in almost every country. Methods of

research vary. Behaviour study, the most recent line of research, developed from studies in decision making in farm management.

Agricultural engineering. Although agricultural engineering did not become a separate discipline until the 20th century, engineers in the 19th century not only constructed useful works for irrigation and other purposes but made significant theoretical contributions. Agricultural engineering tended to focus on farm machinery, originally improvised by farmers, but by the 1930s manufactured industrially on a large scale. Often these machines did not withstand the heavy wear and severe conditions under which they were used. Various institutes undertook machine testing, which today is carried out principally by government organizations using international standards. New work methods developed in the 1930s by agricultural engineering were based on the use of instruments.

Agricultural engineering can be subdivided into material properties, machine design, and the application of technique. This last has been especially significant in livestock farming, with a growing use of actuating and control techniques, and the increasing use of electricity.

Other aspects of agricultural science. Agricultural work science originated during World War I. The existence of chronic rural overpopulation in all of the developed countries in those earlier periods made the rationalization of labour seem to be of little urgency. It was the study of social problems among the rural population in Germany that led to the beginning of agricultural work science and the establishment of special research and university institutes. Other European countries joined in after World War II. The improvement of work procedures, the appropriate use of labour, the analysis of human capacity for work, and the adjustment of mechanized production methods and labour requirements all represent the main objects of research. The influence of mechanization on the worker and the study of training of workers later came to be part of this branch of agricultural science.

Research and university institutes

Before 1930 writings on farm economics considered farm buildings to be merely parts of the farm property, and the age of the building to be more important than the way in which it was used. Improved standards of animal production brought a change, starting in the 1930s and growing rapidly after World War II in most European countries. In the United States, animal husbandry systems, especially their relation to different types of buildings, have been the main object of interest; while equipment, buildings, and methods of stock-raising are treated together in Europe. Economic construction methods and maximum animal performance are the principal goals in both regions.

In some countries, the techniques of land reclamation and water use also are part of agricultural engineering. In other countries, these disciplines are part of general engineering science. Drainage and irrigation are objectives of research and teaching within agricultural engineering, which now includes disposal of agricultural waste.

In some countries, processing technology, especially processing of animal products, such as milk, is included in the agricultural sciences. Such scientific disciplines as meteorology and geodesy, as well as law and history, also are frequently included.

BIBLIOGRAPHY. F. AEREBOE, *Beiträge zur Wirtschaftslehre des Landbaues* (1905), a representative publication of Aereboe, who made farm economics and management an independent discipline; G. FRANZ (ed.), *Deutsche Agrargeschichte*, vol. 1–4 (1962), vol. 5 ed. by H. HAUSHOFER, *Die deutsche Landwirtschaft im technischen Zeitalter* (1963), a general history of agriculture in Germany and Central Europe from the beginning to present times: G. FRANZ (ed.), *Die Geschichte der Landtechnik im 20. Jahrhundert* (1969), the most recent publication on the history of agricultural engineering during the 20th century; T. GOLTZ, *Geschichte der deutschen Landwirtschaft*, vol. 2, *Das 19. Jahrhundert* (1903), a comprehensive history of agriculture during the 19th century; H.C. and A.D. TAYLOR, *The Story of Agricultural Economics in the United States, 1840–1932* (1952), a history of developments in U.S. agricultural economics; UNITED STATES DEPARTMENT OF AGRICULTURE, *Bibliography of Agriculture* (monthly), a listing of agricultural publications from all over the world.

(G.P.)

Agriculture, History of

In spite of industrialization in many parts of the world, agriculture still drafts into its service more of the world's aggregate manpower than all other occupations combined. For information on agriculture in specific countries or regions, see articles such as CHINA; AFRICA; or WISCONSIN. For general information on agriculture in underdeveloped countries, see AGRICULTURAL ECONOMICS; ECONOMIC DEVELOPMENT; LAND REFORM AND TENURE; and ECONOMIC PLANNING. See also such articles as AGRICULTURE, TECHNOLOGY OF; IRRIGATION AND DRAINAGE; and PEST CONTROL.

This article is divided into the following sections:

I. Origins of agriculture
  Earliest beginnings
  Early development
II. Early agricultural societies
III. Agriculture in ancient Asia
  China
  The Indian subcontinent
  Southeast and east Asia
IV. Improvements in agriculture in the West: *c.* 200 BC to AD 1600
  The Roman epoch
  The medieval period: 600–1600
V. Crop-farming changes in western Europe: 1600–1800
  The Norfolk four-course system
  Improvements in technology
  Progress in different countries
VI. The 19th-century power revolution on the farm: *c.* 1792–1914
  Mechanization
  Applying science to farming
  Farming outside Europe
VII. Scientific agriculture: the 20th century
  Developments in power: the internal-combustion engine
  New crops and techniques
  New strains: genetics
  Animal breeding
  Electricity in agriculture
  Pest and disease control in crops
  Economics, politics, and agriculture

I. Origins of agriculture

Agriculture has no single, simple origin. At different times and in numerous places, many plants and animals have been domesticated. How many species passed into or out of domestication in prehistory is not known. Cultivation of foxtail millet in America and domestication of the elk in Scandinavia and the gazelle in the Near East were abandoned long ago. Today, cultivation of bottle gourds, finger millet, and Galla potatoes is on the decline, while efforts proceed to tame the eland, musk ox, and fox. Cultivation of rice, wheat, barley, potatoes, and maize met with great success in the favourable climate that followed the last Ice Age, while reindeer husbandry, by contrast, found these climatic changes unfavourable and lost importance.

EARLIEST BEGINNINGS

Nineteenth-century scholars hypothesized four stages in man's development: (1) a savage stage in which all men were hunter-gatherers; (2) a herdsman or nomad stage during which man domesticated some animals; (3) a farming stage; and (4) civilization. Botanical, archaeological, and zoological researchers have since attempted to determine when and where man first changed from hunter-gatherer to pastoralist or agriculturalist. It was long believed that agriculture began in a single centre in the Near East, about 4000 BC. Modern dating techniques have since disproved this hypothesis; they indicate agriculture already in progress around 7000 BC, and archaeologists now claim evidence of animal domestication thousands of years earlier. Recent work has also shown that some plants were probably cultivated in the

New World, that this development took place in many areas at once, and did not spread from a single originating centre.

Effect of domestication on animals

Many authorities today think man's domestication of plants and animals causes changes in their form and that the presence or absence of such changes may indicate whether the animal or plant was domesticated at some time in the past. On the basis of such evidence, some scholars hypothesize a preliminary agricultural phase of intensive food gathering in the Near East about 9000–7000 BC, when man passed from hunting and gathering to food producing or agriculture. The Natufians of Palestine, who possessed sickles, lived at this time; whether the grain they harvested was sown or wild is not known. At Shanidar, in Iraq, it is claimed that sheep, similar to wild varieties in form and structure, were kept in herds. Furthermore, it has been suggested on somewhat speculative grounds that einkorn wheat (*Triticum monococcum*), emmer wheat (*Triticum dicoccum*), and barley (*Hordeum spontaneum*) were cultivated about 7000 BC at Ali Kosh on the borders of Iraq and Iran.

There seems no compelling reason, however, why these instances should be regarded as the first of their kind. It is possible that domesticated beans (*Phaseolus*), peas (*Pisum*), bottle gourds (*Lagenaria*), and water chestnut (*Trapa*) may have been grown at the Spirit Cave in North Thailand about 7000 BC. In the Americas, pumpkins (*Cucurbita*) and gourds (*Lagenaria*) are known to have existed in domesticated form in northeast Mexico about 7000 BC, and probably beans in the Tehuacán Valley. The bones of a dog, possibly used for hunting about 8500 BC, were discovered in the western United States. In sum, it now seems unlikely that there was either a single or even a very limited number of places of origin of plant and animal domestication, and therefore of agriculture.

From E.C. Curwen and G. Hatt, *Plough and Pasture* (copyright 1953); Murnat Publications, Inc.

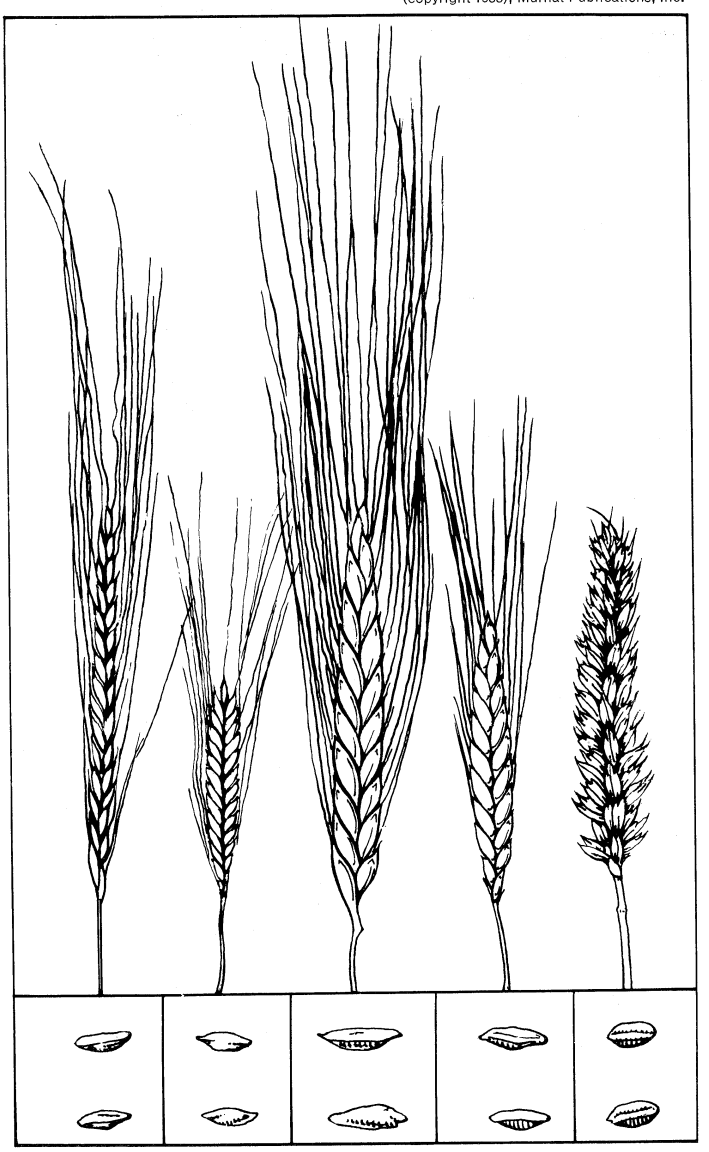


Figure 1: Wild and cultivated wheats. (Left to right) wild einkorn, einkorn, wild emmer, emmer, bread wheat. Scales: Ears, about ½; grains, about 9/10.

EARLY DEVELOPMENT

The development of agriculture was an intensification by man of his food extractive processes. More food could be obtained from a given area of land by encouraging plant and animal species found useful, and discouraging others. This provided for an increased population and gave better opportunity for settled life. Durable houses, as well as tools such as pestles, mortars, and grindstones, all of which had long been known in scattered places, came into more general use. There had even been villages in Stone Age times, among them the hunters' villages at Mezin in the Soviet Union, at Mallaha in Israel, and at Suberde in Turkey. These produced scant archaeological records, however; there exists no satisfactory evidence that they were inhabited by hunter-gatherers.

The Near East. In the 7th millennium before Christ, with heightened agricultural development, villages in the Near East became more numerous. Called the "effective" village stage of development, this period was accompanied by increasingly diversified agriculture, which led to more productive and efficient exploitation of available resources. The earliest known domesticated pigs have been recorded, in association with einkorn, at the village site of Jarmo (Iraq), dated about 6750 BC. The earliest known domesticated cattle date from about 6000 BC, at Argissa and Nea Nikomedeia, in Greece, in association with cultivated einkorn, emmer wheat, and lentils (*Lens esculenta*); and at Knossós on Crete in association with bread wheat (*Triticum aestivum*), emmer, and barley. Hoes or digging sticks were still used to break the ground where necessary. Seeding by treading in with flocks and herds was probably employed at this time. Techniques of food storage, a practice that man shares with many other animals, grew in sophistication; there were pit silos and granaries, sometimes of quite substantial nature. In drier areas, crop irrigation, which greatly increased yield, was developed; and, with the increasing population, more labour was available to carry out wider irrigation projects.

"Effective" village stage

Nitrogen-fixing (fertilizing) crops were also grown; a form of crop rotation came into use either by accident or by design. By this particular means, soil fertility was maintained, and thus additional plant protein was added to the diet.

The Americas. Although domesticates such as peppers, avocados, and amaranths were being grown in the Americas around 7000–5000 BC, villages were not numerous until widespread maize production began about 3500 BC. By then, common beans (*Phaseolus vulgaris*), chili pepper (*Capsicum annum*), and black and white zapotes were also widely cultivated. An early cereal, the foxtail millet, was probably domesticated around 4000 BC in Tamaulipas, Mexico, but was superseded by maize, and its cultivation was abandoned.

China. In China, millet was cultivated by the 5th millennium before Christ; rice was probably grown by the 4th millennium. Wheat was introduced before 1300 BC, but the arid nature of the uplands made progress slow. In the wetter lowlands, wheat could scarcely compete with already domesticated rice. Barley apparently was not introduced until after 1300 BC, nor was the soybean cultivated until perhaps 1100 BC.

Europe. In Europe, archaeologists have given so much attention to domesticates that are today outstandingly successful, such as wheat, barley, cattle, sheep, and pigs, that other species, possibly of major significance in the past, have largely been ignored.

It is generally assumed—apart from the earliest known

Old World domesticated dog at Star Carr in Yorkshire, England (about 7500 BC), cattle in Greece (about 6000 BC), and a possibility of domesticated pigs in the Crimea at an earlier date—that Europe prior to the introduction of agriculture from the East was inhabited by hunter-gatherers. It has indeed been suggested that the grain called fat hen (*Chenopodium album*) was cultivated at Iron Age settlements in Denmark, as well as Gold of Pleasure (*Camelina*) and *Polygonum lapathifolium*. Since cultivation of these grains is regarded as typifying a late form of agricultural development, it follows that agriculture must have taken about 3,000 years to spread from Greece to Denmark and the British Isles. Furthermore, considerable adaptations to new environments must have taken place in human technology and in the Asian species of sheep, goats, and cereals after their introduction into Europe. The effective village stage appears to have begun as soon as cereals and sheep or goats became important in Europe. In some areas, such as Hungary and Switzerland, many villages continued for thousands of years to rely upon indigenous animals for their animal protein.

Spread of agriculture in Europe

Along the Danube River, the great postglacial European forest was penetrated by slash-and-burn agriculture based on cereals. In the more arid regions along the Mediterranean coast, fewer modifications were necessary. The incorporation of indigenous wild stock, wherever available, into domesticated herds doubtless aided their acclimatization, for this practice continued into historic times, in Hungary until the invasion of the Mongols.

II. Early agricultural societies

In the Old World, settled life had developed on higher ground from Iran to Anatolia (site of modern Turkey) and the Levant, and in China in the semi-arid loess plains, but the earliest civilizations based on complex and productive agriculture developed on the alluvia of the Tigris, the Euphrates, and the Nile. When the first villages appeared there is not certain. It is known, however, that villages and townships existed in the Euphrates Valley in the latter part of the 5th millennium before Christ. Soon the population was dispersed over the available area in agricultural units based upon hamlets and villages. Townships provided additional services that the hamlets themselves could not provide. On this basis, Sumerian civilization developed.

Sumer. The Sumerian civilization was marked by an overall increase in wealth rather than by technical or agricultural innovations. The necessary foundations had already been laid. The Early Dynastic Phase began about 3000 BC. In Sumer, barley was the main crop, but wheat, flax, dates, apples, plums, and grapes were also grown. The earliest known white woolly, and therefore carefully bred, sheep and goats were herded and were more numerous than cattle. The sheep were kept mainly for meat and milk; butter and cheese were made. It has been estimated that at Ur, a large town covering some 20 hectares (50 acres) within a cultivated enclave, there were 10,000 animals confined in sheepfolds and stables, of which 3,000 were slaughtered each year. Ur's population of about 6,000 included a labour force of 2,500 who annually cultivated 1,200 hectares (3,000 acres) of land, leaving 1,200 hectares fallow. The work force included storehouse recorders, work foremen, overseers, and harvest supervisors, as well as labourers. Agricultural product was allocated to temple personnel in return for their services, to important people in the community, and to small farmers. The land was plowed by teams of oxen pulling light unwheeled plows, and the grain was harvested with sickles in the spring. Wagons had solid wheels with leather tires held in position by copper nails. They were drawn by oxen or the now-extinct Mesopotamian onager (wild ass), harnessed by collars, yokes, and headstalls, and controlled by reins and a ring through the nose or upper lip and a strap under the jaw. As many as four animals, harnessed abreast to a central pole, pulled a wagon. The horse, probably domesticated about 2700 BC by pastoral nomads in the Ukraine, did not in fact displace the onager in Mesopotamia until about 2300 BC, when bridles and bits were also introduced. Soon after, written instructions appeared for the grooming, exercising, and medication of horses; presumably for breeding purposes, horses were named and records of sires kept. The upper highland areas continued to be exploited by transhumant nomads.

Agriculture in the town of Ur

The Nile Valley. In Egypt, intensified agricultural exploitation apparently did not take place until domesticated animals from the Near East were introduced. By the first quarter of the 5th millennium before Christ, villages existed in al-Fayyūm that kept sheep, goats, and swine and cultivated wheat (emmer), barley, and flax, which was woven into linen. In this dry climate, village silos consisted of pits lined with coiled basketry; crops were harvested with reaping knives slotted with sharp flints. Elsewhere, at al-Badārī in Upper Egypt, animals were also kept; the fact that domesticated animals were wrapped in linen and buried close to villages perhaps indicates that agriculture was closely associated with some form of religious belief.

By predynastic Amratian times, about 3600 BC, agriculture appears to have begun in the valley alluvia of the Nile. By late predynastic times, about 3100 BC, there is evidence of a considerable growth in wealth consequent upon the earlier agricultural development, and accompanied by a more integrated social system.

From depictions on the tombs and other artifacts in dynastic times, it is known that, in addition to present-day domesticates, other animals, such as deer, gazelles, hyenas, and Barbary sheep, were kept either in captivity or under some form of control. Whether or not this man–animal relationship can be regarded as domestication is a matter of semantics, but certainly some aspects of animal husbandry were practiced with these unusual animals. Though it is not known how far back in time such practices existed, some early villages in Egypt were dependent upon gazelle for their source of animal protein, as was the township of Jericho, in Palestine, about 7000 BC. Indeed, such a dependence on gazelle is known to go back to Upper Paleolithic times in Palestine, and the question arises whether this was due to earlier forms of domestication. Recent discoveries at Nahal Oren, in Israel, indicate that it may well have been so. It has also been suggested that millet was a staple crop in early times in Egypt.

By the 4th dynasty, about 3000 BC, agriculture was developed and sophisticated. In contrast to Mesopotamia, where the tendency had been to develop urban cities, the inhabitants going out of the towns to work in the fields, in Egypt the cities tended to be no more than market towns to serve the surrounding countryside. A whole bureaucracy dealt with agriculture. The grand vizier, second only to the pharaoh, stood at its head, the ministry of agriculture under him. There was a chief of the fields and a master of largesse, who looked after the livestock. There were royal domains and temple estates. Between landlord and tenant there was a patriarchal relationship, which, although despotic, was underlain by a strong sense of responsibility to the land. Rent was three and a half bushels of grain to the acre. Irrigation and the waters of the Nile were carefully controlled. Seed corn was lent to tenants, and teams of oxen were lent or hired to them. The land was tilled with a wooden plow drawn by ox or ass traction (Figure 2). The land was plowed twice, once to break the ground, after which the clods were broken up by heavy hoes, and a second time to cover the seed. Six-rowed barley and emmer wheat were the main crops. The seed was sown by a funnel on the plow or, alternatively, was trodden in by sheep. The crops were cut with sickles, which had been improved by the introduction of a curved handle. The harvest produced 11 times the sowing, but it is not known whether or not two crops were grown within the year. The grain was threshed by asses or cattle treading on it on the threshing floor. It was winnowed by tossing in the wind, the chaff blowing away and the grain falling back into the basket, and stored in great silos (Figure 3). Lentils, onions, beans, and flax were other important Egyptian field crops.

Egypt's agricultural bureaucracy

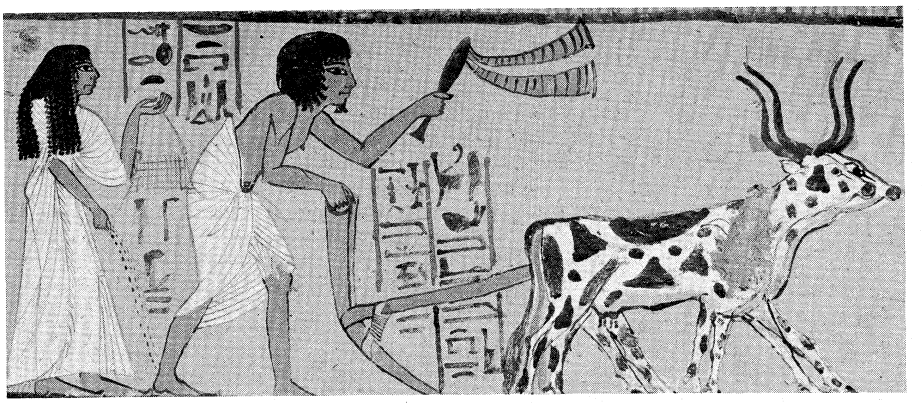


Figure 2: Plowing and sowing in Thebes. Painting from Tomb No. 1, Sennedjem, Thebes, Egypt.
Andre Held

Production of animal protein was also given attention. The wetter areas were exploited by domesticated ducks and geese. The marshes, swamps, wasteland, and stubbles were grazed by numerous herds of cattle, black, piebald, and white, sheep with kempy (coarse) coats, goats, and pigs. One wealthy landlord in the 6th dynasty owned 1,000 cattle, 760 asses, 2,200 goats, and 1,000 sheep. Animal breeding for specialized purposes was developed; one breed of cattle was kept for meat and another for milk; a Saluki-like hunting dog was bred, as were fat-tailed sheep with tails so heavy they were carried by a small cart drawn behind.

Meso-America and Peru. As far as is known, no animals were domesticated in the Americas until a late date. Available evidence seems to indicate that, in spite of the early domestication of some plants, village life did not begin to develop on any scale until 3500 BC or possibly somewhat earlier in Mexico, following domestication of maize. The process of agricultural development was therefore rather slow, occurring in widely dispersed centres, often in areas of poor fertility, sometimes even in deserts. Cacao (chocolate used to make a beverage), tomatoes, and avocado pears were cultivated. Irrigation, terracing, and the construction of islands in lakes increased land usage in drier areas. The land was cleared by chopping and burning, and the seed sown with the aid of fire-hardened digging sticks (Figure 4). Crops were stored in pits or granaries. The maize was prepared by boiling in limewater and by wet grinding. Cornmeal paste was then made into tortillas or flat cakes and gruel. Village life was based on the extended family, composed of parents and their children's families, which provided the labour force. Villages were organized into larger territorial units based on ceremonial centres commonly in the form of flat-topped pyramids. Fine textiles were woven of cotton and paper made from tree bark. Larger territorial units developed early in the 1st millennium AD.

Agriculture in Peru

Much less is known about the formative stage in Peru. Indeed, a civilization based on agricultural production may have begun there hundreds of years before it did in Meso-America, about the beginning of the Christian Era. Maize did not become important as a main crop in Peru until the 9th century before Christ, after perhaps 1,000 years of root crop, bean, and cotton agriculture. Maize was not extensively cultivated until irrigation developed; the potato, domesticated since about 2500 BC, was the chief crop in many areas until the 16th century AD.

At some time before the Inca civilization, the guanaco, ancestor of the llama and alpaca, was domesticated. The llama may have been domesticated as early as 2000 BC. Though the llama was essentially a beast of burden, the alpaca's main product was its fine wool. Another animal, raised for food, was the guinea pig. (E.S.H.)

By courtesy of the United Nations

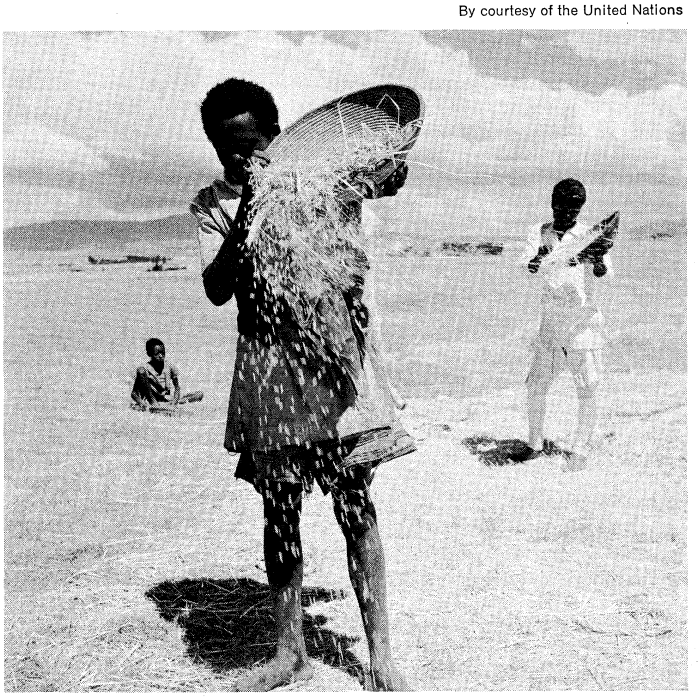


Figure 3: Winnowing grain by the traditional method of tossing it in the wind, Ethiopia.

III. Agriculture in ancient Asia

On his way across the Pamirs in search of Buddhist texts (AD 518) the Chinese pilgrim Sung Yun noted that the crest of the bare, cold, snowy highlands was commonly believed to be "the middle point of heaven and earth":

> The people of this region use the water of the rivers for irrigating their lands; and when they were told that in the middle country (China) the fields were watered by the rain, they laughed and said, "How could heaven provide enough for all?"

Yet, heaven provided. It has long supported more than one-half of humanity on less than a seventh of the earth's land surface along the continental rim to the south and east of the high interior mass with its festoons of mountains. In the so-called golden fringe to a beggar's mantle, from Pakistan through India, Burma, Thailand, Indochina, and eastern China up to the Gulf of Chihli, and the offshore island groups of Japan, Indonesia, the Philippines, and Ceylon (Sri Lanka), lives around 90 percent of the population of Asia. Over 1,000,000,000 people are concentrated in two countries, China and India.

There is no consensus on the origin and progress of plant and animal domestication. According to a leading geographer, N. Vavilov of the Soviet Union, however, of the world's eight independent centres of origin,

> An unusual wealth of original genera, species, and varieties of plants is found in India and China, countries which have contributed almost half of our crop plants. . . .

CHINA

Vavilov thinks that the mountainous regions of central and western China, together with the adjacent lowlands,

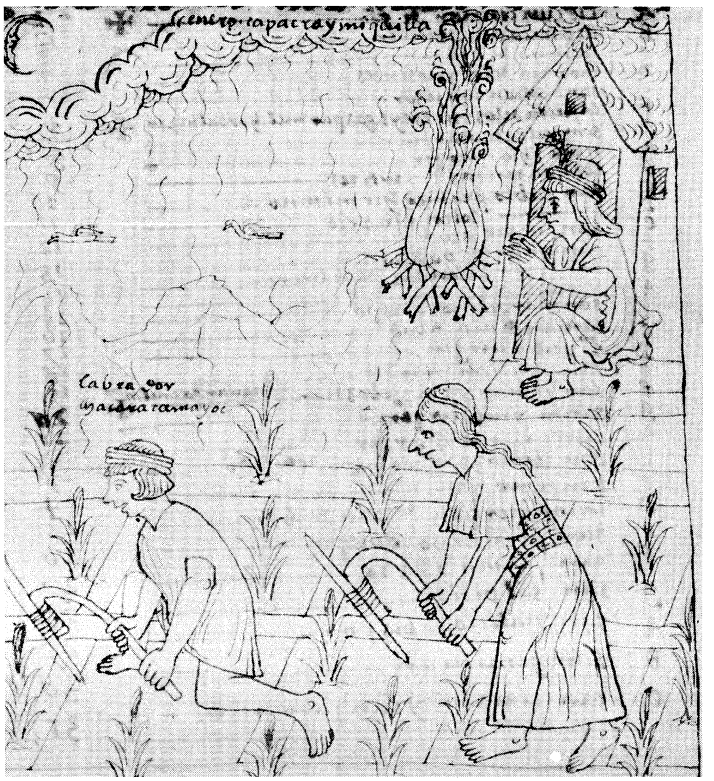


Figure 4: Inca cultivation with fire-hardened digging sticks. Drawing from "Nueva Crónica y Buen Gobierno" by Felipe Guaman Poma de Ayala. In Det Kongelige Bibliotek, Copenhagen.
By courtesy of the Kongelige Bibliotek, Copenhagen

constitute the earliest and largest independent centre of the world's agriculture. From earliest times, land use has been divided into two major regions by the Tsinling Shan, with wheat predominant in the northern realm, and rice in the south. At different periods and places, subsidiary crops have included soybean, kaoliang (a grain sorghum), millet, maize, barley, sweet potato, peanuts, fruits, and vegetables. Cotton, tobacco, sugarcane, tea, and sericulture (silkworms) have been the important cash crops.

Two growing areas of China

Early history. Archaeological information is comparatively scanty. It does not reveal when and how the first farmers settled down to cultivate food crops. Written records, available from about 1750 BC, are more conclusive. Tradition accords the honour of inventing agriculture to the legendary emperor Shen Nung, who is said to have carved a piece of wood into a plowshare and bent another piece to make a handle and taught the world the advantages of plowing and weeding. The chronicler Pai hu t'ung relates that he did so out of necessity, because "the ancient people ate meat of animals and birds." At the time of Shen Nung, there were so many people that the animals and birds became inadequate for people's wants, and therefore Shen Nung taught the people to cultivate. As for the origin of cultivated plants, the problem was resolved by divinity: "Millet rained from heaven. Shen Nung collected the grains and cultivated them."

Apparently the "millet from heaven" rained over the North China Plain in the central region of the Yellow River Valley, which is regarded as the birthplace of agriculture-based civilization in China. From there it spread throughout the country, but it failed to penetrate the northern grassland and plateaus. The Great Wall may be taken to demarcate the approximate frontier between the two basically different ways of life, the agrarian to the south, and the pastoral nomadic of the steppes and deserts on the other side.

The classical-imperial era. In about 335 BC, the Chinese philospher Mencius wrote that, "If the farmer's seasons are not interfered with, there will be more grain in the land than can be consumed." By the 1st century BC, however, wastelands were being reclaimed for cultivation, and there was a demand for limitation of landholdings. Nine years after Christ, the first (unsuccessful) attempt was made to "nationalize" the land and distribute it among the peasants. By the end of the 2nd century AD, severe agrarian crises culminating generally in the downfall of the ruling dynasty had become a recurrent theme of history. Through the centuries, Chinese agriculture has been a continual struggle between man and land to raise more food.

By the 4th century AD, cultivation was more intensive in China than in Europe or the rest of Asia. The major cereal-producing region and the population, however, were shifting rapidly from the wheat and millet area of the North China Plain to the paddies of the lower Yangtze Valley. By the 8th century, the lower Yangtze was exporting enormous quantities of grain into the old Northwest. To carry it, more than 800 miles (about 1,300 kilometres) of new canals were constructed, linking the large rivers into a unified water transport system.

Shift in major growing area

By about AD 1100, the population of South China had probably tripled, while that of the whole country may have exceeded 100,000,000. Consequently, cultivation became extremely intensive, with a family of ten crowded, for example, onto a farm of about 5.5 hectares (14 acres). Again, more new lands were opened to cultivation. Even tanks, ponds, reservoirs, streams, and creeks were reclaimed to be turned into farms. At the same time, complex water-driven machinery came into use for pumping irrigation water onto fields, for draining them, and for threshing and milling grain. A large variety of improved and complicated field implements were also employed; these are described and illustrated in the agricultural literature of the day.

Tools and techniques. The first significant revolution in Chinese agricultural technology had occurred when iron became the metal of the Chinese peasantry, finding its early and large-scale use in the fashioning of agricultural implements. The earliest iron plow found in northern Honan dates from the Warring States period (*c.* 481–221 BC) and is a flat V-shaped iron piece that must have been mounted somewhat insecurely on wooden blades and handles to serve as working edges. It was small, and there is no evidence that draft animals were used. Cattle-drawn plows do not appear until the 1st century BC.

Several improvements and innovations, such as the three-shared plow, the *lou-li* (plow-and-sow) implement, and the harrow, were developed subsequently. By the end of the Sung dynasty in 1279, Chinese agricultural engineering had reached a high state of development.

The common farmer continued to use these early medieval techniques into modern times. His unfenced fields were cultivated by a wooden plow, with or without a cast-iron share and usually drawn by a water buffalo. Harvesting was by sickle or billhook (a cutting tool consisting of a blade with a hooked point fitted with a handle). Sheaves were carried from the field slung at the ends of a pole across a man's shoulders. The grain was threshed by beating on a frame of slats or by flails on the ground. Winnowing was accomplished by tossing the grain in the wind. Rice was dehusked by hand pounding in a mortar or with a hand-turned mill. Irrigation techniques varied. The commonest perhaps was a wooden, square-paddle chain pump with radial treadle operated by foot. Fields were drained by open ditches and diking. Night soil, oil cakes, and ash fertilized the soil.

Over the past millennium, in fact, the revolution in Chinese agriculture has not been in mechanical or chemical technology but rather in the biological sphere: in crops, cropping systems, and land utilization. Under increasing population pressures, cultivation was forced to become more labour intensive and also to expand into the sandy loams, arid hills, and the upper reaches of lofty mountains. Lacking major technological inventions, the Chinese peasant had to expand the area under cultivation by finding suitable crops for inferior land.

Land use. A "three fields in two years" rotation system for wheat and millet was being practiced by the 6th

century AD. Revolutionary changes in land utilization, however, started with the introduction in Fukien of an early-maturing and relatively drought-resistant rice from Champa, in central Indochina. In 1012, when there was a drought in the lower Yangtze and Huai Ho regions, 30,000 bushels of Champa seeds were distributed. Usually a summer crop, the native rice plant required 150 days to mature after transplanting. Not only did this make a second crop difficult, but, because of the plant's soil and water requirements, cultivation was confined largely to the deltas, basins, and valleys of the Yangtze. The original Champa rice, on the other hand, ripened in 100 days after transplanting and required less water.

Introduction of Champa rice

The success of Champa rice initiated the development and dissemination of many more varieties suited to the local peculiarities of soil, temperature, and crop rotation. The first new early-ripening strain to develop required 60 days after transplantation. By the 18th century, there was a 50-day Champa and a 40-day Champa. By 1834, a 30-day variety was available—probably the quickest ripening rice ever recorded. The effect was revolutionary. By the 13th century, much of the hilly land of the lower Yangtze region and Fukien had been turned into terraced paddies. By the close of the 16th century, Champa rice had made double, and sometimes triple, rice cropping common.

A second revolution in land utilization began in the 16th century, with adoption of food crops from the New World, such as maize, sweet potato, potato, and peanuts. These could be grown at drier altitudes and in sandy loams too light for rice and other indigenous cereals. Virgin heights of the Yangtze region and North China were turned into maize and sweet-potato farms. As the population in the mountain districts increased, the potato took over the soils too poor for maize and sweet potato. By the middle of the 19th century, even ravines and remote mountains were being cultivated. Similarly, peanuts penetrated the remote and agriculturally backward areas of Kwantung, Kwangsi, and Yunnan, and the sandbars of Szechwan. Gradually, they brought about a revolution in the utilization of sandy soils along the lower Yangtze, the lower Yellow River, the southeast coast, particularly Fukien and Kwantung, and numerous inland rivers and streams.

Exploiting remote and unpromising regions

Even so, the revolution in land use failed to alter the basic man–land relationship in China. Again in the 18th century, the Emperor rejected demands for limitation of landownership. In an edict (1740), however, he noted that "the population is constantly increasing, while the land does not become any more extensive." He directed his subjects, therefore, to cultivate all and every odd piece of soil,

> on top of the mountains or at the corners of the land. All these soils are suitable either for rice or for miscellaneous crops. . . . no matter how little return the people may receive from cultivation of these lands, it will be always helpful in supplying food provisions for the people.

THE INDIAN SUBCONTINENT

India's most important contribution to world agriculture is rice (*Oryza sativa*), the staple food and crop of most of south, Southeast, and east Asia. Sugarcane, varieties of legumes, and tropical fruit like the mango also are native to the subcontinent.

Archaeological records of this region's prehistory are still incomplete and uneven. Evidence of how the transition from pastoralism to crop farming took place is lacking. Toward the close of the 3rd millennium before Christ, in the alluvial plains of the Indus in West Pakistan, ruins of the twin cities of Mohenjo-daro and Harappā reveal an apparently sudden explosion of an organized, sophisticated urban culture. This society, known as the Indus Valley or Harappan civilization, flourished until around 1750 BC, was more extensive than that of Egypt or Babylonia, and earlier than that of North China. Peculiarly Indian, it was remarkably homogeneous, thoroughly individual and independent, and technically the peer of China and Egypt. Nothing is known, however, about the farm communities that must have produced a surplus of agricultural products sufficient to sustain the population and fill the granaries of the metropolitan centres. At Harappā, for instance, the granary consisted of two blocks of six chambers each. The floor space of the 12 units covered an area of over 90,000 square feet (about 8,100 square metres).

Indus Valley civilization

Barley (*Hordeum vulgare*), a small-seeded, six-rowed variety, and two types of wheat (*Triticum compactum* and *Triticum sphaerococcum*) were evidently the main cereal crops, supplemented by dates, sesame, field peas, and lentils. Humped Indian cattle and a humpless breed, goat, sheep, fowl, and the elephant had been domesticated. Fragments of dyed and woven cotton fabric attest to the antiquity of the cultivation of cotton and to the textile industry for which India was to become famous the world over.

No archaeological or pictorial evidence of farm implements has survived, and the Harappan script is yet to be deciphered. It has been surmised, however, that the cereals could have been sown in the fall on inundated land after the annual flooding of the rivers had receded and then harvested in spring. The system is still in use and involves minimal skill, labour, and equipment. The land does not have to be plowed, fertilized, or irrigated.

It was clearly more than a subsistence economy, for there is proof of river and sea traffic. There was a trading post at Lothal on the Gulf of Cambay with a brick dockyard and elaborate channel and spillway. Two-wheeled bullock carts and light, covered wagons—both are still common—were used for local travel. Caravans of pack oxen were the principal mode of transportation over longer distances.

South India, centre of the later distinctive Tamil culture, constituted a second, initially independent agricultural region. Crops were being raised here during the first half of the 2nd millennium BC. Two varieties of pulses and finger millet (raggee) were cultivated.

To the north and west of the Deccan Plateau lay a third intermediate area. Here, at Lothal and Rangpur, has been found the earliest evidence of rice cultivation, in the later Harrappan period. Subsequently, wheat, cotton, flax, and lentils spread into the region from the Indus Valley, and pulses and millets from the south.

In all the three regions, except for rice, the basic cropping pattern of the 2nd millennium BC has continued into the 20th century.

Early historic period. A fourth agricultural region developed during the 1st millennium BC in the Gangetic Valley. Waves of invading Indo-Aryan tribes presumably destroyed the Harappan cities and penetrated deep into Hindustān proper, the principal seat of the classical Indian civilization. With the Aryans came the horse, coinage, the Brāhmī script, and the whole corpus of Vedic literature. Sources of information henceforth are literary rather than archaeological. The plow, for example, figures in a hymn of the most ancient of the texts, the Ṛgveda: "Harness the plows, fit on the yokes, now that the womb of the earth is ready to sow the seed therein. . . ."

Gangetic Valley culture

Apparently, rice played an important role in growth of population and new settlements. They had spread eastward to the Ganges Delta by the 7th century BC.

In the later Vedic texts (*c.* 1000–500 BC), there are repeated references to iron. Cultivation of a wide range of cereals, vegetables, and fruit is described. Meat and milk products were part of the diet; animal husbandry was important. The soil was plowed several times. Seed was broadcast. Fallowing and a certain sequence of cropping was recommended. Cow dung provided the manure. Irrigation was practiced.

A more secular eyewitness account is available from Megasthenes (*c.* 300 BC), a Greek envoy to the court of the Mauryan Empire. According to him:

> India has many huge mountains which abound in fruit-trees of every kind, and many vast plains of great fertility. . . . The greater part of the soil, moreover, is under irrigation, and consequently bears two crops in the course of the year. . . .
> In addition to cereals, there grows throughout India much millet . . . and much pulse of different sorts, and rice also, and what is called *bosporum* [Indian millet].

And again,

> since there is a double rainfall [*i.e.*, the two monsoons] in the course of each year. . . . the inhabitants of India almost always gather in two harvests annually.

Other sources reveal that the soils and seasons had been classified and meteorological observations of rainfall charted for the different regions of the Mauryan Empire, which comprised nearly the whole subcontinent and territory to the northwest.

A special department of the state supervised construction and maintenance of the irrigation system. The most famous work of the period is the Sudarshana Lake in Kāthiāwār, created by an artificial dam and furnished with conduits. Roads, too, were the government's responsibility. The swifter horse-drawn chariot provided greater mobility than the bullock cart.

The Mughal century (c. 1600). At the climax of the Mughal Empire, with the arrival and presence of the Western powers, a commercial economy based on oceanic trade was evolving. But no technological revolution in cultivating tools or techniques had occurred since roughly the time of the *Upaniṣads* (*c.* 600–300 BC).

The crops were broadly divided into rice zones and wheat and millet zones. Rice predominated in the eastern states, the southwest coast, and in Kashmir. Aside from its original home in Gujarāt, it had spread also to the Punjab and Sind with the aid of irrigation. Wheat grew throughout its "natural" region in north and central India. Millets were cultivated in the wheat areas and in the drier districts of Gujarāt and Khāndesh as well.

Cotton, sugarcane, indigo, and opium were major cash crops; the last two have since passed out of cultivation. Introduced by the Portuguese, cultivation of tobacco spread rapidly. The Malabar coast was the home of spices, especially black pepper, that had stimulated the first European adventures in the East. Coffee had been imported from Abyssinia and became a popular beverage in aristocratic circles by the end of the century. Tea, which was to become the common man's drink and a major export, was yet undiscovered, though it was growing wild in the hills of Assam. Vegetables were cultivated mainly in the vicinity of towns. New species of fruit, such as the pineapple, papaya, and cashew nut, also were introduced by the Portuguese. Quality of mango and citrus fruits was greatly improved.

Spices, coffee, tea, and fruit

Cattle continued to be important for draft and milk. According to a Dutch observer, however, the cows gave far less milk than in his country. Land use never became as intensive as in China and the Far East, although, as noted by Megasthenes, double (and even triple) cropping was fairly common in regions favoured with irrigation or adequate double rainfall. Though the population must have increased many times over since Mauryan times, in the 17th century, virgin land was still abundant; peasants were scarce.

Irrigation from wells, tanks, and canals, however, had greatly expanded. Some new water-lifting devices, such as the Persian wheel (*sāqīyeh*) or a series of leather buckets on an endless rope yoked to oxen, had been adopted and are still widely used.

The plow was the principal implement for tillage. Drawn by oxen, the traditional Indian plow has never had a wheel or moldboard. The part that penetrates the soil is a wedge-shaped block of hardwood. The draft pole projects in front, where it is attached to the neck yoke of the bullocks. A short, upright stilt in the rear serves as a guiding handle. The point of the wedge, to which an iron share may or may not be attached, does not invert the soil. Some plows are so light that the cultivator can carry them daily on his shoulder to and from the fields. Others are heavy, requiring teams of four to six pairs of oxen. Levellers and clod crushers, generally consisting of a rectangular beam of wood drawn by bullocks, are used to smooth the surface before sowing. Among hand tools, the commonest is the *kodali*, an iron blade fitted to a wooden handle with which it makes an acute angle. The Western-style spade or shovel has never been used in the subcontinent.

Drill sowing and dibbling (making small holes in the ground for seeds or plants) are old practices in India. An early-17th-century writer notes that cotton cultivators "push down a pointed peg into the ground, put the seed into the hole, and cover it with earth—it grows better thus." Another simple device was a bamboo tube attached to the plow. The seed was dropped through the tube into the furrow as the plow worked and was covered by the soil in making the next furrow.

Operations of reaping, threshing, and winnowing have remained almost exactly as described in the Vedic texts. Thus, grain is harvested with a sickle. It is bound in bundles and threshed by bullocks treading on it or by hand pounding. To separate the grain from the chaff, it is either sieved with sieves made of grass or bamboo, or, if there is a strong wind, it is winnowed by pouring by hand at a height from a *supa* (winnowing scoop). The grain is then measured and stored. The sickle, sieve, and *supa* remain today as they were more than two millennia ago.

SOUTHEAST AND EAST ASIA

The Chinese agrarian system could not conquer the northern steppes and deserts nor the barren heights of the Tibetan plateau to the west. It did reach out, however, to the plains of Manchuria, Inner Mongolia, the lower valleys of eastern Tibet, and Chinese Turkistan, wherever climatic conditions were favourable. It encountered no comparable barriers to the east. As a result, crops and cultivation techniques, especially of rice culture, spread to Korea and Japan, where they enjoyed notable success. Japan achieved higher yields per acre of rice than did China and most of the world.

Southeast Asia, on the other hand, constitutes an agricultural subcentre supplementing that of the Indian subcontinent. Though it has a wealth of tropical wild and cultivated flora, rice became the staple and predominantly subsistence crop of the region after its introduction probably by Indian colonists around the middle of the 1st millennium. Most of the other key farm products, such as sugar, tea, coffee, cotton, rubber, and maize, also were introduced from the outside by the European colonizers.

Shifting cultivation

Historically, shifting cultivation has been native to all the countries and still survives. As in south Asia, most of the rice is rain fed, and multiple cropping is infrequent, probably because of relatively low population densities in most of the region. With the exception of small pockets, as in Java and the Red River Delta of Tonkin, even wet-rice cultivation techniques have never been as intensive as in South China and Japan; nor have the yields been nearly as high. (K.Na.)

IV. Improvements in agriculture in the West: c. 200 BC to AD 1600

THE ROMAN EPOCH

Crop farming and domestication of animals were well established among the scanty population of western Europe by Roman times. Greek and Roman farming techniques are known from contemporary textbooks that have survived. Methods were dictated to some degree by the Mediterranean climate and the contours of the area. The majority of the crops cultivated today on the Mediterranean coast, wheat, spelt, barley, some millet, the legumes, beans, peas, vetches, chick-peas, lucerne (alfalfa), and lupines, were known at that time. The olive, the vine, and fruit trees were grown, as were turnips and radishes.

The farm. Roman holdings were commonly as small as 1.25 acres, or one-half hectare; the ground was prepared with hand tools, hoes, and mattocks, doubtless edged with bronze or iron. Later, as farming developed, and estates of different sizes came into existence, two writers set out catalogs of the tools, implements, and labour required to exploit a given-size holding. These were Cato the Censor (234–149 BC) and Varro (116–27 BC). Already in Cato's time, emphasis was on production of wine and oil for sale, rather than cultivation of cereal crops, beyond the volume required to feed animals and slaves.

For an olive grove of 240 jugers (150 acres, or 60 hec-

Equipment and labour for an olive grove

tares), Cato estimated necessary equipment as three large carts, six plows and plowshares, three yokes, six sets of ox harness, one harrow, manure hampers and baskets, three pack saddles, and three pads for the asses. Required tools included eight heavy spades, four smaller spades, shovels, rakes, scythes, axes, and wedges. Some 13 people, including an overseer, housekeeper, five labourers, three teamsters, a muleteer, swineherd, and a shepherd responsible for 100 sheep, did the work. Other livestock included three yokes of oxen, three donkeys to carry manure, and one for the olive-crushing mill. The farm was also equipped with oil presses and containers for the oil.

Farm implements. Most Roman-era hand tools were similar in shape to their modern counterparts. The wooden plow was fitted with an iron share and, later, with a coulter (cutter). Though it had no moldboard to turn the soil over, it was sometimes fitted with two small "ears" that helped to make a more distinct rut. Though it could not turn a furrow, it could invert some of the soil if held sideways. It was usually followed by a man with a mattock who broke up clods and cleared the row so seed would fall into it. Two or three such plowings were given each year to land intended for cereals. Manure was spread only after the second plowing. If spread earlier, it would be buried too deep to do any good. The farm included a compost pit where human and animal excrement were placed along with leaves, weeds, and household waste. Water was added from time to time to rot down the mass, and an oak pole was driven into the middle to keep snakes away. Various animal and bird droppings were believed to have different effects on growing plants. Pigeon's dung was valued, but that of aquatic birds avoided. Marl—earth containing lime, clay, and sand—was used in Gaul and possibly Britain.

Seeds were sown by hand, broadcast, or dropped. They were covered with a harrow, which may have had iron teeth or may simply have been a thornbush. A more complex plow, fitted with a wheeled forecarriage, may have been used in Cisalpine Gaul (north Italy) as early as the 1st century after Christ. Traction normally was supplied by a pair of oxen; the Roman historian Pliny the Elder (AD 23–79) mentions as many as eight being used on heavy land. In light soil, only one was necessary, and sometimes asses were used.

Cropping systems. Olive groves and vineyards were permanent; grain and pulses were annuals. Although it was realized that different soils were better suited to some crops than to others, the same piece of land was used for all crops. A specific crop, however, was grown in alternate years in what is known as the crop and fallow system. The fallow land was plowed two or three times during the fallow year to kill the weeds, which typically accumulate where cereal crops are continuously cultivated. Wetland was drained by digging V-shaped trenches, the bottom of which, usually four feet (1.2 metres) deep, was paved with loose stones, willow branches, or bundles of brushwood placed lengthwise, and covered with the replaced soil. Soil was judged by colour, taste, smell, adhesion to the fingers when rubbed, and whether it filled up a hole from which it had been dug or proved too loose.

Then as now, wheat was mostly sown in autumn, though a species known as *Triticum trimestre* was sometimes planted in spring; it ripened in three months. Barley was a spring-sown crop, as were most others. Though the Romans knew that growing lucerne and clover was in some way good for the succeeding crop, they did not know why. Similarly, a crop of lupines was sometimes planted for plowing in as green manure, and occasionally a crop of beans was used in the same way.

Harvesting and processing. The harvest was reaped with a curved sickle, a tool which has changed little since Roman times. In some places, the ears of grain were cut and carried in wicker baskets to the threshing floor. The straw was cut and stacked later. In other areas, the plant was cut lower down, and the grain was threshed from the straw. Another set of tools was used, consisting of a short-handled sickle held in the right hand, with the blade at right angles to the handle. A short-handled hooklike implement held in the left hand was used to draw together enough grain to be cut at one stroke. In Gaul, a reaper was used, a cart with an open back pushed by an animal reversed in the shafts. On the edge of the back, a comblike device was fixed to tear off the ears as the vehicle was pushed through the crop. The grain was threshed in the long-established way, by animals treading it on a firm floor, or by an implement known as a *tribulum*, a wooden framework with bits of flint or metal fixed to the underside, hauled over the grain by an animal. Winnowing was still done by tossing in the air from a winnowing basket when there was a favourable wind to blow away the chaff.

The quern

Corn was ground with a quern, a hand implement made of two stones, a concave base with a convex upper stone fitted into it. Some querns turned in a circle, while others merely rubbed up and down on the grain. Though designed before the end of the Roman period, watermills were uncommon.

Some forage crops were necessary to feed the plow animals and the cattle, sheep, and pigs. Grass was cut for hay, and many hours must have been spent in the woods collecting acorns for winterfeed for the swine. Lucerne (alfalfa) was the best fodder; it helped fertility as well. Lupines and a mixed crop of beans, vetch, and chickpeas and another mixture of barley, vetch, and legumes were also employed. Turnips were grown for man and beast in some regions, notably Gaul.

Livestock. Little attempt was made at selective breeding, and little was possible, for most of the animals spent their time at open range or in the woods. Nevertheless, different breeds of cattle were recognized as native to particular places. They were bred between two and ten years old; two bulls to 60 or 70 cows was the correct proportion. Sheep were important; Greek shepherds garbed some of the very fine wooled sheep in skin coats to keep their fleece clean. Ewes were served at three years old, two if essential. They fed on the stubble after harvest. Transhumance, or seasonal migration in search of pasture, was normal. A supply of clear water near the grazing ground was necessary.

Swine raising

Goats were kept in large herds, with 50 to 100 the optimum. Swine were also important. Very fat animals were preferred, and large numbers of these, whose meat was frequently seen on the Roman table, were kept. Sows were covered (bred) at from 12 to 20 months of age; it was desirable for them to pig in July or August. The best proportion of boars to sows was ten to 100. Herds of 100 to 150 ranged the woods. The bacon produced in Gaul had a reputation for quality; swine also flourished in north Italy and eastern Spain.

The methods of the Roman farmer produced only limited yields, and cereals were regularly imported to Italy from lands more naturally favourable to grain growing: Egypt, Sicily, Sardinia, and Gaul. Yet the Roman methods were basically sound and, with the help of modern mechanical aids, remain to a large extent in force today.

THE MEDIEVAL PERIOD: 600–1600

In the 1,000 years of medieval history, many details of farming in the Western world changed; a surprising number remained much the same. The period falls into two divisions: the first, one of development, lasted until the end of the 13th century; the second, a time of recession, was followed by two centuries of recovery.

Agricultural advances. The most important agricultural advances took place in the countries north of the Alps, in spite of the large population changes and warfare that accompanied the great migrations and the later onslaughts of Northmen and Saracens. Agriculture had, of course, been practiced regularly in Gaul and Britain and sporadically elsewhere in Europe both before and during the Roman epoch. The climate and soils and, perhaps, the social organization compelled different arrangements of land division and the use of more complex tools as men reclaimed more and more land from forest, marsh, and heath to meet the needs of a rising population.

Open-field system. The precise origin of the open-field arrangement, which involves long scattered strips of arable land separated from each other by a furrow, balk (ridge of land left after plowing), or mere (boundary), is obscure (Figure 5). The earliest examples of this system

By courtesy of the R.A.F., Crown copyright reserved

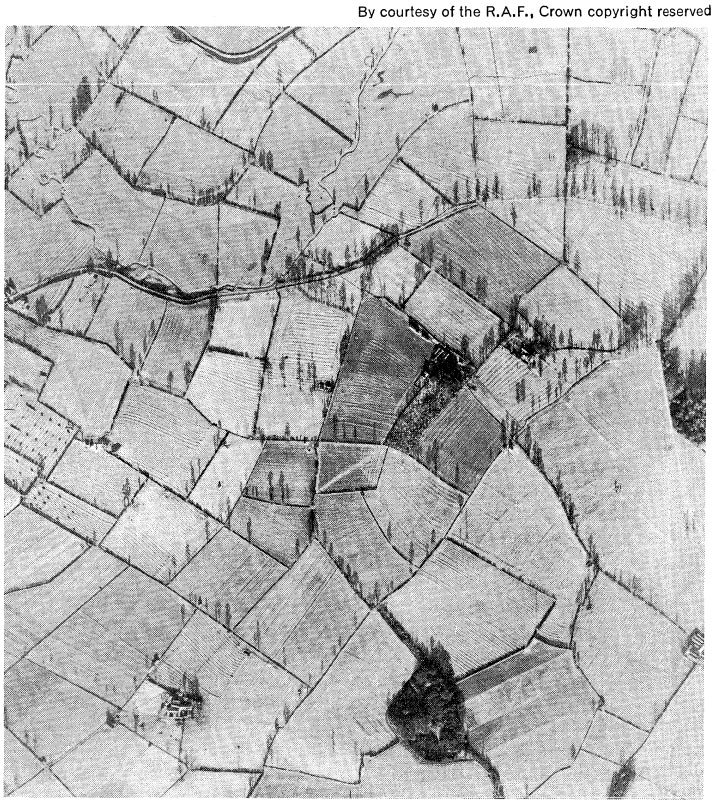


Figure 5: Ridges of the English open-field system as seen today in Leicestershire.

date from about 800, the year Charlemagne was crowned emperor of the West. Usually these strips of land, normally about an acre (0.4 hectare) in size, were laid out in two or three large fields. Each farmer in the village worked a number of these acres; the units forming his holding were scattered among those of other men. The open-field system continued as more land was reclaimed and lasted for many centuries—longer, of course, in some places than in others. Some say the length of each strip was determined by the distance a draft animal, usually an ox, could haul a plow before stopping for a rest. The intermingling of the strips is said to have been the result of a jointly owned plow team and plow supplied by a number of farmers working together, each being allotted a strip in turn. A recent theory holds that in some places the division of fields, which may have originally been rectangular or square, among a number of heirs led to the creation of long, narrow acres.

In theory each man's holding totalled 30 acres (12 hectares), comprised of strips equally divided between the three arable fields. With the passage of time, however, wide variations in the size of holdings came about; many became very small.

Plows and plowing. Besides the different arrangement of the plowland, there were other changes, some of them important. Though Pliny the Elder claimed a wheeled plow was used in Cisalpine Gaul about the time of Christ, there is a good deal of doubt about this. A wheeled asymmetrical plow was certainly in use in some parts of western Europe by the late 10th century. Illuminations of manuscripts and calendars a little later in date show a plow with two wheels fitted with a rudimentary moldboard and a coulter (Figure 6). This plow could invert the soil and turn a true furrow, thus making a better seedbed. Its use left high ridges on the land, traces of which can still be seen in some places.

The horse collar

The horse collar, which replaced the old harness band that pressed upon the animal's windpipe, severely restricting its tractive power, was one of the most important inventions in the history of agriculture. Apparently invented in China, the rigid, padded horse collar allowed the animal to exert its full strength, enabling it to do heavier work, plowing as well as haulage. Many peasants continued to use oxen, however, because horses were more expensive to buy and to keep. Some plowing was done by two oxen as in former times; four, eight, or more were occasionally necessary in very difficult land.

Hand tools. Modifications, slight but important, had been introduced into the design of hand tools. A more effective ax made forest clearance easier and faster. The jointed flail supplanted the straight stick. The scythe was more frequently in use for mowing grass, reaping barley, and similar tasks. The power of the wind was applied to the grinding of grain by the earliest windmills. All these changes and adaptations helped expand the cultivated area and supply food for the growing population.

New lands and crops. Not only were forests cleared and heavy land cultivated, but, in the Netherlands, reclamation from marsh and from the sea was extended. In the midst of the marshes, terps, artificially made patches of higher land on which houses and barns could be built, were made at a very early date. Ditches to drain the fens were dug in the 10th century. Polders, land reclaimed from the sea, are first recorded in the 12th century.

In Spain, the Moors had introduced new crops and a new breed of sheep, the Merino, that was to make Spanish wool famous all over Europe. New crops included sugarcane, rice, cotton, and some subtropical fruits, especially citrus. The vine and the olive flourished in the south, as did the vines the Romans had introduced to the valleys of the Moselle and Rhine.

Widespread expansion of farmed land occurred all over western Europe between the 10th century and the later years of the 13th. German and Dutch settlers were encouraged to take up holdings eastward toward the Baltic countries and south to the Carpathians. In France, new villages were built and new farms carved out of the forest and the waste, while, in England, a great deal of land on the boundaries of the open fields was taken in and cultivated. All this new cultivation was carried out with the same old implements and tools; the same crops were

By courtesy of the Victoria and Albert Museum, London



Figure 6: "The Ploughman" behind a two-wheeled plow drawn by horses. Wood engraving from the "Dance of Death" by Hans Holbein, the Younger, c. 1540. In the Victoria and Albert Museum, London.

cultivated and animals bred as before. In remote and desolate places, monastic organizations created great estates. These estates were formed to feed growing populations rather than to improve technical skills. A new literature of farming arose, directed to the attention of great lords and ecclesiastical magnates rather than to the illiterate majority of husbandmen. Unfortunately, the bright prospects were dimmed in the 14th century by a combination of calamities.

Agricultural recession. What is now called a recession began toward the end of the 13th century. The disasters of the 14th, climatic, medical, and military, followed. Famine resulted from excessively bad weather in 1314, 1315, and 1316; a small recovery followed in 1317. Yields, never high (from six to ten bushels of wheat per acre, or about 120 to 200 litres per hectare; a little more for barley, rye, and oats), were reduced to nothing by the weather. Floods wiped out the reclaimed land in the Netherlands. Pestilence followed famine, bringing suffering both to animals and men. The Black Death broke out in 1348 and is estimated to have killed from one-third to one-half of the population in much of Europe. Renewed outbreaks followed in 1368–69, 1370, 1381–82, and 1396. The Hundred Years' War desolated much of France; other conflicts, accompanied by similar pillage and destruction, broke out elsewhere. The result of all these misfortunes was to be seen in the landscape all over western Europe. Much of the arable land could not be cultivated for lack of workmen; in some regions the countryside was inhabited by a few scattered peasants grubbing a scanty living in the grimmest isolation. Many of the new settlements and some of the old disappeared and became deserted villages.

Effects of the Black Death and Hundred Years' War

The Netherlands. The Netherlands were not as seriously affected as most other countries. The flood destruction was repaired, and a system developed that was to become an example to all Europe. Leguminous and root crops were introduced into the rotation at least as early as the 15th century, and long continuous rotations almost without fallow breaks were employed. Town refuse was added to the supplies of animal manure. The size and milk yield of the Dutch cattle became famous, though possibly exaggerated. Some say that they owed part of their distinction to crosses with animals from Lombardy and Piedmont, which also enjoyed a great reputation. Flemish horses were already renowned for size and strength.

England. In England, when agricultural recovery began in the 15th century, there was no immediate improvement in technique. The profits of the wool trade and, later, of the woollen manufacture induced landowners to increase the size of their flocks. This led to some difficulties. Not only had some arable land fallen down to rough grazing owing to labour scarcity after the diseases and bad seasons of the 14th century, but the profit of wool encouraged enclosure of formerly open fields for grazing; actual destruction of some villages had occurred to increase the area of grazing land. Though there was a considerable outcry against enclosure in the 16th and early 17th centuries, the practice was too profitable to halt. At the same time, farmers began exchanging their scattered plots of land in order to consolidate individual holdings. These consolidated plots were then enclosed with a hedge or fence to prevent subjection to regulation that governed use of the remaining strips. Land was also acquired by purchase for this purpose. None of these changes involved any technological advances in farming, however.

Spain. In Spain, the shepherds, organized in the Mesta, a powerful body with great political influence, came into conflict with the farmers. The annual journey of the sheep from their northern grazing area to the south carried them along an established route, which steadily broadened, with the sheep trespassing upon the farmers' lands and consuming crops. At the same time the Mesta successfully opposed any expansion of the arable land until the mid-16th century.

Italy. In the 14th century, the city-states of Italy were devoted to commerce. There was little emphasis on farming, though some attempts at draining marshes were made, and, in spite of the introduction of rice culture in the north, Italian farming on the whole remained much as it had been in Roman times. In the south great flocks were kept and moved up to the mountains for the summer along well-defined paths.

France. In the 15th century, French farmers made substantial progress toward recovery, but even in France there was little advance in technology. The open-field system was prevalent in the north, and a type of Roman farming suited to the environment was practiced in the south, with lucerne, clover, lupines, and other legumes grown for fodder and to maintain fertility. A fodder crop called Burgundy grass was grown in Burgundy toward the end of the 16th century.

Germany. Many of the German villages depopulated by the disasters of the 14th century were never resettled. Some of them had been established on marginal land, such as sandy heaths, or high up in the mountains. By the middle of the 16th century, the advanced farming of the Netherlands penetrated into the north at the mouth of the Rhine and in Schleswig-Holstein. This is clear from one of the earliest printed books on farming, by a German, Conrad Heresbach. Heresbach described and recommended many of the methods used by the Romans: lupines for green manuring and a rotation of fallow-manured, winter-sown rape with wheat, rye, and spring barley. For the preparation of the seedbed, the destruction of weeds, manuring, sowing, and harvesting, implements that derived from the Roman pattern were used.

Heresbach's book followed somewhat the pattern of Crescentius, who wrote in the 13th century, and in that respect was similar to the growing number of agricultural treatises that were appearing in Spain and France. These were often encyclopaedias of rural life presumably intended for the landowning public. In the late 16th century, Henry IV of France and his minister, Sully, tried to stimulate interest among the lesser nobility in the management of their estates. In England, translations were made of continental works.

V. Crop-farming changes in western Europe, 1600–1800

THE NORFOLK FOUR-COURSE SYSTEM

Of the many changes that took place during this period in agricultural history, few were more important than the Norfolk four-course system, characterized by the disappearance of the fallow year and by a new emphasis on fodder crops. The movement toward change was further intensified by the invention of new farm machines, improvements in farm implements, and scientific interest and new biological theories relating to farm and animal life.

In the Norfolk four-course system, wheat was grown in the first year, followed by turnips in the second, then barley, with clover and ryegrass undersown, in the third. The clover and ryegrass was grazed or cut for feed in the fourth year. The turnips were either employed for feeding cattle in open yards during the winter (some covered yards were built) or for feeding sheep confined in folds set up on the ground. This new system was cumulative in effect, for the fodder crops eaten by the livestock produced large supplies of previously scarce animal manure, and that was richer in nature because the animals were better fed. When the sheep grazed the fields, their urine and droppings fertilized the soil, so that heavier cereal yields were obtained in following years.

Established in Norfolk County, England, and in several other counties before the end of the 17th century, the Norfolk four-course system became fairly general on the newly enclosed farms by 1800, remaining almost standard practice on most British farms for the best part of the following century. The system was used in the Lothians and some other parts of Scotland by about 1800 and, during the first three quarters of the 19th century, was adopted in much of continental Europe.

Enclosure. In order to adopt the Norfolk four-course system, it was first necessary to alter the thousand-year-

old layout of the arable fields. It was virtually impossible for an individual farmer to grow fodder crops on his strips of land in open fields, for at certain seasons (after harvest for example) these fields were opened to grazing by the livestock of the whole community. The improving farmer, who grew clover and ryegrass or other legumes or a root crop, would simply have provided additional feed for his neighbours' as well as his own animals. Such an arrangement was possible, of course, if all the farmers cooperated, a rather unlikely but not absolutely unknown state of affairs. On enclosed land, however, a farmer could cultivate these crops and benefit from his own efforts.

Consequently there was a rapid acceleration of the enclosure movement in England, sometimes by local agreement, by Chancery decree (late 17th century), and by private acts of Parliament in growing numbers (18th century). Some 6,000,000 acres (2,400,000 hectares) of English fields were enclosed between 1700 and 1845. On the Continent, the change, where it was made, took place slightly later. Still today, many peasant farms are worked in strips, though communal regulation has vanished.

Livestock. Where new fodder crops were introduced into the rotation, more and better livestock were bred and kept. Animal size increased, and meat development was greater. In cows, milk yield improved. Better breeding increased fleece and flesh production in sheep and improved the size of pigs. Horses, essential to transport, were carefully bred. Unfortunately, there were disastrous outbreaks of cattle plague that spread throughout Europe in the second and fourth decades of the 18th century, possibly introduced from eastern Europe by cattle driven to markets in the west. As a result of these calamities, however, a veterinary school was founded at Lyons, France, in 1762, and a second at Alfort, France, in 1767. Denmark established its first veterinary school in 1773, and England followed in 1791.

Spread of the Merino sheep

An interesting aspect of livestock breeding was the admiration of all Europe for the Spanish Merino sheep, which were forbidden from export for fear of competition. It proved impossible to maintain the prohibition. Merinos were obtained by Louis XVI of France in 1786. Some had been imported into Sweden even earlier, before 1765. Both Prussia and Saxony raised them by mid-century, the Saxon Merino becoming famous. Austria also secured Merinos, and George III of England obtained a few by the end of the century. Merinos were not a success in England, however.

Though a major source of meat since Roman times, pigs had been raised half wild, feeding in the woods under the care of a swineherd and his dogs. Over a large part of western Europe, this system continued, especially where open-field farming prevailed. In England, however, where population doubled during the 18th century from about 5,000,000 to about 10,000,000, more care was taken in breeding productive swine. Several different breeds could already be distinguished; some were crossed with Chinese animals possibly brought from that distant country in vessels carrying spices and tea.

IMPROVEMENTS IN TECHNOLOGY

Plows and farm machinery. Probably of Dutch origin, the Rotherham plow, the main design of which has remained virtually unchanged to this day, was first put into use in the Netherlands, England, and Scotland during the first half of the 18th century. Details of design varied from area to area. Some Rotherham plows were wheeled, and some were not; some were made of timber; others of iron. The first factory for making plows was established in England in 1783.

Though the problem of mechanizing harvest work was not solved until the 19th century, a few fairly simple threshing machines were designed and put in use by 1800. Similarly, a variety of machines for such tasks as preparing animal feed, chopping turnips, and cutting chaff were designed and used.

The English gentleman Jethro Tull made important contributions during this period, among them the horse-drawn hoe and the seed drill.

In order to save expensive hand labour, Tull designed and fabricated horse-drawn hoes that destroyed the weeds and kept the soil between the rows in good crumbly (friable) condition. Tull believed that friable soil would supply all the nutrients a plant required, and he was so convinced of his theory that he planted wheat in rows spaced widely enough to allow for horse hoeing while the plant was growing. Tull also believed that placing the seed into the ground in small holes (drilling) would permit a much lighter rate of seeding than hand-broadcast methods and still get high yields. In order to place the seed, he designed and used a seed drill.

Tull was not the first man to invent a seed drill, however. A seed dropper, consisting of a tube attached to a plow through which seed could be dropped by hand at regular intervals, had been attached to an ancient Babylonian plow. A similar system, but with two tubes behind two plowshares set in a rectangular frame, was employed in 17th-century China, and probably at a very much earlier date. An Italian described a modern form of seed drill in the late 16th century. An Austrian, Locatelli, described a *sembrador*, or seed drill attached to a plow, which was either used or proposed for use in Spain. After Tull, inventors were numerous, and, at the end of the 18th century, the drilling of seed, though not commonplace, was accepted. Throughout this period, however, the ordinary farmer continued to prepare his seedbed by plowing three times a year, occasionally more, and broadcasting seed by hand and harrowing it in. All these techniques were thousands of years old.

Irrigation and drainage. Walter Blith, a captain in Cromwell's army, was interested in land drainage and irrigation in the form of water meadows, pieces of low, flat land capable of being kept in fertility by being overflowed from some adjoining stream. During his lifetime considerable drainage work was carried out in eastern England, where it has continued to the present day. At the same time, water meadows were constructed in southwest England. This system was well known and of long standing in Germany. Similarly, there were sporadic attempts to drain wetlands in France and Italy; maintenance of the drainage system was essential to parts of the Netherlands. Colonization of reclaimed lands in eastern England by Dutch settlers also introduced some novel crops to that area, especially rape and cole.

Water meadows

Crops from the New World. The age of exploration that began in the mid-15th century led to discovery of edible plants previously unknown to Europeans. The Arabs had introduced sugarcane and rice to some parts of southern Europe. The voyagers to North America returned with maize, tobacco, and the turkey. South America supplied the potato, cocoa, quinine, and some vegetable drugs, while coffee, tea, and indigo came from the Orient.

Though the tropical plants could not be cultivated in the temperate European climate, maize was rapidly adopted in southern Europe. Imported first to Ireland, the potato came into general use elsewhere by the end of the 18th century.

These novel crops expanded the dietary range in some places, and importation of the new beverages may have helped reduce the very large consumption of alcoholic drinks. Tea drinking was regarded as harmful by some fervent propagandists at the end of the 17th century and later, and, in the 1820s, William Cobbett even condemned the potato. By that time, however, the crop had become the staple food of the Irish rural population and was raised on a moderate scale in Lancashire and Cheshire for consumption by their growing manufacturing population, as it was in some parts of Germany, France, the Low Countries, and Italy.

Agricultural textbooks and improvement societies. The small number of vernacular farming textbooks printed in Europe during the 16th century increased in the 17th and became much larger in the 18th, indicating a developing interest in commercial farming. Population growth, especially in the urban centres, and the rise of industry opened new opportunities for the enterprising landlord and farmer, even though transport was difficult.

English writers on farming were more numerous than those in other countries. The French paid more attention to gardening, and the Thirty Years' War (1618–48) brought a long interruption in agricultural development in Germany and adjacent countries.

The Flemish and Dutch produced no agricultural texts, declaring that they were more interested in farming than in writing about it. Indeed, the system of crop rotation in Brabant and Flanders became the pattern for the rest of Europe. In 1645 an Englishman, Sir Richard Weston, wrote the first description of this system, which featured crops such as clover or roots, rotating with industrial crops such as flax, with no fallow break. Adoption in England and other parts of western Europe increased agricultural production between 1600 and 1800.

Crop rotation in Brabant and Flanders

The problem of educating the ordinary farmer occupied the minds of improving landlords who were the governing class in every European country. Improvement societies formed in Great Britain, and a semi-official Board of Agriculture was established there in 1793, financed in part by a government grant. It lasted until 1822. One of the first tasks it undertook was to make and publish a survey of each county in the kingdom in order to make the best local methods known to other districts. A similar plan had been tried in France in 1759. Other countries in western Europe engaged in similar activities. A Royal Society was established in Denmark in 1769 and made practical researches. The Italian Societa di Georgofili, founded in 1753, tried to stimulate technical improvement and planned large drainage schemes. The first similar Swiss society was established at Zürich in 1747. Frederick William I of Germany founded an agricultural school at Halle in 1727; and the first agricultural high school was begun, with royal encouragement, at Möglin, near Berlin, in 1806.

PROGRESS IN DIFFERENT COUNTRIES

Italy. Although Italy was the first modern country where a four-course rotation was advocated, traditional farming persisted there into the 18th century and even later. Italian farmers, however, had accepted maize, potatoes, tomatoes, and tobacco and relied upon these and other subtropical products for subsistence and sale. The vine continued to be important, as did the olive. The north was more advanced than the less fertile south; a wide variety of crops was grown there, often in long rotations with a reduction of the proportion of fallow. An occasional fodder crop of turnips, clover, and ryegrass was cultivated. Shortages of fodder encouraged adoption of the Norfolk four-course system. Plows and other implements, however, remained primitive. The mulberry tree for the support of silkworms was important in the Milanese and Piedmont area. Rice was also grown on irrigated land in that area. High yielding milch cows were also kept. Transhumance was practiced, with sheep on the lower lands in winter and on the mountain pastures in summer.

Spain. Spain remained backward, partly because of its physical and climatic conditions and partly from social pressure. The Merino sheep, fine horses, and mules continued to deserve their high repute. Some subtropical crops were cultivated in the south and east—sugarcane, cotton, vines, olives, figs, raisins, and oranges—but the arid and stony land of the centre and elsewhere remained substantially unproductive and thinly populated by a poor class of métayers (sharecroppers).

Britain. The Industrial Revolution in Great Britain drew many people from food production; this had happened before in the mercantile and manufacturing Low Countries. Since these people still had to be fed, there existed a strong incentive toward increased farm productivity. These countries could not, however, grow enough food to satisfy local demands, and they sought to import food, especially grain. Thus industrialization in England stimulated cereal growing in Prussia, Poland, and parts of Russia.

Germany. Increased grain production in Prussia was also stimulated by the personal efforts of Frederick the Great, who set out to repair some of the damage caused by the Thirty Years' War and generally to improve conditions in his country. He ordered the rearrangement of the open-field farms into separate enclosed holdings and did away with common grazing, doubtless to encourage better animal selection and breeding. He bred horses for the army and gave bulls and rams to favoured landowners. Elsewhere in the German congeries of small states, the princes enforced improvements by royal edicts. Fodder crops were introduced. In addition to keeping the animals indoors in winter, summer stall feeding was introduced in some areas, resulting in preservation of the animal manure formerly dropped on the commons.

Rehabilitation by Frederick the Great

Sugar beet had been grown for feed in Germany, but it was not until 1747 that sugar was successfully extracted from this root and not until 1802 that the first factory for making beet sugar was built in Silesia. This industry, however, did not really begin to grow until the time of the Napoleonic Wars and later.

Scandinavia. The countries of Scandinavia were influenced by the spirit of improvement, but less spectacularly. Danish farmers were somewhat hindered in the introduction of novel methods by political restrictions, and it was not until the end of the 19th century that serious advances in technique took place. Sweden had some improvers who imported Merino sheep, English cattle, and Angora goats and who studied plow design. Nevertheless, Swedish farming in general remained almost static until after 1757, when some exchange of strips made enclosed farms possible. Potatoes became a staple.

North America. Changes in farming technique that increased productivity, introduced primarily in Great Britain and the Netherlands before 1800, made British farming the example to the world, even the distant lands of America. That continent, north and south, had indeed supplied Europe with very valuable plants. By the end of the 18th century, the production of North America was sufficient to supply some of the necessities of a warring Europe. The New England settlements, like those of the Southern states, were expanding toward the west, as land was claimed from the forest. Agriculture produced small surpluses for export. The Southern states had, of course, always exported staples like tobacco, cotton and (Louisiana) sugar; but the processes by which these plants were cultivated remained primitive.

VI. The 19th-century power revolution on the farm: c. 1792–1914

This period was characterized by the partial mechanization of agriculture in western Europe, especially in Great Britain, and in the new lands of Australia, New Zealand, and North America, where wild, uncultivated, and virtually unoccupied land was made to yield huge plant and animal crops.

MECHANIZATION

Though the first steps had been taken earlier, it was not until after 1850 that mechanization took hold in western Europe and the newly settled countries. A variety of machines were slowly coming into use when the French revolutionary wars broke out in the 1790s. An efficient seed drill had been devised but still required demonstration in the 1830s to convince farmers of its value. A few threshing machines were in use before 1800, and they became steadily more popular until, in the 1830s, farm labourers in England rebelled against them because the machine robbed them of their winter employment. The speed with which the thresher was adopted is rather surprising, as there was a surplus rather than a shortage of labour at the time.

The reaper. Yet even an ample supply of labour could not always cope with the harvest by hand methods. Local labour frequently had to be supplemented by travelling gangs or by small tradesmen from neighbouring towns. This unsatisfactory situation proved an incentive to invention. Though reaping machines had been proposed before, it was not until Patrick Bell in Scotland and Cyrus Hall McCormick (Figure 7) in the United States produced their designs that this machine became a practical reality. After the Great Exhibition of London in 1851,

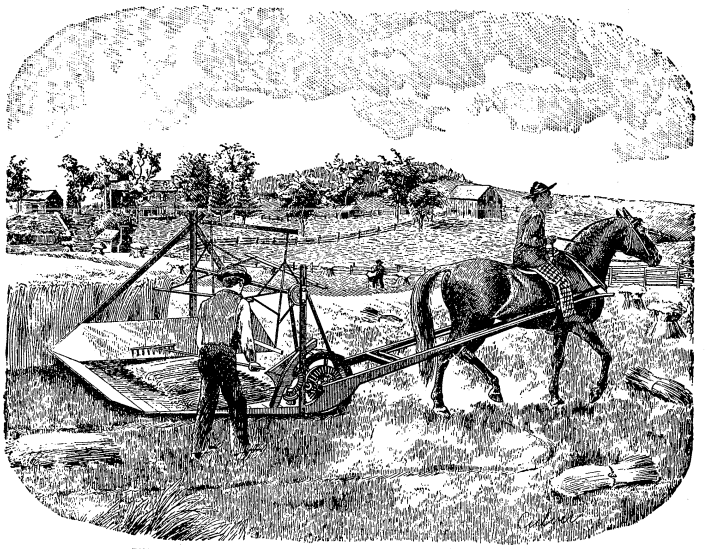


Figure 7: The original McCormick reaper, 1831; from R.L. Ardrey, "American Agricultural Implements," 1894.
By courtesy of the Library of Congress, Washington, D.C.

The combine harvester

the reaper slowly came into general use. About this time an animal-pushed combine harvester, which stripped the grain from the plant like the Gallo-Roman reaper of classical times, was devised. This machine was used successfully in South Australia, where more grain was grown than could be harvested manually by the few men available there.

Plows and plowing. Before harvesting machines could be used, however, the ground had to be prepared, and the fundamental instrument for this purpose was the plow. Though attempts had been made in the 17th and 18th centuries to develop a mathematical theory of plow design, in a less esoteric way practical men had made significant improvements. Prominent among these was the English inventor Robert Ransome, who patented a cast-iron share in 1785 and a self-sharpening share in 1803. Later he designed a plow with standard parts that could be removed and replaced in the field, a double plow (*i.e.*, with two shares), and other patterns.

Plowing the Great Plains

As settlers in the United States moved westward, plowing of the black prairie soils, high in organic matter, posed a special problem to the farmers who had cast-iron and iron-patched plows. The ingenuity of John Deere, an Illinois blacksmith and plow maker of the 1830s, resulted in a new kind of plow that was made entirely of steel except for the braces, beam, and handles. The one-piece share and moldboard of his first steel plow was cut from a mill-saw blade and shaped over a wooden form. This greatly improved implement not only made possible the effective plowing of the black prairie soils but also considerably lessened the animal power needed to turn the soil.

A mole plow, characterized by a cartridge-shaped "mole" attached to the bottom of its wide cutting blade, and intended for plowing a drain in heavy wetland, was successfully fabricated and used in the late 18th century. The plow, pulled by a capstan or team of animals, was a great help in draining wetland. Invention of a method for manufacturing tubular clay pipes also aided and simplified drainage. Use of these pipes spread rapidly beginning about 1850; large areas of land were thus drained all over Europe. Drainage by this method continues today.

Steam-powered equipment. The first attempts at steam-powered plowing took place in the 1830s. Though ingenious, the early apparatus was impractically heavy and cumbersome. A successful steam plow was made in the 1860s, after which steam power was widely adopted, especially on large farms (Figure 8). Supplementary implements were produced about this time, including hay rakes, hay-loaders, and potato spinners. Barn machinery, housed in especially designed buildings, was driven by a steam engine through a system of belts and pulleys or in more primitive style by horses. Threshing drums and supplementary machines were often hauled about the countryside by steam traction engines and worked on a contract basis by itinerant gangs of labourers.

Steam power spread rapidly but thinly over Europe. The first steam plow was worked in The Netherlands in 1862. Though milking machines were built, they were not really successful until the end of the century. The cream separator and milk cooler, both little changed since the time of their invention, were introduced in the 1880s.

Progress in mechanization in the second half of the 19th century was not, of course, confined to agriculture. Transport and communications improved enormously. Railroads were built all over the world, making agriculture possible in areas where farmers previously had no way to carry their goods to market. Steamships facilitated the movement of goods and the spread of ideas.

APPLYING SCIENCE TO FARMING

New fertilizers. Mechanical developments were paralleled by scientific discovery. Research in plant physiology and nutrition, begun in the 18th century, continued and grew in scope. The brilliant English chemist Sir Humphry Davy summarized the current state of agricultural knowledge in a series of lectures, published as *Elements of Agricultural Chemistry* in 1813. Though Davy's lectures probably had little impact on the ordinary farmer, a few

By courtesy of J I Case Company



Figure 8: Self-propelled steam engines were extensively used during the latter part of the 19th century for operating grain-threshing machines and other stationary farm machines such as corn husker–shredders and shellers.

enterprising pioneers were soon conducting field experiments with new fertilizers. The beneficial effects of saltpetre (nitrate of potash) had been known since the 17th century. When a firm began to import saltpetre from India in the 1820s, interested farmers bought and used it. Imported Chilean nitrate of soda was more plentiful, however, and more often used. Peruvian guano, an organic fertilizer, was imported and extensively used, even though sellers often diluted it with useless material to increase its bulk. Ground bones and other kinds of waste had been used in the late 18th century for restoring phosphate to pastures in dairy areas. Perhaps it was the use of these bones and their cracking and soaking in sulfuric acid that led to the production of superphosphate of lime in the 1840s. As the 19th century continued, more mineral fertilizers were discovered and marketed. With the development of the chemical industry, other fertilizers were manufactured. Basic slag, a waste product of the iron industry, was applied to grasslands with success. Gypsum was also tried but was later superseded by more effective products.

Peruvian guano

Agricultural research and education. Scientists of the 18th century had established the principles that governed plant life. About 1837 the English agriculturalist Sir John Bennet Lawes began to experiment with the effects of manures on plants and crops. In 1842 he patented a process for treating phosphate rock to produce superphosphate and thus initiated the synthetic fertilizer industry. In the following year, Lawes enlisted the services of the English scientist J.H. Gilbert, with whom he worked for more than half a century, performing experiments on crop and animal nutrition. The work of Lawes' agricultural experiment station became world renowned. At the same time similar work was carried out in France, Germany, and the United States.

Though a considerable number of books on agricultural subjects appeared during the 19th century, their effect was perhaps less great than the measures taken to promote agricultural education in most European countries. Though schools for farmers had been established in some German states in the 18th century, the first professorships of rural agriculture and economy were established at Oxford (1790) and Edinburgh (1797). Though similar events took place in France and Germany in the 19th century, perhaps the key date in the history of agricultural research and education is 1862, when the United States Congress set up the Department of Agriculture and provided for colleges of agricultural and mechanical arts in each state.

Disease. An important spur to research was the great prevalence of plant disease, whioh at intervals became epidemic. Two crops—the potato and the grapevine—were virulently attacked in the 1840s, and irreparable damage resulted. When the potato blight struck Ireland, where the potato had long been a staple food crop, it caused widespread famine. The vine growers of southern Europe suffered several disastrous epidemics of the vine fungus.

FARMING OUTSIDE EUROPE

European colonies. The 19th century brought European colonial expansion and plantation farming to many areas of the world. Coffee was grown widely in Ceylon until disease destroyed the plants and caused their replacement by tea bushes. Sugarcane was raised on a large scale in the East and West Indies, in the United States, and Hawaii. Cocoa was carried from South America to colonies in West Africa, where it prospered. Natural rubber was tapped in the Congo and Brazil and was later cultivated in well-designed plantations in the Malay Peninsula. Palm oil was collected from trees in West Africa and the South Seas. Yet these spectacular developments had little impact on the native agriculture of the underdeveloped regions in general; ancient patterns of cultivation remained largely unchanged.

America and Australia. By 1900 much of the western United States had been settled; great livestock ranches and wheat farms had been established there and in Argentina. A large dairy and sheep industry had grown up in Australia and New Zealand. Imported Merinos were the foundation stock of the vast flocks of the Australian hinterland. Devon cattle were imported from England to South Africa, where sheep were also important.

The Great Plains of the United States, lying between the Mississippi River and the Rocky Mountains, were first used by the open-range cattle industry, whose heyday lasted from about 1866 to 1886. To improve their lean, lanky longhorn, breeders imported Herefords, Durham Shorthorns, and other fine European cattle. As the range became more and more heavily occupied, and as wheat farmers began to till the soil of the Great Plains, it became necessary to put up fences. This was accomplished with a remarkable invention, barbed wire, a new, cheap, and rapid kind of fencing, used in preference to post and rail or sod and stone walls, which had been seen in Europe for centuries.

Barbed wire

Grain farmers rapidly followed the ranchers, and farming spread from Oklahoma north into Canada. Because hand labour was scarce, cultivation and harvesting were rapidly mechanized. A combine harvester hauled by a huge team of mules was used in California in the late 1880s. Different from the combine harvester used in South Australia, this was a prototype of today's machine. Iron and steel plows were produced in large numbers, and steam traction engines were used as power units for many mechanized operations such as threshing. Though seed was usually drilled, yields were often low because of a shortage of manure. Furthermore, because manure was scarce, a three-year rotation was practiced, with two crops followed by a fallow year. In the Southern states, export staples were produced in addition to subsistence crops.

America's prodigious new supplies of meat and grain were generally exported from the Western to the Eastern states, as well as to Europe, where they lowered prices paid to European farmers but played an important role in feeding industrial populations. Fresh meat and dairy produce from Argentina and Australia were carried thousands of miles in newly developed refrigerator ships.

The huge quantities of inexpensive food produced by the newly settled countries created a difficult problem for the European farmer. The arable and mixed farming common in most European countries simply could not compete. In Great Britain, much cropland was planted with grass, and farmers began to raise pedigree animals and dairy produce. On the Continent, where many farmers occupied peasant holdings, a kind of subsistence farming continued. Denmark and The Netherlands turned to production of dairy produce and high-quality bacon, feeding pigs on waste dairy materials. (G.E.F.)

VII. Scientific agriculture: the 20th century

Agricultural technology has developed more rapidly in the 20th century than in all previous history. Though the most important developments during the first half of this century took place in the industrial countries, especially the United States, the picture has changed somewhat since the 1950s. With the coming of independence, former colonies in Africa and Asia have initiated large-scale efforts to improve their agriculture. In many cases they have used considerable ingenuity in adapting Western methods to their own climates, soils, and crops (see also AGRICULTURE, TECHNOLOGY OF).

DEVELOPMENTS IN POWER: THE INTERNAL-COMBUSTION ENGINE

The internal-combustion engine brought major changes to agriculture in most of the world. In advanced regions it soon became the chief power source for the farm.

The tractor. The first applications to agriculture of the four-stroke-cycle gasoline engine, invented in the 1860s, were as stationary engines, at first in Germany, later elsewhere. By the 1890s, stationary engines were mounted on wheels to make them portable, and soon a drive was added to make them self-propelled. The first successful gasoline tractor was built in the United States in 1892. Within a few years several companies were

manufacturing tractors in Germany, the United Kingdom, and the United States.

By 1907 there were about 600 gasoline-powered tractors in use in the United States. By 1920, stimulated by the demand for food during World War I, the number had reached 246,000. In 1940 it was 1,567,000; in 1950, 3,394,000. By the mid-1950s the number of tractors on American farms exceeded the number of horses and mules. Excluding garden types, there were 4,810,000 tractors on United States farms in 1968, compared with 1,821,000 in the Soviet Union; 353,700 in the United Kingdom; and 1,293,000 in West Germany. There were 14,816,000 tractors in the entire world.

Major changes in tractor design, greatly increasing efficiency, include the power takeoff in 1918, the all-purpose, or tricycle-type, tractor in 1924, rubber tires in 1932, and more versatile transmissions since 1954. Power takeoff permits direct transmission of power from the tractor's own engine, through a special shaft and under control of the tractor operator.

Unit machinery. In recent years there has been an increase in the use of self-propelled machines in which the motive power and the equipment for performing a particular task form one unit. Though the grain combine is the most important of these machines, self-propelled units are also in use for spraying, picking cotton, bailing hay, picking corn (maize), and havesting tomatoes, lettuce, sugar beets, and many other crops. These machines are faster, easier to operate, and above all, have lower labour requirements than those powered by a separate tractor.

Grain combine. The first successful grain combine, a machine that cuts ripe grain and separates the kernels from the straw, was built in the United States in 1836. Lack of an adequate power unit and the tendency of combined grain to spoil because of excessive moisture limited its development, however. Large combines, powered by as many as 40 horses, were used in California in the latter part of the 19th century. Steam engines replaced horses on some units as a power source, but, about 1912, the gasoline engine began to replace both horses and steam for pulling the combine and operating its mechanism. A one-man combine, powered by a two-plow-sized tractor (*i.e.*, one large enough to pull two plows), was developed in 1935. This was followed by a self-propelled machine in 1938.

Numbers of combines in different countries

In 1970, the United States had 850,000 grain combines; in 1968, the Soviet Union had 581,000 combines; Canada, an estimated 160,000; West Germany, 152,000; France, 117,000; Australia, 68,000; the United Kingdom, 60,000; and Romania, 47,000.

Mechanized equipment for corn. Corn, or maize, the most important single crop in the United States and extremely important in many other countries, is grown commercially with the aid of equipment operated by tractors or by internal-combustion engines mounted on the machines. Using a tractor, three-bottom plow, ten-foot (three-metre) disk, four-row planter, a cultivator, a corn-head attachment for a combine (*i.e.*, a device especially adapted for removing ears of corn from the stalk), and trucks, four to six man-hours of labour are required to farm one acre (0.4 hectare) producing 75 bushels (about 2,800 litres) of corn. In 1900, 38 man-hours were required in the United States to produce 26 bushels (about 900 litres), using horsepower rather than tractor power.

Corn pickers came into use in the United States Corn Belt after World War I and have been even more widely adopted since World War II. Horse-drawn equipment, used for cutting corn in some areas from about 1890, has almost entirely disappeared. Corn pickers vary in complexity from the snapper-type harvester, which removes the ears from the stalks but does not husk them, to the picker-sheller, which not only removes the husk, but shells the grain from the ear. The latter are often used in conjunction with dryers.

Mechanized equipment for cotton. Mechanization has reduced the labour needed to grow one acre of cotton from 150 hours with one-row horse-drawn equipment to 25 hours using tractor-drawn or tractor-mounted equipment and machine picking, even though the yield per acre has more than doubled. Equipment includes tractor, two-row stalk-cutter, 14-foot (4.2-metre) disk (to shred the stalks), bedder (to shape the soil into ridges or seed beds), planter, cultivator, sprayer, and two-row harvester.

Cotton fibre is harvested by a stripper-type harvester, developed in the 1920s, or by a picker. The stripper strips the entire plant of both open and unopened bolls and collects many leaves and stems. Though a successful cotton picker that removed the seed cotton from the open bolls and left the burrs on the plant was invented in 1927, it did not come into use until after World War II. Strippers are used mostly in dry regions, while pickers are employed in humid, warm areas. The two-row tractor-mounted stripper harvests a bale of cotton (roughly 500 pounds, or 225 kilograms) in about 30 minutes when operating in cotton, yielding three-fourths of a bale per acre (about 422 kilograms per hectare), compared with one man snapping by hand in 30 hours. The pickers are either single-row machines mounted on tractors or two-row self-propelled machines. A one-row mechanical picker can replace 50 or more hand pickers.

Where strippers and pickers are used

Tomato harvesting equipment. The self-propelled mechanical tomato harvester, developed in the early 1960s by engineers working in cooperation with plant breeders, now handles virtually all packing tomatoes grown in California. In the 1960s the average crew using a machine harvested 0.35 ton (315 kilograms) per hour per worker. An efficient hand picker averaged 380 pounds (170 kilograms) per hour. A total of 61 man-hours of labour were required to harvest an acre of tomatoes yielding 21.5 tons (19,350 kilograms) using a machine, compared with 113 hours using handworkers. Harvesters using electronic sorters can further reduce labour requirements.

Automobiles, trucks, and airplanes. The automobile and truck have also had a profound effect upon agriculture and farm life. Motortrucks appeared on American farms between 1913 and 1920. By 1970 there were 3,185,000 trucks on United States farms—slightly more than one per farm. The 1960 World Census of Agriculture showed that nations with large farms and comparatively limited populations, such as Canada, New Zealand, and South Africa, made extensive use of trucks. On the other hand, such nations as Belgium and Japan, with small holdings and dense populations, reported very little such equipment.

Trucks have changed patterns of production and marketing of farm products. The farmer hauls seed, fertilizer, and other supplies from town, while such items as fertilizer, gasoline, and feed are delivered to the farm by trucks. Trucks go into the fields as part of the harvest equipment and haul the crops to markets, warehouses, or packing and processing plants. Most of the livestock is trucked to market.

Early experiments with airplanes in agriculture

The airplane may have been used agriculturally in the United States as early as 1918 to distribute poison dust over cotton fields afflicted with the pink bollworm. While records of this experiment are fragmentary, it is known that airplanes were used to locate and map cotton fields in Texas in 1919. In 1921, a widely publicized dusting experiment took place near Dayton, Ohio. Army pilots, working with Ohio entomologists, dusted a six-acre (2.4-hectare) grove of catalpa trees with arsenate of lead to control the sphinx caterpillar. The experiment was successful. It and others encouraged the development of dusting and spraying, mainly to control insects, disease, weeds, and brush. In more recent years, in recognition of the possible long-term harmful effects of some of the chemicals, dusting and spraying have been subject to various controls.

Airplanes are also used to distribute fertilizer, to reseed forest terrain, and to control forest fires. Many rice growers use planes to seed, fertilize, and spray pesticides, and even to hasten crop ripening by spraying hormones from the air.

During heavy storms, airplanes have dropped baled hay

to cattle stranded in snow. Airplanes have also been used to transport valuable breeding stock, particularly in Europe. Valuable and perishable farm products are frequently transported by air. Airplanes are especially valuable in such large agricultural regions as western Canada and Australia, where they provide almost every type of service to isolated farmers.

NEW CROPS AND TECHNIQUES

New crops and techniques are, in reality, modifications of the old. Soybeans, sugar beet, and grain sorghums, for example, all regarded as "new" crops, are new only in the sense that they are grown today in wider areas and have different uses from those of earlier times. Such techniques as terracing, dry farming, and irrigation are nearly as old as agriculture itself, but their widespread application is still increasing productivity in many parts of the world.

New crops. *The soybean.* This is an outstanding example of an ages-old crop that, because of the development of new processes to make its oil and meal more useful, is widely produced today. In the Orient, where the soybean originated long ago, over half the crop is used directly for food, and less than a third is pressed for oil. Its high protein and fat content makes it a staple in the diet, replacing or supplementing meat for millions of people.

Though first reported grown in America in 1804, the soybean remained a rare garden plant for nearly 100 years. Around the beginning of the 20th century, when three new varieties were introduced from Japan, farmers began growing it for hay, pasture, and green manure. In the early 1930s a method of processing soybean oil that eliminated a disagreeable odour from the finished product was developed. World War II brought an increased demand for edible oil. The food industry began using soybean oil for margarine, shortening, salad oil, mayonnaise, and other food products and is presently its chief user. Manufacturers of paints, varnishes, and other drying oil products are the most important nonfood users.

Solvent process for soybean oil

Development of the solvent process of extracting soybean oil has greatly increased the yield. A 60-pound (27-kilogram) bushel of soybeans processed by this method will yield 10.5 pounds (about 4.5 kilograms) of oil and 45 pounds (20 kilograms) of meal. Soybean meal and cake are used chiefly for livestock feed in the United States. The high protein content of the meal has made it an attractive source of industrial protein, and, with proper processing, it is an excellent source of protein for humans.

World production of soybeans in 1970 was 1,535,000,000 bushels. Of this total, 75 percent was produced in the United States, 17 percent in China, and between 1 and 2 percent each in Brazil, the Soviet Union, and Indonesia.

Development of new varieties suited for different parts of the world is possible by means of hybridization. This permits isolating types superior in yielding ability, resistance to lodging (breakage of the plant by wind and rain) and shattering (of the bean), adaptation to suit various requirements for maturity, and resistance to disease. Hybridization, however, has not yet led to spectacular gains in yields.

Sorghum. Just as the soybean was used for many centuries in Asia before its introduction into the Western world, so sorghum was a major crop in Africa. Sorghum is fifth in importance among the world's cereals, coming after wheat, rice, maize, and barley. It is called Guinea corn in West Africa, kafir corn in South Africa, durra in The Sudan, and mtama in East Africa. In India it is known as juar, jowar, cholam, and great millet and is called kaoliang in China. In America, it is often called milo or milo maize, while the sweet-stemmed varieties are referred to as sorgo.

Sorghum probably had been domesticated in Ethiopia by 3000 BC. From there, it spread to West and East Africa, and then southward. Traders from Africa to the East carried sorghum as provisions on their dhows. It is likely that sorghum thus reached India, where cultivation began between 1500 and 1000 BC. Other traders carried sorghum to China and the Far East. The amber sorghums or sorgos, useful for forage and syrup, may have moved by sea, while the grain sorghums probably moved overland. The movement to the Mediterranean and the Middle East also began through traders.

Sorghum reached the New World through the slave trade. Guinea corn and chicken corn came from West Africa to America as provisions for the slaves. Other types were introduced into the United States by seedsmen and scientists from about 1870 to 1910. Seed was sometimes sold to farmers as a new, highly productive variety of maize. It was not until the 1930s, after the value of the plant as grain, forage, and silage for livestock feeding had been recognized, that acreage began to increase. Yields rose markedly in the late 1950s, after successful hybridization of the crop. Better yields led in turn to increased acreage.

Sorghum sugar

Chinese ambercane was brought from France to the United States in 1854 and was distributed to farmers. While the cane provided good forage for livestock, promoters of the new crop were most interested in refining sugar from the sorghum molasses, a goal that persisted for many years. While refining technology has been perfected, the present cost of sorghum sugar does not permit it to compete with sugar from other sources.

The Far East devotes the largest areas to sorghum, followed by Africa and North America. In 1970 the United Nations Food and Agriculture Organization reported 37,800,000 hectares (93,430,000 acres) in sorghum, of which the Far East, excluding China, planted 19,400,000 hectares; Africa, 7,150,000; and North America, 5,600,000. India produced 10,000,000 metric tons; China, an estimated 19,300,000 (1968); the United States, 17,700,000; Nigeria, 3,500,000; The Sudan, 1,350,000; Argentina, 4,070,000; and Mexico, 2,350,000. Production of grain sorghum in the United States more than quadrupled between the early 1950s and the late 1960s, the dramatic increase due primarily to higher productivity resulting from successful hybridization.

Large amounts of sorghum grain are eaten every year by people of many countries. If the world population continues to grow as projected, food is likely to be its most important use. Most of the sorghum is ground into flour, often at home. Some is consumed as a whole-kernel food. Some of the grain is used for brewing beer, particularly in Africa.

Sugar beet. The sugar beet is much newer as a crop than either soybeans or sorghum. Although beets had been a source of sweets among ancient Egyptians, Indians, Chinese, Greeks, and Romans, it was not until 1747 that a German apothecary, Andreas Marggraf, obtained sugar crystals from the beet. Some 50 years later, Franz Karl Achard, son of a French refugee in Prussia and student of Marggraf, improved the Silesian stock beet—probably a mangel-wurzel—as a source of sugar. He erected the first pilot beet-sugar factory at Cunern, Silesia (now Poland), in 1802. Thus began the new use for sugar of a crop traditionally used as animal feed.

Napoleonic experimentation with beet sugar

When during the Napoleonic Wars continental Europe was cut off from West Indies cane sugar, further experimentation with beet sugar was stimulated. In 1808 a French scientist, Benjamin Delessert, used charcoal in clarification, which insured the technical success of beet sugar. On March 25, 1811, Napoleon issued a decree that set aside 80,000 acres (32,000 hectares) of land for the production of beets, established six special sugar-beet schools to which 100 select students were given scholarships, directed construction of ten new factories, and appropriated substantial bounties to encourage the peasants to grow beets. By 1814, 40 small factories were in operation in France, Belgium, Germany, and Austria. Although the industry declined sharply after Napoleon's defeat, it was soon revived. For the last third of the 19th century, beets replaced cane as the leading source of sugar. By 1968, however, 42,700,000 tons of cane sugar entered the market, compared with 33,700,000 tons of beet. Beet-sugar production is usually subsidized in western Europe and the United States.

Since World War II, major changes have taken place in sugar-beet production in the United States and, to a lesser extent, in the Soviet Union, Germany, and other nations with a substantial production. These changes may be illustrated by developments in the United States.

In 1931 the California Agricultural Experiment Station and the United States Department of Agriculture undertook a cooperative study of the mechanization of sugar-beet growing and harvesting. The goal in harvesting was a combine; that is, a machine that would perform all the harvesting operations—lifting from the soil, cutting the tops, and loading—in one trip down the row. By the end of World War II, four different types of harvesters were being manufactured.

The spring and summer operations—planting, blocking (cutting out all plants except for clumps standing ten or 12 inches apart), thinning, and weeding—did not yield so easily to mechanization, largely because the beet seed, a multigerm seedball, produced several seedlings, resulting in dense, clumpy, and somewhat irregular stands. In 1941 a machine for segmenting the seedball was developed. The problem was solved in 1948, when a plant with a true single-germ seed was discovered in Oregon. Now precision seed drills could be used, and plants could be first blocked and then cultivated mechanically using a cross-cultivating technique; that is, cultivating the rows up and down and then across the field. During World War I, 11.2 hours of labour were required to produce a ton (900 kilograms) of sugar beets; in 1964, 2.7 hours were needed.

The world's major producer of beet sugar is the Soviet Union. In 1970–71 that nation produced 9,300,000 tons, compared with 3,500,000 tons in 1953. The United States produced 3,200,000 tons in 1970–71; France, 2,600,000 tons; West Germany, 2,100,000; Poland, 1,600,000; Italy, 1,200,000; and the United Kingdom, 1,000,000 tons.

New techniques. As the development of the sugar beet shows, new techniques may bring particular crops into new prominence. This discussion, however, is confined to three that, in some forms, are old yet today are transforming agriculture in many parts of the world.

Terracing. Terracing, which is basically grading steep land, such as hillsides, into a series of level benches, was known in antiquity and was practiced 500 years ago in such divergent areas as Peru and central Africa. Today, terracing is of major importance in Japan, Mexico, and parts of the United States, while many other nations, including Israel, Australia, South Africa, Colombia, and Brazil are increasing productivity through the inauguration of this and other soil-conserving practices.

Colombia provides an example of the modern need for terracing. For many years, the steep slopes used for producing the world-renowned Colombian coffee have been slowly eroding. During the 1960s experimental work showed that contour planting and terracing would help preserve the land. Further south, the Brazilian state of São Paulo created a terracing service in 1938. Since then, the program has become a full conservation service.

Irrigation. The usefulness of a full-scale conservation project is seen in the Snowy Mountains Project of Australia, the world's driest continent, where three river systems are being diverted to convert hundreds of miles of arid but fertile plains to productive land. Intensive soil-conservation methods have been undertaken wherever the natural vegetation and soil surface have been disturbed. Drainage is controlled by stone and steel drains, grassed waterways, absorption and contour terraces, and settling ponds. Steep slopes are stabilized by woven wickerwork fences, brush matting, and bitumen sprays, followed by revegetation with white clover and willow and poplar trees. Grazing is strictly controlled to prevent silting of the reservoirs and damage to slopes. The two main products of the plan are power for new industries and irrigation water for agriculture, with recreation and a tourist industry as important by-products.

Snowy Mountains Project

The Snowy Mountains Project of Australia is a modern successor, so far as irrigation is concerned, to practices that have provided water for crops almost from the beginnings of agriculture. The simplest irrigation device was to dip water from a well or spring and pour it on the land. Many types of buckets, ropes, and, later, pulleys were employed. The ancient swipe, or shadoof, a long pole pivoted from a beam, with a weight at one end to lift a full bucket of water at the other, is still in use. Conduction of water through ditches from streams was practiced widely in the Middle East, in Africa, and in America, where ancient canal systems can be seen. A conduit the Romans built 2,000 years ago to provide a water supply to Tunis is still in use.

Expanding irrigation is an absolute necessity to extend crop acreage in significant amounts; it may be the most productive of possible improvements on present cropland. Sufficient water at the proper time makes possible the full use of technology in farming—including the proper application of fertilizers, suitable crop rotations, and the use of more productive varieties of crops.

About 160,000,000 hectares (400,000,000 acres) were being irrigated throughout the world in the mid-1960s, amounting to 14 percent of the arable land. This acreage might be increased. First, there is the possibility of making wider use of irrigation in districts that already have a high rate of output. Second, there is the possibility of irrigating nonproductive land, especially in arid zones. The greatest immediate economic returns might well come from irrigating productive districts, but irrigation of arid zones has a larger long-range appeal. Most of the arid zones, occupying more than one-third of the landmass of the globe, are in the tropics. Generally, they are rich in solar energy, and their soils are rich in nutrients, but water is lacking.

Supplemental irrigation in the United States, used primarily to make up for poor distribution of rainfall during the growing season, has increased very substantially since 1939. This irrigation is carried on in the humid areas of the United States almost exclusively with sprinkler systems. The water is conveyed in pipes, usually laid on the surface of the field, and the soil acts as a storage reservoir. The water itself is pumped from a stream, lake, well, or reservoir. American farmers first used sprinkler irrigation about 1900, but the development of lightweight aluminum pipe with quick couplers meant that the pipe could be moved easily and quickly from one location to another, resulting in a notable increase in the use of sprinklers after 1945.

Leading nations by irrigated acreage are China, India, the United States, and the Soviet Union. India, where irrigation has been practiced since ancient times, illustrates some of the problems. In 1967 India had over 27,500,000 hectares (65,000,000 acres) under irrigation, about 20 percent of its cultivated area, and the area is increasing. Both large dams, with canals to distribute the water, and small tube, or driven, wells, made by driving a pipe into water or water-bearing sand, controlled by individual farmers, are in use. Some are affected by salinity, however. Tube wells help in these instances by lowering the water table and by providing sufficient water to flush away the salts. The other major problem is persuading Indian farmers to level their lands and build the small canals needed to carry the water over the farms.

Irrigation in India

Dry farming. The problem of educating farmers to make effective use of irrigation water is found in many areas. An even greater educational effort is required for dry farming; that is, crop production without irrigation where annual precipitation is less than 20 inches.

Dry farming as a system of agriculture was developed in the Great Plains of the United States early in the 20th century. The system included deep fall plowing, subsurface packing of the soil, thorough cultivation both before and after seeding, light seeding, and alternating-summer fallow, with the land tilled during the season of fallow as well as in crop years. Though none of the steps were novel, their systematic combination was new. Systematic dry farming has continued, with substantial modifications, in the Great Plains of Canada and the United States, in Brazil, in South Africa, in Australia, and elsewhere. It is under continuing research by the Food and Agriculture Organization of the United Nations.

The direction of change. While no truly new crop has been developed by modern man, new uses and new methods of cultivation of known plants may be regarded as new crops. There is, for example, a potential for the development of subsistence and special-use plants into new crops. New techniques, too, are the elaboration and systematization of practices from the past. There is undoubted future promise in crop farming in the "package" of modern agricultural technology, which usually includes machines, improved (often hybrid) seeds and breeding stock, careful tillage, adequate fertilizer, productive use of water, and the natural or chemical control of weeds, fungi, and insects.

NEW STRAINS: GENETICS

The application of genetics to developing new strains of plants and animals has brought major changes in agriculture since the 1920s. Genetics as the science dealing with the principles of heredity and variation in plants and animals was established only at the beginning of the 20th century. Its application to practical problems came later.

Early work in genetics. The modern science of genetics and its application to agriculture has a complicated background, built up from the work of many men. Nevertheless, Gregor Johann Mendel (1822–84) is generally credited with its founding. Mendel, a monk in Brunn, Austria (now Brno, Czechoslovakia), purposefully crossed garden peas in his monastery garden. He carefully sorted the progeny of his parent plants according to their characteristics and counted the number that had inherited each quality. He discovered that when the qualities he was studying, including flower colour and shape of seeds, were handed on by the parent plants, they were distributed among the offspring in definite mathematical ratios, from which there was never a significant variation. Definite laws of inheritance were thus established for the first time. Though Mendel reported his discoveries in an obscure Austrian journal in 1866, his work was not followed up for a third of a century. Then in 1900, investigators in The Netherlands, Germany, and Austria, all working on inheritance, independently rediscovered Mendel's paper.

Mendel's experiments

By the time Mendel's work was again brought to light, the science of genetics was in its first stages of development. The word genetics comes from "genes," the name given to the minute quantities of living matter that transmit characteristics from parent to offspring. By 1903, scientists in the United States and Germany had concluded that genes are carried in the chromosomes, nuclear structures visible under the microscope. In 1911 a theory that the genes are arranged in a linear file on the chromosomes and that changes in this conformation are reflected in changes in heredity was announced.

Genes are highly stable. During the processes of sexual reproduction, however, means are present for assortment, segregation, and recombination of genetic factors. Thus, tremendous genetic variability is provided within a species. This variability makes possible the changes that man can bring about within a species to adapt it to his specific uses. Occasional mutations (spontaneous changes) of genes also contribute to variability. By the 1960s, evidence indicated that nucleic acid is the fundamental hereditary substance. Experiments with deoxyribonucleic acid (DNA) may be reaching the bases of the continuity of life.

Development of new strains of plants and animals did not, of course, await the science of genetics, and some advances were made by empirical methods even after the application of genetic science to agriculture. The U.S. experimenter Luther Burbank (1849–1926) developed the Burbank potato as early as 1873 and continued his plant-breeding research, which produced numerous new varieties of fruits and vegetables, without any formal knowledge of genetic principles. In some instances, both practical experience and scientific knowledge contributed to major technological achievements. An example is the development of hybrid maize or corn.

Maize. Maize originated in America, probably having been first developed by Indians in the highlands of Mexico. It was quickly adopted by the European settlers, Spanish, English, and French. The first English settlers found the northern Indians growing a hard-kerneled, early-maturing flint variety that kept well, though its yield was low. The Indians in the south central area of English settlement grew a soft-kerneled, high-yielding, late-maturing dent corn. There were doubtless many haphazard crosses of the two varieties. In 1812, however, John Lorain, a farmer living near Philipsburg, Pennsylvania, consciously mixed the two and demonstrated that certain mixtures would result in a yield much greater than that of the flint, yet with many of the flint's desirable qualities. Other farmers and breeders followed Lorain's example, some aware of his pioneer work, some not. The most widely grown variety of the Corn Belt until recent times, Reid's Yellow Dent, originated from a fortuitous mixture of a dent and a flint variety.

Origins of maize

At the same time, other scientists besides Mendel were conducting experiments and developing theories that were to lead directly to hybrid corn. In 1876 Charles Darwin published the results of experiments on cross- and self-fertilization in plants. Carrying out his work in a small greenhouse in his native England, the man who is best known for his theory of evolution found that inbreeding usually reduced plant vigour and that crossbreeding restored it.

Darwin's work was studied by a young United States botanist, William James Beal, who probably made the first controlled crosses between varieties of maize for the sole purpose of increasing yields through hybrid vigour. Beal's achievement was without knowledge of the genetic principle involved. In 1908 George Harrison Shull concluded that self-fertilization tended to separate and purify strains while weakening the plants, but that vigour could be restored by crossbreeding the inbred strains. Another scientist found that inbreeding could increase the protein content of maize, but with a marked decline in yield. With knowledge of inbreeding and hybridization at hand, there was yet to be developed a technique whereby hybrid maize with the desired characteristics of the inbred lines and hybrid vigour could be combined in a practical manner. In 1918 Donald F. Jones of the Connecticut Agricultural Experiment Station discovered the answer, the "double cross."

The double cross

The double cross was the basic technique used in developing modern hybrid corn and has been used by commercial firms since. Jones's invention was to use four inbred lines instead of two in crossing. Simply, inbred lines A and B made one cross, lines C and D another. Then AB and CD were crossed, and a double-cross hybrid, ABCD, was the result. This hybrid became the seed that changed much of American agriculture. Each inbred line was constant both for certain desirable and for certain undesirable traits, but the practical breeder could balance his four or more inbred lines in such a way that the desirable traits outweighed the undesirable. Foundation inbred lines were developed to meet the needs of varying climates, growing seasons, soils, and other factors. The large hybrid seed-corn companies undertook complex applied-research programs, while state experiment stations and the United States Department of Agriculture tended to concentrate on basic research.

The first hybrid corn involving inbred lines to be produced commercially was sold by the Connecticut Agricultural Experiment Station in 1921. The second was developed by Henry A. Wallace, a future secretary of agriculture and vice president of the United States. He sold a small quantity in 1924 and, in 1926, organized the first seed company devoted to the commercial production of hybrid corn.

Many Midwestern farmers began growing hybrid corn in the late 1920s and 1930s, but it did not dominate maize production until World War II. In 1933, 1 percent of the total maize acreage was planted with hybrid seed. By 1939 the figure was 15 percent, and in 1946 it rose to 69. The percentage was 96 in 1960. The average per-acre yield of corn rose from 23 bushels (2,000 litres per hectare) in 1933, to 72 bushels (6,260 litres per hectare) in 1970. The United States in 1970 devoted 57,357,000 acres

(23,212,000 hectares) to maize, followed by Brazil with 26,934,000 acres (10,900,000 hectares), Mexico with 20,633,000 acres (8,350,000 hectares), India with 14,826,000 acres (6,000,000 hectares), South Africa with 12,849,000 acres (5,200,000 hectares), and the Soviet Union with 11,367,000 acres (4,600,000 hectares). As hybrid strains suitable for more areas are developed, world production seems certain to increase.

The techniques used in breeding hybrid corn have been successfully applied to grain sorghum and several other crops. New strains of most major crops are developed through plant introductions, crossbreeding, and selection, however, because hybridization in the sense used with corn and grain sorghums has not been successful. By 1968 attempts to apply the technique to wheat had resulted in making the genetic mechanisms available, but an economic means for obtaining the hybrid seed was still lacking.

Wheat. Advances in wheat production during the 20th century include improvements through the introduction of new varieties and strains; careful selection by farmers and seedsmen, as well as by scientists; and crossbreeding to combine desirable characteristics (see Figure 1 for a comparison between wild and cultivated wheat). The adaptability of wheat enables it to be grown in almost every country of the world. The ten countries with the highest average production are the Soviet Union, the United States, China, Canada, France, India, Italy, Turkey, Australia, and Argentina. In most of the developed nations producing wheat, endeavours of both government and wheat growers have been directed at bringing science to wheat breeding.

Canadian Marquis wheat

The development of the world-famous Marquis wheat in Canada, released to farmers in 1900, came about through sustained scientific effort. Sir Charles Saunders, its discoverer, followed five principles of plant breeding: (1) the use of plant introductions; (2) a planned crossbreeding program; (3) the rigid selection of material; (4) evaluation of all characteristics in replicated trials; and (5) testing varieties for local use. Marquis was the result of crossing a wheat long grown in Canada with a variety introduced from India. For 50 years, Marquis and varieties crossbred from Marquis dominated hard red spring wheat growing in the high plains of Canada and the United States and were used in other parts of the world.

In the late 1940s, a short-stemmed wheat was introduced from Japan into a more favourable wheat-growing region of the United States Pacific Northwest. The potential advantage of the short, heavy-stemmed plant was that it could carry a heavy head of grain, generated by the use of fertilizer, without falling over or "lodging." Early work with the variety was unsuccessful; it was not adaptable directly into American fields. Finally, by crossing the Japanese wheat with acceptable varieties in the Palouse Valley in Washington, the first true semidwarf wheat in the United States to be commercially grown under irrigation and heavy applications of fertilizer resulted. This first variety, Gaines, was introduced in 1962, followed by Nugaines in 1966. The varieties now grown in the United States commonly produce 100 bushels per acre (8,800 litres per hectare), and world records of more than 200 bushels per acre have been established.

The Rockefeller Foundation in 1943 entered into a cooperative agricultural research program with the government of Mexico, where wheat yields were well below the world average. By 1956, per-acre yield had doubled, mainly because of newly developed varieties sown in the fall instead of spring and the use of fertilizers and irrigation. The short-stemmed varieties developed in the Pacific Northwest from the Japanese strains were then crossed with various Mexican and Colombian wheats. The new dwarf lines were released in 1961. By 1965 the new Mexican wheats were established and were gaining an international reputation.

Improving Mexican wheat

Experimentation with Mexican wheats was closely watched. During the crop year 1967–68, there were at least 18 other countries growing Mexican wheats on large acreages. Once the seeds' adaptability to local conditions was established, several nations, particularly India, Pakistan, and Turkey, imported them from Mexico in thousands of tons, and at prices only marginally higher than world market prices for wheat. These seeds are transforming wheat production in the areas of those nations where irrigation or high rainfall permits the use of fertilizer.

Rice. The success of the wheat program led the Rockefeller and Ford foundations in 1962 to establish the International Rice Research Institute at Los Baños in the Philippines. A research team assembled some ten thousand strains of rice from all parts of the world and began outbreeding. Success came early with the combination of a tall, vigorous variety from Indonesia and a dwarf rice from Taiwan. The strain IR-8 has proved capable of doubling the yield obtained from most local rices in Asia. The new strain is responsive to fertilizer and is early maturing. In some areas, the early maturity means that an extra crop of rice can be grown each year.

Future prospects. The application of genetics to agriculture since World War II has resulted in substantial increases in the production of many crops. This has been most notable in hybrid strains of maize and grain sorghum. At the same time, crossbreeding has resulted in much more productive strains of wheat and rice, which may help end the threat of starvation in Asia. It may be that genetics will influence agricultural production in the future to an even greater extent than it has in the past.

ANIMAL BREEDING

The goal of animal breeders in the 20th century has been to develop types of animals that will meet market demands, be productive under adverse climatic conditions, and be efficient in converting feed to animal products. At the same time, producers have increased meat production by improved range management, better feeding practices, and the eradication of diseases and harmful insects. The world production of meat has been increasing steadily since World War II. In 1963, the world total reached a record of 151,000,000,000 pounds (50,400,000,000 kilograms), 38 percent above the 1952–56 average, and by 1970 the total was 177,000,000,000 pounds (80,300,000,000 kilograms).

World meat production

While the number of livestock in relation to the human population is not significantly lower in less developed than in more developed regions, there is much lower productivity per animal and thus a much lower percentage of livestock products in diets. Primitive breeding practices usually prevail in the less developed regions, while great care is given to animal breeding in the more developed regions of North America, Europe, Australia, and New Zealand.

The advances made in developing new, highly productive strains of crops through the application of genetics have not been matched by similar advances in livestock. Except for broiler chickens in the United States, little progress has been made in improving the efficiency with which animals convert feed to animal products. The potential exists. Research on the breeding and nutrition of poultry, for example, makes it possible to produce chickens of 3.5 pounds (1,575 grams) for market in 65 days, compared with the 91 days needed before the research findings were applied.

Hogs. Advances in animal breeding have been made by careful selection and crossbreeding. These techniques are not new. The major breeds of English cattle, for example, were developed in the 18th and early 19th centuries by selection and crossbreeding. The Poland China and Duroc Jersey breeds of swine were developed in the United States in the latter part of the 19th century by the same means.

The hogs developed in the United States in the latter part of the 19th and first part of the 20th century were heavy, fat-producing animals that met the demands for lard. During the 1920s, however, lard became less important as a source of fat because of increasing use of cheaper vegetable oils. Meat-packers then sought hogs yielding more lean meat and less fat, even though market prices moved rather slowly toward making their production profitable.

Breeding leaner hogs

At the same time, Danish, Polish, and other European breeders were crossbreeding swine to obtain lean meat and vigorous animals. An outstanding new breed was the Danish Landrace, which in the 1930s was crossed with several older United States breeds, eventually giving rise to several new, mildly inbred lines. These lines produced more lean meat and less fat, as well as larger litters and bigger pigs.

Sheep. Similar crossbreeding, followed by intermating and selection with the crossbreeds, brought major changes in the sheep industries of New Zealand and the United States. The goal in New Zealand was to produce more acceptable meat animals, while that in the United States was to produce animals suited to Western range conditions and acceptable both for wool and mutton.

During the late 19th century, several New Zealand sheep breeders began crossing Lincoln and Leicester rams with Merino ewes. Early in the 20th century, the Corriedale had become established as a breed, carrying approximately 50 percent Australian Merino, with Leicester and Lincoln blood making up the remainder. The Corriedale was successfully introduced into the United States in 1914. Since World War II, a more uniform lamb carcass has been developed in New Zealand by crossing Southdown rams with Romney ewes.

With different objectives in view, breeders in the United States in 1912 made initial crosses between the long-wool mutton breed, the Lincoln, and fine-wool Rambouillets. Subsequent intermating and selection within the crossbreds led to a new breed, the Columbia. Both the Columbia and the Targhee, another breed developed in the same way as the Columbia, have been widely used. They are suited to the Western ranges, and they serve reasonably well both as wool and meat animals.

Beef cattle. Changes in beef cattle, particularly the establishment of new breeds, have resulted from selective linebreeding and from crossbreeding. The Polled Shorthorn and the Polled Hereford breeds were established by locating and breeding the few naturally hornless animals to be found among the horned herds of Shorthorns and Herefords, first established as distinctive breeds in England. It is of particular note that the originator of the Polled Herefords made an effort to locate naturally hornless Herefords and begin linebreeding with them after he had studied Charles Darwin on mutations and variations and how they could be made permanent by systematic mating.

Three new breeds of beef cattle

Three new breeds originating in the United States were developed for parts of the South where the standard breeds lacked resistance to heat and insects and did not thrive on the native grasses. The first of these breeds, the Santa Gertrudis, was developed on the King Ranch in Texas by crossbreeding Shorthorns and Brahmans, a heat- and insect-resistant breed from India. The Santa Gertrudis cattle carry approximately five-eighths Shorthorn blood and three-eighths Brahman. They are heavy beef cattle and thrive in hot climates. Increasing numbers are exported to South and Central America each year, where they are used to upgrade the native cattle.

The Brangus breed was developed in the 1930s and 1940s by crossing Brahman and Angus cattle. The breed has been standardized with three-eighths Brahman and five-eighths Angus breeding. The Brangus generally have the hardiness of the Brahman for Southern conditions but the improved carcass qualities of the Angus.

The Beefmaster was developed in Texas and Colorado by crossbreeding and careful selection, with the cattle carrying about one-half Brahman blood and about one-fourth each of Hereford and Shorthorn breeding. Emphasis was given to careful selection, emphasizing disposition, fertility, weight, conformation, hardiness, and milk production.

Artificial breeding. An increase in milk production per cow in the 20th century has come through better nutrition and artificial breeding. Artificial breeding permits the use of proven sires, developed through successive crosses of animals of proven merit. An Italian scientist experimented successfully with artificial insemination in 1780, but its practical usefulness was not demonstrated until the 20th century. A Soviet scientist, Ilya Ivanov, established the Central Experimental Breeding Station in Moscow in 1919, to continue work that had begun some 20 years earlier. As early as 1936, more than 6,000,000 cattle and sheep were artificially inseminated in the Soviet Union.

After the Soviets reported their successes, scientists in many nations experimented with artificial breeding. Denmark began with dairy cattle in the 1930s. The first group in the United States began work in 1938. Statistics show that the milk and butterfat production of proved sires' daughters, resulting from artificial breeding, is higher than that of other improved dairy cattle. Furthermore, a single sire can serve 2,000 cows a year, as compared with 30 to 50 in natural breeding.

In summary, crossbreeding and careful selection, combined with such techniques as artificial insemination, better feeding, and control of diseases and pests, have made substantial contributions to livestock production in the 20th century. More intensified research, particularly in hybridization, offers promise of help to livestock producers in meeting increasing world demands for their products. (W.D.R.)

ELECTRICITY IN AGRICULTURE

Electric power has had an impact on modern agriculture that promises to be more significant than that of either steam or gasoline, because electricity in its nature is far more versatile than the earlier power sources.

Electricity and plants. Experiments in electrical stimulation of plant growth and development were being carried out as early as the 18th century. In 1746 an Edinburgh physician published an account of his experiments claiming that the "electrification of myrtle trees" resulted in accelerated shoot and leaf production and earlier flowering. An early French experimenter found in 1873 that "by conveying atmospheric electricity to plants their fertility was increased and their general appearance much enhanced." Many other investigations followed, and it was demonstrated beyond question that the growth of plants was actually stimulated by electricity. Efforts have continued along these lines in many countries, particularly the United States, Great Britain, and Italy, on a range of cereal and vegetable crops, but the results have not yet been translated into commercial practice.

Light. The effect of light on plant growth and development had interested botanists for many years, and development of a range of electric lamps gave added impetus to their investigations. Work in England, the United States, and France in the late 19th century ultimately resulted in the development of plant lighting techniques of immense value to horticulture, both in scientific study and commercial practice.

Electrical cooperatives. It was, however, the development of the electric motor that really gained the interest of the farming community. Some authorities saw its value to farmers as early as 1870. Despite the obvious advantages of the other, more available power sources, progressive farmers in a number of countries were determined to exploit the possibilities of electricity on their farms. To get electricity, farmers formed cooperatives, which either bought bulk power from existing facilities or built their own generating stations.

It is believed that the first such cooperatives were formed in Japan in 1900, followed by similar organizations in Germany in 1901. Multiplying at a considerable rate, these farmer cooperatives not only initiated rural electrification as such but provided the basis for its future development.

Rural electrification

From these small beginnings the progress of rural electrification, though necessarily slow, steadily gained impetus until, in the 1920s, public opinion compelled governments to consider rural electrification on a national basis. Today in the most advanced countries more than 90 percent of all rural premises—domestic, commercial, industrial, and farms—have an adequate supply of electricity.

Early applications of electricity were of necessity restricted to power and some lighting, although the full value of lighting was not completely realized for many

years. Electric motors were used to drive barn machinery, chaffcutters and root cutters, cattle-cake and grain crushers, and water pumps. Electricity's ease of operation and low maintenance requirements soon showed considerable savings in time and labour. It was not long before the electric motor began to replace the mobile steam engine on threshing, winnowing, and other crop-processing equipment outside the barn.

In the fields, a number of electrically driven, rope-haulage plowing installations, some of them quite large, came into use in several European countries. These systems, however, did not stand the test of time or competition from the mobile internal-combustion-driven tractor.

Applications of electricity in agriculture did not increase greatly until the 1920s, when economic pressures and the increasing drift of labour from the land brought about a change in the whole structure of agriculture. This change, based on new techniques of intensive crop production resulting from the development of a wide range of mechanical, electrical and electromechanical equipment, was the start of the evolution of agriculture from a labour-intensive industry to the capital-intensive industry of today, and in this electricity played a major part.

Modern applications. Modern applications of electricity in farming range from the comparatively simple to some as complex as those in the manufacturing industries. They include conditioning and storage of grain and grass; preparation and rationing of animal feed; and provision of a controlled environment in stock rearing houses for intensive pig and poultry rearing and in greenhouses for horticultural crops. Electricity plays an equally important part in the dairy farm for feed rationing, milking, and milk cooling; all these applications are automatically controlled.

The engineer and farmer have combined to develop electrically powered equipment for crop conservation and storage to help overcome weather hazards at harvesttime and to reduce labour requirements to a minimum. Grain can now be harvested in a matter of days instead of months and dried to required moisture content for prolonged storage by means of electrically driven fans and, in many installations, with electrical heaters. Wilted grass, cut at the stage of maximum feeding value, can be turned into high-quality hay by forced ventilation in the barn with very little risk of spoilage loss from inclement weather.

Conditioning and storage of such root crops as potatoes, onions, carrots, and beets, in especially designed stores with forced ventilation and temperature control, and of fruit in refrigerated stores are all electrically based techniques that minimize waste and maintain top quality over longer periods than was possible with traditional methods.

Controlling storage temperatures

The most significant changes in the pattern of agriculture over recent years have been the degree to which specialization has been adopted and the scale of farm enterprises. Beef cattle by the thousand are raised in enclosures and fed carefully balanced rations by automatic equipment (Figure 9). Pigs by the thousand and poultry by the tens of thousands are housed in special buildings with controlled environments and feed automatically with complex rations.

Dairy herds of up to 1,000 cows are machine-milked in milking parlours, then individually identified and fed appropriate rations by complex electronic equipment. The milk passes direct from the cow into refrigerated bulk milk tanks ready for immediate shipment.

(A.W.Gr.)

PEST AND DISEASE CONTROL IN CROPS

Beginnings of pest control. Wherever agriculture has been practiced, pests have attacked, destroying part or even all of the crop. In modern usage, the term pest includes animals (mostly insects), fungi, plants, bacteria, and viruses. Man's efforts to control pests have a long history. Even in Neolithic times (about 7000 BC), farmers practiced a crude form of biological pest control involving the more or less unconscious selection of seed from resistant plants. Severe locust attacks in the Nile Valley during the 13th century BC are dramatically described in the Bible, and, in his *Natural History*, the Roman author Pliny the Elder describes hand picking of insects from plants and spraying. But not until the 17th and 18th centuries was scientific study of pests undertaken. The first successful large-scale conquest of a pest by chemical means was the control of the vine powdery mildew (*Unciluna necator*) in Europe in the 1840s. The disease, brought from the New World, was controlled first by spraying with lime sulfur and, subsequently, by sulfur dusting.

First successful chemical pest control

Another serious epidemic was the potato blight that caused famine in Ireland in 1845 and some subsequent years and severe losses in many other parts of Europe and the United States. Insects and fungi from Europe became serious pests in the United States, too. Among these were the European corn borer, the gypsy moth, and the chestnut blight, which practically annihilated that tree.

The first book to deal with pests in a scientific way was John Curtis' *Farm Insects*, published in 1860. Though

By courtesy of New Holland Division of Sperry Rand Corp., Pennsylvania

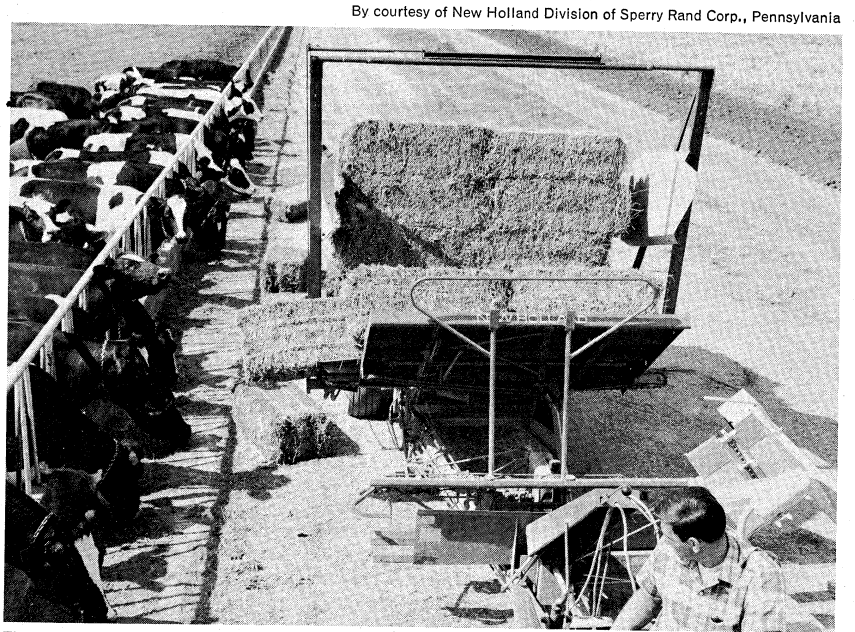


Figure 9: Automatic bale wagons load, haul, and distribute individual bales of hay into cattle feedlots.

farmers were well aware that insects caused losses, Curtis was the first writer to call attention to their significant economic impact. The 19th century also saw a successful battle for control of the Colorado beetle (*Leptinotarsa decemlineata*) of the western United States. When miners and pioneers brought the potato into the Colorado region, the beetle fell upon this crop and became a severe pest, spreading steadily eastward and devastating crops, until it reached the Atlantic. It crossed the ocean and eventually established itself in Europe. But a U.S. entomologist in 1877 found a practical control method consisting of spraying with water-insoluble chemicals such as London Purple, paris green, and calcium and lead arsenates.

Other pesticides soon developed included nicotine, pyrethrum, derris, quassia, and tar oils, first used in 1870 against the winter eggs of the *Phylloxera* plant louse. The Bordeaux mixture fungicide (copper sulfate and lime), discovered accidentally in 1882, was used successfully against vine downy mildew and is still employed to combat it and potato blight. Since many insecticides available in the 19th century were comparatively weak, other pest-control methods were used as well. A ladybird beetle was imported from Australia to California, where it controlled the cottony-cushion scale, then threatening to destroy the citrus industry. A moth introduced into Australia destroyed the prickly pear, which had made millions of acres of pasture useless for grazing. In the 1880s the European grapevine was saved from destruction by *Phylloxera* by the simple expedient of grafting it onto certain resistant American rootstocks.

This period is characterized by increasing awareness of the possibilities of avoiding losses from pests, by the rise of firms specializing in pesticide manufacture, and by development of better application machinery.

Pesticides as a panacea: 1942–62. In 1942 Dr. Paul Müller of Switzerland discovered the insecticidal properties of a synthetic chlorinated organic chemical, dichlorodiphenyltrichloroethane, first synthesized in 1874 and subsequently known as DDT. Müller received the Nobel Prize for his discovery. DDT was far more persistent and effective than any previously known insecticide. Originally a mothproofing agent for clothes, it soon found use among the armies of World War II for killing body lice and fleas. It stopped a typhus epidemic threatening Naples. Müller's work led to discovery of other chlorinated insecticides, including aldrin, introduced in 1948; chlordane (1945); dieldrin (1948); endrin (1951); heptachlor (1948); methoxychlor (1945); and toxaphene (1948).

Origin of DDT

Research on poison gas in Germany during World War II led to the discovery of another group of yet more powerful insecticides and acaricides (killers of ticks and mites)—the organophosphorus compounds, some of which had systemic properties; that is, the plant absorbed them without harm and became itself toxic to insects. The first systemic was octamethylpyrophosphoramide, trade named Schradan. Other organophosphorus insecticides of enormous power were also made, the most common being diethyl-*p*-nitrophenyl monothiophosphate, named parathion. Though low in cost, these compounds were highly toxic to man and warm-blooded animals. Moreover, the products can poison by absorption through the skin, as well as through the mouth or lungs. Thus spray operators must wear respirators and special clothing. Systemic insecticides need not be carefully sprayed, however; the compound may even be absorbed by watering the plant.

Though the advances made in the fungicide field in the first half of the 20th century were not as spectacular as those made with insecticides and herbicides, certain dithiocarbamates, methylthiuram disulfides, and thaladimides were found to have special uses. It began to seem that almost any pest, disease, or weed problem could be mastered by suitable chemical treatment. Farmers foresaw a pest-free millennium. Crop losses were cut sharply; locust attack was reduced to a manageable problem; and the new chemicals, by killing carriers of human disease, saved the lives of millions of people.

Problems appeared in the early 1950s. In cotton crops standard doses of DDT, parathion, and similar pesticides were found ineffective and had to be doubled or trebled. Resistant races of insects had developed. In addition, the powerful insecticides often destroyed natural predators and helpful parasites along with harmful insects. Insects and mites can reproduce so rapidly that often the destruction of natural predators meant that a few survivors from a treatment, unchecked in breeding, soon produced worse outbreaks of pests or a population explosion to pest status of previously harmless insects. Concern began to be expressed about the presence of pesticide residues in food, man, and wildlife. It was found that many large birds retained considerable quantities of DDT in their bodies, accumulated along their natural food chains. This disquiet was epitomized in 1962 by the publication in the United States of a book entitled *Silent Spring*, whose author, Rachel Carson, attacked the indiscriminate use of pesticides, drew attention to various abuses, and stimulated a reappraisal of pest control. Thus began a new "integrated" approach, which was in effect a return to the use of all methods of control in place of a reliance on chemicals alone.

Integrated control. Research into biological methods was undertaken by governments; commercial firms could not help, since there was no end product to sell. In many countries, however, plant breeders began to patent new pest-resistant plant varieties.

Biological pest control

One method of biological control involved the breeding and release of males sterilized by means of gamma rays. Though sexually potent, such insects have inactive sperm. Released among the wild population, they mate with the females, who either lay sterile eggs or none at all. The method was used with success against the screwworm, a pest of cattle, in Texas. A second method employed lethal genes. It is sometimes possible to introduce a lethal or weakening gene into a pest population, leading to the breeding of intersex (effectively neuter) moths or a predominance of males. Experiments are now underway to reduce the gypsy moth by this means. Various studies were also initiated leading to chemical identification of substances attracting pests to the opposite sex or to food, thus enabling traps to be set up only attracting the pest species. Finally, certain chemicals were fed to insects to sterilize them. Used in connection with a food lure, these could lead to the elimination of a pest from an area. Chemicals tested so far, however, are considered too dangerous to man and mammals for any general use.

Some countries (notably the United States, Sweden, and the United Kingdom) have now partly or wholly banned the use of DDT because of its persistence and accumulation in human body fat (though whether it is doing any harm there is not yet known) and its effect on wildlife. New pesticides of lesser human toxicity have now been found, one of the most used being mercaptosuccinate, trade named Malathion. A recent important discovery is the systemic fungicide, absorbed by the plant and transmitted throughout it, making it resistant to certain diseases.

The majority of pesticides are sprayed on crops as solutions or suspensions in water. Spraying machinery has developed from the small hand syringes and "garden engines" of the 18th century to the very powerful "autoblast machines" of the 1950s, applying up to some 400 gallons per acre (4,000 litres per hectare). Though spraying suspended or dissolved pesticide was effective, it involved moving a great quantity of inert material for only a small amount of active ingredient. Low-volume spraying was invented about 1950, particularly for the application of herbicides, in which ten or 20 gallons (40 or 80 litres) of water, as fine drops, would carry the pesticide. Recently, ultralow-volume spraying has been introduced; four ounces (about 110 grams) of the active ingredient itself (usually Malathion) are applied to an acre (0.4 hectare) from aircraft. The spray as applied is invisible to the naked eye. (G.O.)

ECONOMICS, POLITICS, AND AGRICULTURE

As the 20th century opened, western Europe was recovering from a depression during which most of the nations

had turned to protecting agriculture through tariffs. The major exceptions were Great Britain, Denmark, and The Netherlands. The first decade of the century saw an increasing demand for agricultural products, a result of industrialization and population growth, while agriculture in the United States, France, Germany, and many other nations reached a state of equilibrium in relation to the rest of the economy.

During World War I farmers benefitted from high prices, which persisted for a short period after the war because of a shortage of food supplies. European agriculture recovered from wartime losses in land fertility, livestock, and capital, however, and by the early 1920s was fully productive. Elsewhere, production had expanded greatly during the war and afterward continued to rise. After the war prices declined, and, as farmers sought to compensate by expanding their output, the fall in prices was aggravated.

Agricultural tariffs, generally suspended during the war, were gradually reintroduced. The Great Depression of the 1930s brought a new wave of protectionism, leading some industrial nations to look toward self-sufficiency in food supplies. In countries such as France, Germany, and Italy, where agriculture was already protected, the tariff structure was reinforced by new and more drastic measures, while countries such as Britain, Denmark, The Netherlands, and Belgium abandoned free trade and began to support their farmers in a variety of ways. The United States first raised tariffs and then undertook to maintain the prices of farm products. Major exporters of farm products, such as Argentina, Brazil, Australia, Canada, and New Zealand, tried a number of plans to maintain prices.

The milling ratio

One of the most effective of the nontariff measures was the "milling ratio" for wheat or, less often, rye, under which millers were legally obliged to use a certain minimum percentage of domestically produced grain in their grist. This device was first used in Norway in 1927. In 1929, both France and Germany adopted it, and, from 1930 on, it became customary in Europe, and in some non-European nations as well. Milling ratios continued in many nations up to World War II and have been reintroduced in several since that time.

Import quotas, although used earlier in a limited way, became a major protective device during the 1930s. Starting with France, by 1939 they had been adopted on a large scale, mainly for agricultural products, by most European and several non-European nations.

The most radical measures were undertaken in Germany under Hitler, where the Nazi government, seeking self-sufficiency in food, fixed farm prices at rather high levels and maintained complete control over imports.

In the fall of 1931, Great Britain abandoned the free-trade policies followed for nearly a century and turned to protectionism. The new policies emphasized industrial rather than agricultural products. A commodity-by-commodity approach for agriculture through subsidies, marketing schemes, and import restrictions provided help for some farmers. Competition from the Dominions was strong, however.

In 1932 Great Britain negotiated the Ottawa Agreements with the Dominions, which included comparatively free access of Dominion agricultural products to the British market. The agreement helped considerably to alleviate the difficulties the Dominions faced during the Depression. In addition, export subsidies were adopted by several of the Dominions: for wheat by Canada and Australia; for dairy products by New Zealand; and for several products by South Africa.

In the United States, the Agricultural Marketing Act of 1929 established the Federal Farm Board to promote orderly marketing through cooperatives. The Tariff Act of 1930 imposed virtually exclusive tariffs on many products, including those of the farm. These acts, comparatively ineffective in meeting the pressures of the Great Depression, were followed by the Agricultural Adjustment Act of 1933 and subsequent legislation that included restoring the balance between production and consumption, restoration of farm purchasing power, and the attainment of parity, the ratio between prices farmers paid and those they received during the period 1910–14.

Some of the exporting nations adopted extreme measures during the Depression in an attempt to maintain prices for their commodities. Brazil burned surplus coffee stocks, destroying over 8,000,000,000 pounds (3,600,000,000 kilograms) of coffee over ten years beginning in 1931. An Inter-American Coffee Agreement, signed in 1940, assigned export quotas to producer nations for shipment to the United States and other consuming nations and was effective during World War II. Other commodity agreements have met with very limited success.

During World War II, agricultural production declined in most of the European countries, shipping became difficult, and trade channels shifted.

In contrast, agriculture in the United States, undisturbed by military action and with assurance of full demand and relatively high prices, increased productivity. The United States, Great Britain, and Canada cooperated in a combined food board to allocate available supplies. The United Nations Relief and Rehabilitation Administration (UNRRA) was organized in 1943 to administer postwar relief, while the Food and Agriculture Organization (FAO) of the United Nations, established in 1945, was to aid in agricultural development throughout the world.

United Nations activity

By the end of World War II, food production in most of the countries of Europe had fallen below the prewar level. Through assistance given primarily by the United States and United Nations recovery was rapid. Western Europe was greatly helped from 1948 on by United States aid under the Marshall Plan, administered through the Organization for European Economic Co-operation (OEEC). In September 1961, this organization was replaced by the Organization for Economic Co-operation and Development (OECD), with Canada and the United States, and later Japan, as full members. Its present agricultural programs deal, for example, with economic policies, standardization, manpower, and development.

During the 1950s and 1960s, agricultural production increased markedly in western Europe, in the United States, and in most of the nations that ordinarily export large quantities of agricultural products. The Soviet Union increased production, in part by opening new lands, and Chinese agriculture became more productive.

Most of the Western nations offer some type of protection to their farmers—price supports, import quotas, and plans for handling surplus production. On the other hand, many of the developing nations have had food deficits, with little in the way of exportable goods to pay for food imports. Several regional and international organizations have been established in an effort to deal with both types of problems.

Individual farmers in the nations where commercial agriculture is important have been forced to make changes to meet the problems of world surpluses and low world prices for farm products. Thus, in most nations, farmers have increased productivity through adopting advanced technology. This has permitted each worker, generally speaking, to farm larger areas. In some nations, commercialization has led to farming by large-scale corporations. Thus, the world tendency is toward larger farms. The farm operated by a single family, however, is still the dominant unit of production in most of the world.

Since the farmer is so vitally affected by both economics and politics, his problems have become a part of economic and legislative planning and action. Many authorities argue that the decentralization of industry, possible through modern communications, transport facilities, and sources of energy, would provide people with a choice of where to live and how to make a living and would make a vital contribution to the problems of agriculture by providing surplus labour a better opportunity to transfer out of farming. (W.D.R.)

BIBLIOGRAPHY

Origins of agriculture and early agriculture societies: SIR JOSEPH HUTCHINSON (ed.), *Essays on Crop Plant Evolution* (1965), an account of the early developments of maize, sorghum, the temperate cereals, potatoes, forage grasses, and legumes; GRAHAME CLARK, *World Prehistory*, 2nd ed. (1969),

an authoritative outline of prehistory on a worldwide scale; E.S. HIGGS (ed.), *Papers in Economic Prehistory* (1972), contains a reconsideration of theories on the origins of agriculture and some results of recent research; R.S. MACNEISH, "The Origins of American Agriculture," *Antiquity*, 39:87–94 (1965), a brief account of the origins of New World agriculture; PING-TI-HO, "The Loess and the Origin of Chinese Agriculture," *Am. Hist. Rev.*, 75:1–36 (1969), a consideration of the evidence for early agriculture in China; HENRI FRANKFORT, *The Birth of Civilization in the Near East* (1951), an account of agriculture at the beginning of civilization in the Old World.

Agriculture in ancient Asia: Si-yu-ki: Buddhist Records of the Western World, trans. by SAMUEL BEAL (1884, reprinted 1968), travel accounts of early Chinese Buddhist pilgrims to India in the first millennium including Shih-Fa-hian, Sung-Yün, and Hiuen Tsiang; MABEL PING-HUA LEE, *The Economic History of China* (1921), a history of Chinese agriculture with emphasis on soil depletion—first part deals with history, and the second translates relevant source selections from the "Chinese Encyclopaedia" with comments; D.C. TWITCHETT, *Land Tenure and the Social Order in T'ang and Sung China* (1962), a brief but critical and scholarly survey of the development of Chinese studies since 1825, with special discussion of research on the social and economic history of the T'ang and Sung periods; KWANG-CHIH CHANG, *The Archaeology of Ancient China*, rev. ed. (1968), a modern text interpreting prehistoric and protohistoric archaeological evidence in the historical framework of cultural development until 221 BC (illustrated, with bibliography); PING-TI-HO, *Studies on the Population of China, 1368–1953* (1959), a scholarly study of population growth and of interacting variables, such as migrations, land utilization and tenure, and food-production techniques (extensive data tables, bibliography, and notes); N.I. VAVILOV, "The Origin, Variation, Immunity and Breeding of Cultivated Plants," *Chronica Botanica*, vol. 13 (1949–50), selected writings of Vavilov, one of the world's outstanding contributors to the theory of genetics, plant breeding, and study of plant variation, systematics, and evolution (illustrated, with selected bibliography); V. GORDON CHILDE, *New Light on the Most Ancient East*, 4th ed. (1952, reprinted 1969), an excellent introduction to the Indus Valley civilization in comparison with the prehistories of Egypt and Mesopotamia—from the Nile to the Indus (illustrated, with bibliography); BRIDGET and RAYMOND ALLCHIN, *The Birth of Indian Civilization* (1968), a useful general history from the Early Stone Age until 550 BC (maps, plates, list of radioactive-carbon dates, and bibliography); J.W. MCCRINDLE, *Ancient India As Described by Megasthenes and Arrian* (1926), the first translation from the Greek, originally published in 1877 (introduction, notes, and map of ancient India); IFRAN HABIB, *The Agrarian System of Mughul India (1556–1707)* (1963), an excellent modern text on the subject that covers cultivation techniques, crops, land tenure, village communities, and revenue administration (bibliography, contemporary and modern, under subject headings).

Agriculture in the West from 200 BC to AD 1900: G.E. FUSSELL, *Farming Technique from Prehistoric to Modern Times* (1966), a general review of the history of agricultural tools and techniques with many illustrations and an extensive bibliography; JEAN-PHILIPPE LEVY, *L'Économie antique* (1964; Eng. trans., *The Economic Life of the Ancient World*, 1964), a general description of the economies of the Greco-Oriental world before Alexander, the Hellenistic Age, early Roman Empire, and the later Roman Empire; FRITZ M. HIECHELHEIM, *Wirtschaftsgeschichte des Altertums* (1938; Eng. trans., *An Ancient Economic History*, 2nd ed., 3 vol., 1958–70), extensive and detailed information on ancient agriculture; see also standard editions of such classical authors as Cato, Columella, Hesiod, Pliny, Varro, and Xenophon; GEORGES DUBY, *L'Économie rurale et la vie des campagnes dans l'Occident médiéval*, 2 vol. (1962; Eng. trans., *Rural Economy and Country Life in the Medieval West*, 1968), a classic work on agriculture from the 9th to 15th centuries; see also ROBERT LATOUCHE, *Les Origines de l'économie occidentale, IVe–XIe siècles* (1956; Eng. trans., *The Birth of Western Economy*, 1961); LYNN WHITE, JR., *Medieval Technology and Social Change* (1962); B.H. SLICHER VAN BATH, *De agrarische Geschiedenis van West-Europa, 500–1850* (1960; Eng. trans., *The Agrarian History of Western Europe, A.D. 500–1850*, 1963); WILHELM ABEL, *Geschichte der deutschen Landwirtschaft vom frühen Mittelalter bis zum 19. Jahrhundert* (1962); RAYMOND GROMAS, *Histoire agricole de la France des origines à 1939* (1947); MARC BLOCH, *Les Càracteres originaux de l'histoire rurale francaise*, new ed., 2 vol. (1952–56); and PAUL LINDEMANS, *Geschiedenis van de Landbouw in België*, 2 vol. (1952); CHRISTABEL S. ORWIN and EDITH H. WHETHAM, *History of British Agriculture, 1846–1914* (1964), deals with agricultural policy, social change, and technical and scientific developments during this period in England; LORD ERNLE (R.D. PROTHERO), *English Farming Past and Present*, 6th ed. (1961), a classic work describing six centuries of British agriculture; FREIHERR VON DER GOLTZ, *Geschichte der deutschen Landwirtschaft*, 2 vol. (1902–03), another classic, but deals with agriculture in Germany; see also ALBERT DEMOLON, *L'Évolution scientifique et l'agriculture francaise* (1946); MICHEL AUGE-LARIBE, *La Révolution agricole* (1955); Z.W. SNELLER (ed.), *Geschiedenis van de Nederlandse Landbouw, 1795–1940*, 2nd ed. (1951); HEINZ HAUSHOFER, *Die deutsche Landwirtschaft im technischen Zeitalter* (1963); CESARE LONGOBARDI, *Land-Reclamation in Italy* (1936); MARIO BANDINI, *Cento anni di storia agraria italiana* (1957); L.C. GRAY and E.K. THOMPSON, *History of Agriculture in the Southern United States to 1860*, 2 vol. (1933, reprinted 1941); P.W. BIDWELL and JOHN I. FALCONER, *History of Agriculture in the Northern United States, 1620–1860* (1941); N.S.B. GRAS, *A History of Agriculture in Europe and America*, 2nd ed. (1940, reprinted 1968); REYNOLD M. WIK, *Steam Power on the American Farm* (1953); WAYNE D. RASMUSSEN (ed.), *Readings in the History of American Agriculture* (1960); CLARK C. SPENCE, *God Speed the Plow: The Coming of Steam Cultivation to Great Britain* (1960); and RONALD H. CLARK, *The Development of the English Traction Engine* (1960).

The 20th century: ARCHIE A. STONE and HAROLD E. GULVIN, *Machines for Power Farming*, 2nd ed. (1967), a fine general text describing modern farm machinery, including power sources, tilling-equipment applicators for fertilizer, seeders, pest-control machinery, and harvesting equipment; DANIEL FAUCHER, *Le Paysan et la machine* (1954), a history of farm machinery; HUBERT MARTIN, *The Scientific Principles of Crop Protection*, 5th ed. (1964), a classic work by the world's leading authority on pesticides; GEORGE ORDISH, *Untaken Harvest* (1951), the first book to describe the economics of plant losses and their control. See also the bibliographies for the articles AGRICULTURE, TECHNOLOGY OF; IRRIGATION AND DRAINAGE; PEST CONTROL; ECONOMIC DEVELOPMENT; and AGRICULTURAL ECONOMICS.

(E.S.H./G.E.F./K.Na./W.D.R./A.W.Gr./G.O.)

Agriculture, Technology of

Agricultural technology can best be defined as the application of techniques to control the growth and harvesting of animal and vegetable products. This technology has a long history, dating back to the beginnings of mankind. The present article is concerned primarily with relatively modern techniques; historical aspects are covered in AGRICULTURE, HISTORY OF.

Many of the specialized techniques of agriculture are covered in greater detail in other articles. For example, for ornamental vegetation, see HORTICULTURE; for irrigation technology, see IRRIGATION AND DRAINAGE; for control of pests, see PEST CONTROL; for control of weeds, see WEED CONTROL. Likewise, the special field of stock farming is covered in LIVESTOCK AND POULTRY FARMING; breeding is discussed in two articles: ANIMAL BREEDING; PLANT BREEDING. Many of the articles on individual classes of farm products contain some information on farming techniques; included are TOBACCO PRODUCTION; FIBRES, NATURAL; VEGETABLES AND VEGETABLE FARMING; FRUITS AND FRUIT FARMING; CEREALS AND OTHER STARCH PRODUCTS.

This article is divided into the following sections:

I. General techniques
  Soil preparation
    Tilling
    Fertilizing and conditioning soil
  Cropping systems
    Crop rotation
    Monoculture
  Crop protection
    Chemical control of insects
    Nonchemical control of insects
    Control of plant diseases and nematodes
    Weed control
II. Specialized crop-farming techniques
  Dryland farming
    Fallow system and tillage techniques
    Fertilizer use
    Crops and planting methods
  Tropical farming
    Plant-pest problems
    Water management
    Mechanical problems

Other specialized techniques
Hydroponics
Greenhouses
III. Harvesting and crop processing
Harvesting machinery
Crop-processing machinery
IV. Weather and agricultural technology
Weather information
Other weather-research benefits
Observing climatic elements
The degree day
Weather effects
Solar radiation
Photosynthesis
Photoperiodism
Weather conditions and controls
Temperature
Frost
Irrigation
Rainmaking
Humidity
Wind
V. Pollution and agricultural technology
Air pollution
Air pollution damage to agriculture
Air pollution by agriculture
Soil and water pollution
Pollutants damaging to agriculture
Pollutants from agriculture
Monitoring pesticides
VI. The future of agricultural technology
Agricultural experimentation and research
Crop research and development
Livestock research and development
Sponsors of research
Agricultural technology and the world food crisis
Expansion of croplands
Utilization of forest and grazing land
Other possible food sources
Technology and population growth

I. General techniques

SOIL PREPARATION

Mechanical processing of soil so that it is in the proper physical condition for planting is usually referred to as tilling; adding nutrients and trace elements is called fertilizing. Both processes are important in agricultural operations.

Tilling. Tillage is the manipulation of the soil into a desired condition by mechanical means; tools are employed to achieve some desired effect (such as pulverization, cutting, or movement). Soil is tilled to change its structure, to kill weeds, and to manage crop residues. Soil-structure modification is often necessary to facilitate the intake, storage, and transmission of water and to provide a good environment for seeds and roots. Elimination of weeds is important, because they compete for water, nutrients, and light. Crop residues on the surface must be managed in order to provide conditions suitable for seeding and cultivating a crop.

Generally speaking, if the size of the soil aggregates or particles is satisfactory, preparation of the seedbed will consist only of removing weeds and the management of residues. Unfortunately, the practices associated with planting, cultivating, and harvesting usually cause destruction of soil structure. This leaves preparation of the seedbed as the best opportunity to create desirable structure, in which large and stable pores extend from the soil surface to the water table or drains, ensuring rapid infiltration and drainage of excess or free water and promoting aeration of the subsoil. When these large pores are interspersed with small ones, the soil will retain and store moisture also.

Desirable soil structure

Seedbed-preparation procedures depend on soil texture and the desired change in size of aggregates. In soils of coarse texture, tillage will increase aggregate size, provided it is done when only the small pores are just filled with water; tillage at other than this ideal moisture will make for smaller aggregates. By contrast, fine-textured soils tend to form clods; these require breakage into smaller units by weathering or by machines. If too wet or too dry, the power requirements for shattering dry clods or cutting wet ones are prohibitive when using tillage alone. Thus, the farmer usually attempts tillage of such soils only after a slow rain has moistened the clods and made them friable (crumbly).

Some soils require deepening of the root zone to permit increased rate of water intake and improved storage. Unfavourable aeration in zones of poor drainage also limits root development and inhibits use of water in the subsoil.

Tillage, particularly conventional plowing, may create a hardpan, or plow sole; that is, a compacted layer just below the zone disturbed by tillage. Such layers are more prevalent with increasing levels of mechanization; they reduce crop yields and must be shattered, allowing water to be stored in and below the shattered zone for later crops.

Primary tillage equipment. The equipment used to break and loosen the soil for a depth of six to 36 inches (15 to 90 centimetres) is sometimes called primary tillage equipment. It includes moldboard, disk, rotary, chisel, and subsoil plows.

The moldboard plow

The moldboard plow is adapted to the breaking of many soil types. It is well suited for turning under and covering crop residues. There are hundreds of different designs, each intended to function best in performing certain tasks in specified soils. The part that breaks the soil is called the bottom or base; it is composed of the share, the landside, and the moldboard.

When a bottom turns the soil, it cuts a trench, or furrow, throwing to one side a ribbon of soil that is called the furrow slice. When plowing is started in the middle of a strip of land, a furrow is plowed across the field; on the return trip, a furrow slice is lapped over the first slice. This leaves a slightly higher ridge than the second, third, and other slices. The ridge is called a back furrow. When two strips of land are finished, the last furrows cut leave a trench about twice the width of one bottom, called a dead furrow. When land is broken by continuous lapping of furrows, it is called flat broken. If land is broken in alternate back furrows and dead furrows, it is said to be bedded or listed.

Different soils require different-shaped moldboards in order to give the same degree of pulverization of the soil. Thus, moldboards are divided into several different classes, including stubble, general-purpose, general-purpose for clay and stiff-sod soil, slat, blackland, and chilled general-purpose (Figure 1). The blackland bottom is used, for example, in areas in which the soil does not scour easily; that is, leave the surface of the plow clean and polished.

By courtesy of John Deere

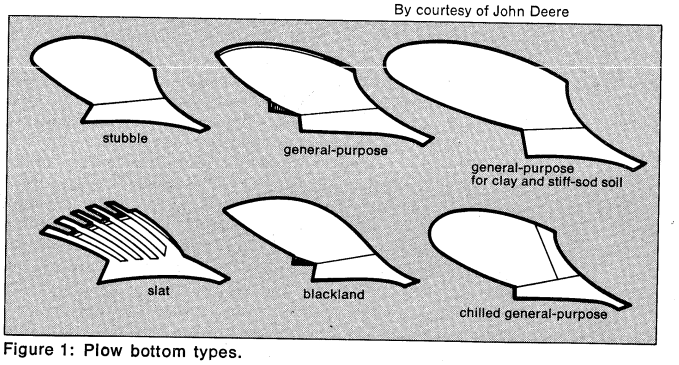


Figure 1: Plow bottom types.

The share is the cutting edge of the moldboard plow. Its configuration is related to soil type, particularly in the down suction, or concavity, of its lower surface. Generally, three degrees of down suction are recognized: regular for light soil, deep for ordinary dry soil, and double-deep for clay and gravelly soils. In addition, the share has horizontal suction, which is the amount its point is bent out of line with the landside. Down suction causes the plow to penetrate to proper depth when pulled forward, while horizontal suction causes the plow to create the desired width of furrow.

Moldboard-plow bottom sizes refer to width between share wing and the landside. Tractor-plow sizes generally

range from ten to 18 inches (25 to 45 centimetres), although larger, special-purpose types exist.

On modern mechanized farms, plow bottoms are connected to tractors either as trailing implements or integrally. One or more bottoms may be so attached (Figure 2). They are found paired right and left occasionally

By courtesy of Robert E. Stewart

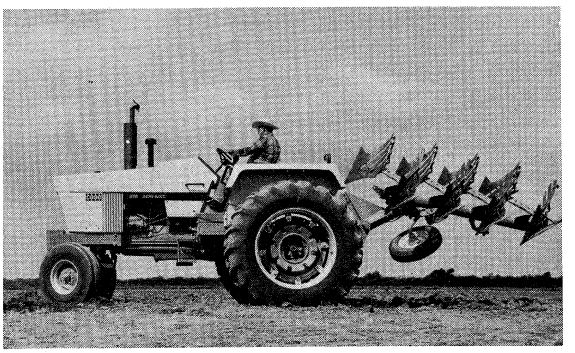


Figure 2: Tractor equipped with integrally mounted four-bottom two-way slat plow.

(two-way), with the advantage of throwing the furrow slice in a constant direction as the turns are made. A variation is the middlebreaker, or lister, which is a bottom equipped with both right- and left-handed moldboards.

The disk plow

The disk plow employs round, concave disks of hardened steel, sharpened and sometimes serrated on the edge, with diameters ranging from 20 to 38 inches (50 to 95 centimetres). It reduces friction by making a rolling bottom in place of a sliding one. Its draft is about the same as that of the moldboard plow. The disk plow works to advantage in situations where the moldboard will not, as in sticky nonscouring soils; in fields with a plow sole; in dry, hard ground; in peat soils; and for deep plowing. The disk-plow bottom is usually equipped with a scraper that aids in pulverizing the furrow slice. Disk plows are either trailed or mounted integrally on a tractor.

The rotary plow's essential feature is a set of knives or tines rotated on a shaft by a power source. The knives chop the soil up and throw it against a hood that covers the knife set. These machines can create good seedbeds, but their high cost and extra power requirement have limited general adoption, except for the small garden tractor.

The chisel plow is equipped with narrow, double-ended shovels, or chisel points, mounted on long shanks. These points rip through the soil and stir it but do not invert and pulverize as well as the moldboard and disk plows. The chisel plow is often used to loosen hard, dry soils prior to using regular plows; it is also useful for shattering plow sole.

Subsoil plows are similar in principle but are much larger, since they are used to penetrate soil to depths of 20 to 36 inches (50 to 90 centimetres). Tractors of 60 to 85 horsepower are required to pull a single subsoil point through a hard soil at a depth of 36 inches. These plows are sometimes equipped with a torpedo-shaped attachment for making subsurface drainage channels.

Secondary tillage. Secondary tillage, to improve the seedbed by increased soil pulverization, to conserve moisture through destruction of weeds, and to cut up crop residues, is accomplished by use of various types of harrows, rollers, or pulverizers, and tools for mulching and fallowing. Used for stirring the soil at comparatively shallow depths, secondary-tillage equipment is generally employed after the deeper primary-tillage operations; some primary tillage tools, however, are usable for secondary tillage. There are five principal types of harrows: the disk, the spike-tooth, the spring-tooth, the rotary cross-harrow, and the soil surgeon. Rollers, or pulverizers, with V-shaped wheels make a firm and continuous seedbed while crushing clods. These tools often are combined with each other.

Harrows and pulverizers

When moisture is scarce and control of wind and water erosion necessary, tillage is sometimes carried out in such a way that crop residues are left on the surface. This system is called trash farming, stubble mulch, or subsurface tillage. Principal equipment for subsurface tillage consists of sweeps and rod weeders. Sweeps are V-shaped knives drawn below the surface with cutting planes horizontal. A mounted set of sweeps provided with power lift and depth regulation is often called a field cultivator.

The typical rod weeder consists of a frame with several plowlike beams, each having a bearing at its point. Rods are extended through the bearings, which revolve slowly under power from a drive wheel. The revolving rod runs a few inches below the surface and pulls up vegetative growth; clearance of the growth from the rod is assisted by its rotation. Rod weeders are sometimes attached to chisel plows.

Some control of weeds is obtained by tillage that leaves the middles between crop rows loose and cloddy. When a good seedbed is prepared only in the row, the seeded crop can become established ahead of the weeds. Plowing with the moldboard plow buries the weed seeds, retards their sprouting, and tends to reduce the operations needed to control them. If weed infestations become bad, they can be reduced somewhat by undercutting.

Since rainfall amount and distribution seldom match crop needs, farmers usually prefer tillage methods that encourage soil-moisture storage at times when crops are not growing. From the soil-moisture standpoint, any tillage practice that does not control weeds and result in greater moisture intake and retention during the storage period is probably unnecessary or undesirable.

Minimum tillage. The use of cropping systems with minimal tillage is usually desirable, because intensive tillage tends to break down soil structure. Techniques such as mulching also help prevent raindrops from injuring the surface structure. Excessive tillage leaves the soil susceptible to crusting, impedes water intake, increases runoff, and thus reduces water storage for crop use. Intensive vegetable production in warm climates where three crops per year may be grown on the same land may reduce the soil to a single-grain structure that facilitates surface cementation and poor aeration.

The loosening and granulating actions of plowing may improve soil structure if done when the moisture content is optimum; if not so timed, however, plowing can create unfavourable structure. The lifting and inversion of the furrow slice likewise may not always be desirable, because in many cases it is better to leave a trashy surface.

The concept of minimum tillage has received much attention. One type of minimum tillage consists in seeding small grain in sod that has been relatively undisturbed. Narrow slits are cut in the sod and seed and fertilizer placed in the breaks thus formed. Soil normally subject to erosion can be planted to grain this way while still retaining the erosion resistance of the sod. The technique has been successful in preparing winter grazing in southeastern portions of the United States. In another type of minimum tillage, the land is broken and planted without further tillage in seedbed preparation. One approach involves breaking the land and planting seeds in the tractor tracks (wheel-track planting); the tractor weight crushes clods and leaves the seed surrounded by firm soil. Another method consists of mounting a planter behind the plow, thus planting without further traffic and leaving a loose seedbed that is satisfactory in areas where post-planting rains may be heavy. In some areas, where winter rain often comes after wheat is drilled, a rotation of wheat following peas has been successful. After the peas have been harvested, the field is rough plowed, and fall wheat is then drilled in directly. All these methods minimize expense and land preparation, tending to leave the soil rough, which reduces erosion and increases water intake. Somewhat similar systems are employed with row crops, where chemical weed control assists in reducing need for cultivation.

Seeding in narrow slits in sod

Mulch tillage. Mulch tillage has been mentioned already; in this system, crop residues are left on the sur-

face, and subsurface tillage leaves them relatively undisturbed. In dryland areas, a maximum amount of mulch is left on the surface; in more humid regions, however, some of the mulch is buried. Planting is accomplished with disk openers that go through several inches of mulch. Since mulch decomposition may deprive the crop of nitrogen, extra fertilizer is often placed below the mulch in humid areas. In rainy sections, intercropping extends the protection against erosion provided by mulches. Intercrops are typically small grains or sod crops such as alfalfa or clover grown between the rows of a field crop that reach maturity shortly after the field crop has been established and furnish mulch cover for a long time (Figure 3).

Intercropping as protection against erosion

By courtesy of the U.S. Department of Agriculture



Figure 3: Intercropping to reduce soil erosion, near Fresno, Ohio.

If growth of the intercrop competes with the main crop for moisture and nutrients, that growth may be killed at seeding time or soon thereafter by undercutting with sweeps.

Tillage in dry areas must make maximum use of scanty rainfall. The lister (double-mold board) plow, or middlebreaker, is here used to make water-impounding ridges that promote infiltration. The special problems of dryland farming will be considered later (see *II. Specialized crop-farming techniques: Dryland farming*).

Fertilizing and conditioning soil. Soil fertility is the quality of a soil that enables it to provide compounds in adequate amounts and proper balance to promote growth of plants when other factors (such as light, moisture, temperature, and soil structure) are favourable. Where fertility of a soil is not good, natural or manufactured materials may be added to supply the needed plant nutrients; these are called fertilizers, although the term is generally applied to largely inorganic materials other than lime or gypsum. Fertilizer grade is a conventional expression that indicates the percentage of plant nutrients in a fertilizer; thus, a 10–20–10 grade contains 10 percent nitrogen, 20 percent phosphoric oxide, and 10 percent potash. The green plant, however, requires more nutrients than these.

Essential plant nutrients. In total, the plant has need of at least 16 elements, of which the most important are carbon, hydrogen, oxygen, nitrogen, phosphorus, sulfur, potassium, calcium, and magnesium.

The plant obtains carbon, hydrogen, and oxygen from water; carbon dioxide and other nutrients are taken up from the soil. Although the plant contains sodium, iodine, and cobalt, these are apparently not essential. This is also true of silicon and aluminum.

Overall chemical analyses indicate that the total supply of nutrients in soils is usually high in comparison with the requirements of crop plants. Much of this potential supply, however, is bound tightly in forms that are not released to crops fast enough to give satisfactory growth. Because of this, the farmer is interested in measuring the available nutrient supply as contrasted to the total quantities. This point will be considered later.

The solid content of soils is broadly classified as organic and inorganic. Materials of organic origin range from fresh plant tissue to the more or less stable black or brown degradation product (humus) formed by biological decay. The organic matter is a potential source of nitrogen, phosphorus, and sulfur; it contains more than 95 percent of the total nitrogen, 5 to 60 percent of the total phosphorus, and 10 to 80 percent of the total sulfur. These three elements are cycled through the entire environment of living things (the biosphere). The soil organic matter can be considered as one of the storage points in these cycles. Where nonlegumes are grown in the absence of fertilizer or manures, the crop must gain its nitrogen supply from the organic matter; only a part, however, of the needed phosphorus and sulfur is so supplied.

Organic and inorganic soils

The inorganic or mineral fraction, which comprises the bulk of most soils, is derived from rocks and their degradation products. The power to supply plant nutrients is much greater in the larger particles, sand and silt, than in the fine particles, or clay. The minerals that comprise the sand and silt in soil contain most of the elements essential for plant growth as a part of their structure. The difficulty is that these minerals decompose so slowly in soil that the rate of supply of the nutrient elements is usually insufficient for good growth of plants.

When the available supply of a given nutrient becomes depleted, its absence becomes a limiting factor in plant growth, and the addition of this nutrient to the soil will increase yields of dry matter. Excessive quantities of some nutrients may cause decrease in yield, however.

Determining nutrient needs. Determination of a crop's nutrient needs is an essential aspect of fertilizer technology. The appearance of a growing crop may indicate need of fertilizer; in some plants, however, the need for more or different nutrients may not be easily observable. If such a problem exists, its nature must be diagnosed, the degree of deficiency must be determined, and the amount and kind of fertilizer needed for a given yield must be found. There is no substitute for detailed examination of plants and soil conditions in the field, followed by simple fertilizer tests, quick tests of plant tissues, and analysis of soils and plants.

Sometimes plants show symptoms of poor nutrition. Chlorosis (general yellow or pale-green colour), for example, indicates lack of sulfur and nitrogen. Iron deficiency produces white or pale-yellow tissue. Symptoms can be misinterpreted, however. Plant disease can produce appearances resembling mineral deficiency, as can various organisms. Drought or improper cultivation or fertilizer application each may create deficiency symptoms.

After field diagnosis, the conclusions may be confirmed by experiments in a greenhouse or by making strip tests in the field. In strip tests, the fertilizer elements suspected of being deficient are added, singly or in combination, and the resulting plant growth observed. Next, it is necessary to determine the extent of the deficiency.

Strip tests of fertilizer elements

An experiment in the field can be conducted by adding nutrients to the crop at various rates. The resulting response of yield in relation to amount of nutrient supplied will indicate the supplying power of the unfertilized soil in terms of bushels or tons of produce. If the increase in yield is large, this practice will show that the soil has too little of a given nutrient. Such field experiments may not be practical, because they can cost too much in time and money. Soil-testing laboratories are available in most areas; they conduct chemical soil tests to estimate the availability of nutrients. Commercial soil-testing kits give results that may be very inaccurate, depending on techniques and interpretation. Actually, the most accurate system consists of laboratory analysis of the nutrient content of plant parts, such as the leaf. The results, when correlated with yield response to fertilizer application in

field experiments, can give the best estimate of deficiency. Further development of remote sensing techniques, such as infrared photography, are under study and may ultimately become the most valuable technique for such estimates.

The economics of fertilizers. The practical goal is to determine how much nutrient material to add. Since the farmer wants to know how much profit to expect if he buys extra fertilizer, the tests are interpreted as an estimation of increased crop production that will result from nutrient additions. Cost of nutrients must be balanced against value of crop or even against alternative procedures, such as investing the money in something else with a greater potential return. The law of diminishing returns is well exemplified in fertilizer technology. Past a certain point, equal inputs of chemicals produce less and less yield increase. The goal of the farmer is to use fertilizer in such a way that the most profitable rate is employed.

Fertilizers can aid in making profitable changes in farming. Operators can reduce costs per unit of production and increase the margin of return over total cost by increasing rates of application of fertilizer on principal cash and feed crops. They are then in a position to invest in soil conservation and other improvements that are needed when shifting acreage from surplus crops to other uses.

Farm manure. Among sources of organic matter and plant nutrients, farm manure has been of major importance in past years. Manure is understood to mean the refuse from stables and barnyards, including both excreta and straw or other bedding material, while the term fertilizer refers to chemicals. Large amounts of manure are produced by livestock; such manure has value in maintaining and improving soil because of the plant nutrients, humus, and organic substances contained in it.

As manure must be managed carefully in order to derive the most benefit from it, some farmers may be unwilling to expend the necessary time and effort. Manure must be carefully stored to minimize loss of nutrients, particularly nitrogen. It must be applied to the right kind of crop at the proper time. Also, additional fertilizer may be needed, such as phosphoric oxide, in order to gain full value of the nitrogen and potash that are contained in manure.

Fertilizer value of manure

Manure is fertilizer graded as approximately 0.5–0.25–0.5 (percentages of nitrogen, phosphoric oxide, and potash), with at least two-thirds of the nitrogen in slow-acting forms. Commercial fertilizer equivalent to one ton (900 kilograms) of average manure can be purchased at a fairly low price. Furthermore, the expense of applying 100 pounds (45 kilograms) of 10–5–10 fertilizer is much less than the cost of applying 20 times as much manure. On properly tilled soils, the returns from fertilizer usually will be greater than from an equivalent amount of manure. The application of manure to a crop cannot be controlled as readily as can granulated fertilizer. In general, manure does not provide all the plant nutrients needed and fails to provide any that cannot be supplied by artificial fertilizers. Thus, there is a tendency to discount the value of manure as fertilizer. In underdeveloped countries, however, where artificial fertilizer may be costly or unavailable and where labour is relatively cheap, manure is attractive as a fertilizer.

The main benefits of manure are indirect. It supplies humus, which improves the soil's physical character by increasing its capacity to absorb and store water, by enhancement of aeration, and by favouring the activities of lower organisms. Manure incorporated into the topsoil will help prevent erosion from heavy rain and slow down evaporation of water from the surface. In effect, the value of manure as a mulching material may be greater than is its value as a source of essential plant nutrients.

Green manuring. In reasonably humid areas, the practice of green manuring can improve yield and soil qualities. A green-manure crop is grown and plowed under for its beneficial effects, although during its growth it may be grazed. These green crops are usually annuals, either grasses or legumes, whose roots bear nodule bacteria capable of fixing atmospheric nitrogen. Among the advantages of green-manure crops are the addition of nitrogen to the soil, increase in general fertility level, reduction of erosion, improvement of physical condition, and reduction of nutrient loss from leaching. Disadvantages include the chance of not obtaining a satisfactory growth; the possibility that the cost of growing the manure crop may exceed the cost of applying commercial nitrogen; possible increases in disease, insects, and nematodes (parasitic worms); and possible exhaustion of soil moisture by the crop.

Advantages of green-manure crops

Green-manure crops are usually planted in the fall and turned under in the spring before the summer crop is sown. Their value as a source of nitrogen, particularly that of the legumes, is unquestioned for certain crops such as potatoes, cotton, and corn; for other crops, such as peanuts (which themselves are legumes), the practice is questionable. Farmers are gradually turning away from growing green-manure crops except where the crop may also serve as winter cover for the land.

Compost, peat, and sludge. Compost, peat, and sludge are used in agriculture and gardening as soil amendments rather than as fertilizers, because they have a low content of plant nutrients. They may be incorporated into the soil or mulched on the surface. Heavy rates of application are common.

Compost, or synthetic manure, is basically a mass of rotted organic matter made from waste-plant residues. Addition of nitrogen during decomposition is usually advisable. The result is a crumbly material that when added to soil does not compete with the crop for nitrogen. When properly prepared, it is free of obnoxious odours. Composts commonly contain about 2 percent nitrogen, 0.5 to 1 percent phosphorus, and about 2 percent potassium; if phosphate and potash are added while composting, those values are higher. The nitrogen of compost becomes available slowly and never approaches that available from inorganic sources. This slow release of nitrogen reduces leaching and extends availability over the whole growing season. Composts are essentially fertilizers with low nutrient content, which explains why large amounts are applied. The maximum benefits of composts on soil structure (better aggregation, pore spacing, and water storage) and on crop yield usually occur after several years of use.

In practical farming, the use of composted plant residues must be compared to the use of fresh residues. More beneficial soil effects usually accrue with less labour by simply turning under fresh residues; also, since one-half the organic matter is lost in composting, fresh residues applied at the same rate will cover twice the area that composted residues would cover. In areas where commercial fertilizers are expensive, labour is cheap, and implements are simple, however, composting meets the needs and is a logical practice.

Peat and peat moss

Peat, composed of prehistoric plant remains that have accumulated under airless conditions in bogs, is a widely used organic soil amendment. Peat moss, the remains of sphagnum plants, is probably its most common form; it contains less than 1 percent nitrogen, with phosphorus and potassium below 0.1 percent. It is highly acid, with pH between 3 and 4.5 (a pH value of 7 is neutral and one above 7 basic). Peat improves the water-storage capability of soils and gives better structure to fine soils. Heavy applications of peat is usually the practice. It is used mostly by specialty-crop producers and on lawns and gardens.

Sewage sludge is the solid material remaining from the treatment of sewage. Its value for soil improvement depends on the method used for treating the sewage. Activated sludge, which results from aerobic (oxygen) treatment, contains 5 to 6 percent nitrogen and 1 to 3.5 percent of phosphorus. After suitable processing, it is sold as fertilizer and soil amendment for use on lawns, parks, and golf courses. It is rarely used in farming.

Liming. Liming to reduce soil acidity is practiced extensively in humid areas where rainfall leaches calcium and magnesium from the soil, thus creating an acid condition. Calcium and magnesium are major plant nutri-

ents supplied by liming materials. Ground limestone is widely used for this purpose; its active agent, calcium carbonate, reacts with the soil to reduce its acidity. The calcium is then available for plant use. The typical limestones, especially dolomitic, contain magnesium carbonate as well, thus also supplying magnesium to the plant.

Another liming material is basic slag, a by-product of steel manufacture; its active ingredient is calcium silicate. Marl and chalk are soft, impure forms of limestone and are sometimes used as liming materials, as are oyster shells. Calcium sulfate (gypsum) and calcium chloride, however, are unsuitable for liming, for, although their calcium is readily soluble, they leave behind a residue that is harmful.

Lime is applied by mixing it uniformly with the surface layer of the soil. It may be applied at any time of the year on land plowed for spring crops or winter grain or on permanent pasture. After application, plowing, disking, or harrowing will mix it with the soil. Such tillage is usually necessary, because calcium migrates slowly downward in most soils. Lime is usually applied by trucks specially equipped and owned by custom operators.

Methods of application. Fertilizers may be added to soil in solid, liquid, or gaseous forms, the choice depending on many factors. Generally, the farmer tries to obtain satisfactory yield at minimum cost in money and labour.

Liquid or solid application of manure

Manure can be applied as a liquid or a solid. When accumulated as a liquid from livestock areas, it may be stored in tanks until needed and then pumped into a distributing machine or into a sprinkler irrigation system. The method reduces labour, but the noxious odours are objectionable. The solid-manure spreader conveys the material to the field, shreds it, and spreads it uniformly over the land. The process can be carried out during convenient times, including winter, but rarely when the crop is growing.

Application of granulated or pelleted solid fertilizer has been aided by improved equipment design. Such devices, depending on design, can deposit fertilizer at the time of planting, side-dress a growing crop, or broadcast the material. Fertilizer attachments are available for most tractor-mounted planters and cultivators and for grain drills and some types of plows. They deposit fertilizer with the seed when planted, without damage to the seed, yet the nutrient is readily available during early growth. Placement of the fertilizer varies according to the types of crops; some crops require banding above the seed, while others are more successful when the fertilizer band is below the seed.

The use of liquid and ammonia fertilizers is growing, particularly of anhydrous ammonia, which is handled as a liquid under pressure but changes to gas when released to atmospheric pressure. Anhydrous ammonia, however, is highly corrosive, inflammable, and rather dangerous if not handled properly; thus, application equipment is quite specialized. Typically, the applicator is a chisel-shaped blade with a pipe mounted on its rear side to conduct the ammonia five to six inches (13 to 15 centimetres) below the surface. Pipes are fed from a pressure tank mounted above. Mixed liquid fertilizers containing nitrogen, phosphorus, and potassium may be applied directly to the surface—by field sprayers where close-growing crops are raised. Large areas can be covered rapidly by use of aircraft, which can distribute both liquid and dry fertilizer.

The future for fertilizers. Future trends in fertilizer technology may be predicted by extrapolating from current developments. Mixtures and materials with high percentages of plant nutrients will dominate the field.

New methods of providing nitrogen

Better ways of providing nitrogen, the most expensive of the three major nutrients, will be forthcoming, including increased use of anhydrous ammonia, ammonium nitrate, and urea. Nonleachable nitrogen, for example, can be obtained through the urea–formaldehyde (urea-form) reaction, and ammonium metaphosphate offers a concentrated liquid product. Micronutrients, or trace elements, specific to particular geographical areas will come into increasing use, as will custom mixing and bulk selling of mixtures containing several nutrients based on reliable soil and plant data.

"Complete environment" seeding in which seed, fertilizer, and water are incorporated in a biodegradable (decomposable in the soil) tape may come into use; with the tape planted, no further fertilizer or water will be needed until growth is well established. Such techniques using biodegradable tapes have already been developed on a small scale for use by home gardeners. Finally, larger and more precise fertilizing machines will be developed and adopted.

CROPPING SYSTEMS

The kind and sequence of crops grown over a period of time on a given area of soil can be described as the cropping system. It may be a pattern of regular rotation of different crops or one of growing only one crop year after year on the same area.

Crop rotation. Early agricultural experiments showed the value of crop rotations that included a legume sod crop in the regular sequence. Such a system generally maintains productivity, aids in keeping soil structure favourable, and tends to reduce erosion. Alfalfa, sweet clover, red clover, and Ladino clover are considered effective for building up nitrogen. Some legumes, however, do not leave nitrogen behind in the soil because it is deposited as protein in the harvested seed; soybeans are an example. Turning under the top growth of a legume aids in adding nitrogen. Though yields of grains are higher when they are rotated with legumes, it is difficult to determine how much of the improvement depends on the nitrogen added by the legume and how much on improved soil structure or fewer insects and disease.

The determination of the best rotation depends upon whether the crops compete with each other (*i.e.*, if growing one crop lowers the yield of its successor) or complement each other; and the output of one crop on a given acreage leads to increased output of the other. This desirable complementary relationship exists only when one crop or soil-management practice concurrent with it provides nutrient or conditions required by the other crop. In this circumstance, grasses and legumes may complement grains or row crops by furnishing nitrogen, controlling erosion and pests, and improving soil structure to such an extent that greater production is achieved. The reverse can also occur; in certain prairie soils, continuous growing of deep-rooted legumes depletes soil moisture, and subsequent forage yield is improved by frequent plowing of the sod and planting of corn. In high-rainfall or irrigated areas, forage stands deteriorate from winter killing, disease, or grazing, to a point where a year of grain in the rotation allows an improved stand of forage later. Fallow (idle) land is complementary to wheat and other small grains in subhumid areas such as the Great Plains of the United States; such rotation is quite beneficial to wheat yield. Complementary relationships between crops can be terminated by the application of the physical law of diminishing returns, however, and give way to competition.

Rotation as a benefit to wheat yield

Both long-range and short-range profits motivate the farmer as cropping systems are examined in relationship to soil erosion. Excessive loss of soil to streams, rivers, and reservoirs is unacceptable to public policy as well as economically damaging to the farmer, and crop rotations that promote erosion are minimized. Soil losses are least from fields in continuous sod and most from continuous row crops. If row crops are grown in rotation with sod, the erosive susceptibility of row crops is reduced over a period of time. Peanuts, potatoes, tobacco, cotton, sugar beets, and some vegetables, and similar row crops that require frequent cultivation (intertillage) and leave minimal post-harvest residue are most likely to permit serious erosion. Less erosive are row crops such as corn (maize), sugarcane, and grain sorghum, which require less cultivation and leave more residue. Small grains such as wheat, oats, barley, and rye usually permit less erosion than the row crops. Among sod crops, grasses or grass–

legume mixtures are less erosive than pure stands of legumes such as alfalfa. Fortunately, cropping systems that tend to control soil erosion usually tend also to give better yields than systems that promote excessive erosion. This results from increased availability of water to the plants and increased amounts of nutrients, which in erosive systems are washed away and lost.

Monoculture. The practice of growing the same crop each year on a given acreage, monoculture, has not been generally successful in the past, because nonlegume crops usually exhaust the nitrogen in the soil, with a resulting reduction in yields; this is particularly true in humid regions. The advent of low-cost nitrogen fertilizers, however, has induced reconsideration of the possible advantages of monoculture. These advantages can best be discussed in terms of a hypothetical general farm where it may be desirable to produce several different kinds of crops: the question to be answered is whether monoculture can do better than rotational systems in producing these crops while still maintaining productivity.

Advantages of monoculture. First, if different kinds of soil exist on the farm, a monoculture system may permit each crop to be grown on the soil best suited to it. Forage crops, for example, could be confined to steep land to minimize erosion; intertilled crops could be planted on the better soils with gentle slopes. Wet areas could be used continuously for crops not requiring early-spring field operations, while dry soils could be used for drought-resistant crops such as sorghums or winter small grains.

Second, the fertility level of the soil can be adjusted to fit one crop more precisely than it can be adjusted to fit all the crops in a rotation.

Third, where continuous cropping is practiced and perennial forage crops are used, regular reseedings are avoided. This is an advantage, because each seeding is accompanied by the possibility of failure.

Flexibility in monoculture crop planning

Fourth, systems based on monoculture usually offer greater flexibility in planning the system to meet year to year changes in the need for various crops. Part of the acreage can be shifted from one crop to another without upsetting the total farm cropping plan.

Disadvantages of monoculture. On the other hand, requirements for successful monoculture are more demanding of management skill than are sod-based rotations. The entire nitrogen need of nonlegume crops must be met by purchased fertilizers or by use of manure. Closer attention to soil erosion is necessary, except for perennial sod. Soil-structure problems can become severe where intertilled crops are grown continuously. In monoculture, the farmer is completely dependent on chemical insecticides, disease-resistant plant varieties, soil fumigation, and similar methods of controlling insects and diseases that are usually controlled by crop rotation.

Thus, the choices of cropping systems that maintain good productivity, minimize soil losses, and are in harmony with demand and desired business organization are not easily made. The growing use of systems analysis will undoubtedly aid in rational selection among the bewildering array of possibilities.

CROP PROTECTION

Crops are vulnerable to attack, damage, and competition. Insects, plant disease, nematodes, rodents, weeds, air pollution, and other enemies reduce crop yields and deny man the use of some of his farm-stored crops.

Insects, for example, can destroy a crop in a relatively short time. Control measures for many years have engaged the attention of farmer and scientist, yet full success has not been achieved, and the battle continues. The problem is further complicated by the fact that control measures not only kill unwanted insects, but also may harm honey bees as well as the parasites and predators that destroy insect pests.

At least 10,000 species of insects are classed as unwanted. Of these, several hundred species are particularly destructive and require some degree of control. They destroy food as well as the forage, pasture, and grain needed to produce livestock; and, in addition, they carry and transmit many diseases of plants and animals (see also PEST CONTROL).

Chemical control of insects. Insecticides generally are effective, cheap, and safe if handled correctly; the good derived from them, however, can be partly offset by adverse effects. Chlorinated hydrocarbon insecticides such as DDT, for example, may leave residues toxic to beneficial insects, fish, and other wildlife; the insecticides may be found in meat and milk, or they may persist in the soil. Another problem is that some species of insects build up resistance to chlorinated hydrocarbon, organic phosphate, and carbamate insecticides. These disadvantages can be overcome only by persistent search for new and safer insecticides accompanied by wide use of nonchemical insect control.

Adverse effects of some insecticides

A wide range of organophosphate and carbamate materials is now available. These can be applied to avoid most of the problems related to residues. Malathion and carbaryl, for example, are used to control insects in areas where persistent materials might appear later in meat or milk and can also be applied in areas where fish and wildlife might be affected. Those two chemicals offer a broad range of toxicity to insect pests. Unlike chlorinated hydrocarbons, they can be applied up to within a day or so of harvest without harm to many crops; they are dangerous, however, to those who apply them and must be handled with care.

Some insecticides are effective in very small amounts. This fact has stimulated development of ultralow-volume technology, where special equipment permits dispersal of low volumes of undiluted chemicals, which offers cost advantages as well as drastic reduction of the chemicals in the environment. For example, six to 16 ounces (170 to 450 grams) per acre of Malathion may be effective against grasshoppers, boll weevil, cereal-leaf beetle (*Oulema melanopus*), mosquitoes, and the beet leafhopper (*Circulifer tenellus*). Formulation of chemicals in granules rather than sprays offers some advantages in use and applications; among others, it reduces the amount needed and also lessens the chance of adverse effects on beneficial insects and wildlife.

Systemic insecticides

Certain insects that attack cotton, vegetables, and forage crops may be controlled by chemicals absorbed by the plant. Called systemics, they are placed with the seed at planting time. The chemical is taken up by the plant, and insects die when they attempt to feed on the leaf or stem. Beneficial insects that do not feed on the plant remain unharmed.

Nonchemical control of insects. *Mechanical and cultural controls.* Light traps that give off radiation that attracts insects have been under test for many years. They have been somewhat successful in controlling the codling moth (*Carpocapsa pomonella*) and the tobacco hornworm (*Protoparce sexta*).

Use of reflective aluminum strips, placed like a mulch in vegetable fields, has reduced or prevented aphid attack and thus protected cucumbers, squash, and watermelons from mosaic diseases. This technique may supplant insecticides, which frequently do not kill aphids quickly enough to prevent crop losses from virus transmitted by them.

For stored products, heat or cold can control many insects that frequent such places. Also, changing the proportions of oxygen, nitrogen, and carbon dioxide in the storage atmosphere can provide control.

Recently, it was discovered that, if adult Indian-meal moths (*Plodia interpunctella*) were exposed to certain wavelengths of sound during the egg-laying period, their reproduction was reduced by 75 percent. The sound waves had a similar effect on flour beetles (*Tribolium* species). Further development is needed, but this method offers potential as a nonchemical control. Other types of physical energy can also kill insects. Light waves, high-frequency electric fields, high-intensity radio frequencies, and gamma radiation have been investigated; some offer promise.

Certain cultural practices can prevent or reduce insect crop damage. These include destruction of crop residues,

deep plowing, crop rotation, use of fertilizers, stripcropping, irrigation, and scheduled planting operations. Such practices are useful but cannot be relied upon entirely to eliminate severe infestations.

Biological controls. The question of using biological controls has always been of considerable public interest. The control agents include parasites, predators, diseases, protozoa, and nematodes that attack the insect pests. Biological controls cannot replace insecticides entirely, because nature provides for survival of both beneficial and destructive insects. Before the population of a parasite or predator can expand, a high population of the host species must also be present. Sometimes the control agents are far outnumbered by the pest insect. Parasites and predators have furnished good control of the Japanese beetle (*Popillia japonica*), European corn borer (*Pyrausta nubilalis*), alfalfa aphid (*Therioaphis maculata*), alfalfa weevil (*Hypera postica*), and several others.

Microbial agents as controls

Microbial agents can be used for control. There exist about 1,100 viruses, bacteria, fungi, protozoa, rickettsiae, and nematodes that parasitize insects. Many pathogens are specific to a particular insect but are harmless to man and domestic animals. It is a possibility that insect pathogens can be produced, packaged, distributed, and applied in much the same way as insecticides.

The ideal solution to insect-control problems is to plant crop varieties that are resistant to attack. The only difficulty is that such varieties are not universally available, and development entails a very long process.

Sterilization of male insects by gamma radiation and their release into a population of wild insects is a promising approach. It has proved successful in control of screwworms and fruit flies, replacing chemicals in some areas. Chemical attractants, which lure insects into contact with small amounts of insecticide or a sterilant, also offer much promise.

All aspects of insect control considered, it is possible that "integrated control," coordinated employment of more than one method, may be the answer. Combining resistant varieties with a systemic insecticide that leaves the parasites and predators unharmed, for example, has been successful in combatting the spotted alfalfa aphid in California. Preliminary reduction of heavy infestation by chemical spray combined with bait, followed by the sterile-insect technique, provides another example of integrated control. Use of sex attractant in light traps, plus special management of postharvest residues, has controlled the tobacco hornworm. Other examples might be cited, but the principal value of such control methods lies in using less insecticide and thus contributing to maintenance of a good environment.

Control of plant diseases and nematodes. Insects, of course, are not the only agents hazardous to crops. Plant diseases and the microscopic worms called nematodes have the potential of creating wholesale destruction of crops, especially those grown in regions of wide weather fluctuation. In fact, these plant pests sometimes limit the kinds and varieties of crops that can be grown. The damage they cause may sometimes be mistaken for that caused by unfavourable weather. Epidemics may destroy crops completely.

As with insects, control of plant diseases and nematodes covers a broad spectrum of measures: use of chemicals, resistant varieties, quarantine, forecasting and warning, cultural practices, heat treatment, and others. Furthermore, most plant virus diseases are transmitted by insect carriers, so control of insects is linked to control of disease.

Nematodes and plant disease can at times be controlled fairly well by crop rotation, deep plowing, and burning of stubble and debris that remain after harvest. Though burning destroys aboveground organisms and permits economical control by chemicals, it contributes to air pollution and destroys organic matter. In another technique, propane-gas flame is applied to living plants as well as stubble to kill disease spores. A virus disease of sugarcane is controlled by heating diseased cuttings in hot-air ovens. Stem rot disease of peanuts can be controlled by plowing under dead plant debris or by planting the seed on a raised bed followed by application of a pre-emergence weed killer.

Predicting plant epidemics

Successful control of epidemic plant disease may depend on prompt application of chemicals before the disease outbreak. Many governments operate plant-disease forecasting and warning services for farmers. The service is based primarily on analysis of temperature, rainfall, humidity, and dew—all factors that can create conditions favourable to disease outbreaks.

Weed control. Weed control is vital to agriculture, because weeds decrease yields, increase production costs, interfere with harvest, and lower product quality. Weeds have other disadvantages, also, as they impede irrigation water-flow, interfere with pesticide application, and harbour disease organisms (see also WEED CONTROL).

Early methods of weed control included mowing, flooding, cultivating, smothering, burning, and crop rotation. Though these methods are still important, other means are perhaps more typical today, particularly the use of herbicide (plant-killing) chemicals. Another technique is to introduce insects that attack only the unwanted plant and destroy it while leaving the crop plants unharmed.

The inadequacy of the cultural, mechanical, and biological control systems, however, stimulated the rapid development of chemical usage since World War II. Herbicides have had an impact on crop production, changing many cultural and mechanical agricultural operations.

Herbicides are formulated as wettable powders, granular materials, emulsions, and solutions. Any of them may be applied as a spot treatment, broadcast, placed in bands, or put directly on a specific plant part. When formulated as solutions or emulsions, the chemical is mixed with water or oil.

Spraying is the most common method, permitting extremely small amounts to be applied uniformly because of dilution. Sprays can be accurately directed underneath growing plants, and calibration and rate control are easier with spray machines than with granular applicators. Granular formulations have advantages under some conditions, however. The use of herbicides must be integrated into the overall farm program because the optimum date and application rate depend on the crop stage, the weed stage, weather conditions, and other factors.

Careful use of herbicides in farm production lowers cost, resulting in a more economical product for the consumer. Herbicides cut the costs of raising cotton, for example, by reducing labour requirements for weed control up to 60 percent. Herbicides replace hand labour in growing crops, labour that is no longer available in developed nations at costs the farmer can afford. Machines for chemical application are widely available.

When used as directed, herbicides are generally safe, not only for the operator but also for wildlife and livestock. The greatest difficulty lies in accidental injury to crop plants resulting from drift and from residues in the soil, particularly if residues enter water courses.

The future of chemical pesticides and herbicides is under debate by those who manufacture, sell, and use them and by those who are concerned about environmental quality. The value of an assured food and fibre supply at reasonable cost is undeniable, and chemicals contribute much toward this. These substances also cause undesirable effects upon the environment, however, and indeed can be toxic to a wide range of organisms. This fact will demand an increasing amount of care in using chemicals, perhaps enforced by law, along with increasing use of nonchemical control techniques.

II. Specialized crop-farming techniques

DRYLAND FARMING

Dryland farming refers to production of crops without irrigation in regions where annual precipitation is less than 20 inches (50 centimetres). Where rainfall is less than 15 inches (40 centimetres) per year, winter wheat is the most favoured crop, although spring wheat is planted in some areas where severe winter killing may occur. (Grain sorghum is another crop grown in these areas.)

Dryland-farming crops

Where some summer rainfall occurs, dry beans are an important crop. All dryland crop yield is mainly dependent on precipitation, but practices of soil management exert great influence on moisture availability and nutrient supply.

Where rainfall exceeds 15 inches, the variety of crop possibilities is increased. In areas of favourable soils and moisture, seed alfalfa is grown, as is barley. Some grass seed may be grown, particularly crested wheat grass of various types.

Fallow system and tillage techniques. Dryland farming is made possible mainly by the fallow system of farming, a practice dating from ancient times. Basically, the term fallow refers to land that is plowed and tilled but left unseeded during a growing season. The practice of alternating wheat and fallow assumes that by clean cultivation the moisture received during the fallow period is stored for use during the crop season. Available soil nitrogen increases and weeds are controlled during the fallow period. One risk lies in the exposure of soil while fallow, leaving it susceptible to wind and water erosion. Modern power machinery has tended to reduce this risk.

Procedures and kinds of tillage that are comparatively new have proved effective in controlling erosion and improving water intake. Moldboard and disk plows are being replaced with chisels, sweeps, and other tools that stir and loosen the soil but leave the straw on the surface. Where the amount of straw or residue remaining from the previous crop is not excessive, this trashy fallow system works well, and tillage implements are designed for it.

Contour tillage helps to prevent excessive runoff on moderate slopes. Broad terraces can aid in such moisture conservation. Steeper slopes are planted to permanent cover.

Improving storage of moisture

Compacted zones at a depth of five to eight inches (13 to 20 centimetres) can be caused by tillage. As such zones interfere with storage of moisture, they can be controlled by growing deep-rooted alfalfa at intervals, or the compacted zone can be broken by fall tillage with chisels or sweeps set to a depth just below the zone of compaction. Such deep tillage will result in reduced runoff and deeper moisture penetration.

When using power machinery in dryland farming, the timing of operations is important. The soil is broken in the fall or early spring before weeds or volunteer grain can deplete the moisture. Use of a rod weeder or similar equipment during fallow can control the weeds. Planting is timed to occur during the short period in fall or spring when temperature and moisture are favourable.

Fertilizer use. Fertilizer is an important component of dryland technology. For example, 20 pounds per acre (22 kilograms per hectare) of nitrogen are recommended where rainfall is less than 13 inches (33 centimetres), ranging up to 60 pounds per acre (67 kilograms per hectare) where more rain is available; those figures refer to the production of wheat, but they are applicable to other dryland-farming areas. Where average annual precipitation is less than 12 inches (30 centimetres), the use of nitrogen is limited to years where moisture outlook is exceptionally favourable. Nitrogen fertilizer can be applied either in fall or spring. Band placement or broadcast techniques are utilized. Good results are obtained from broadcast spring application of nitrate fertilizer, and fall application of ammonia has also been successful. Local climates and rainfall patterns also determine choice of fertilizer and time of application.

Crops and planting methods. Alfalfa grown for seed on drylands is planted in rows, usually two to three feet (60 to 90 centimetres) apart; cultivation between rows is required during the first year. Alfalfa is also grown for forage where favourable. This practice builds nitrogen and organic matter, while improving soil structure. These legumes can be rotated with wheat if rain is between 16 and 18 inches (40 and 45 centimetres) and will increase the yield of wheat.

Other crops, such as cotton, peanuts, and grain sorghum, can be grown successfully in dryland agriculture. Where those crops are produced on sandy soils, special techniques are necessary to reduce soil blowing and drifting. Cotton and peanuts do not produce sufficient crop residues for protection from wind erosion, while sorghum does. For this reason, many farmers in such areas use various row combinations of cotton or peanuts with grain sorghum; two rows of grain sorghum and four to eight rows of peanuts in alternating strips is a popular technique. Another is to use a two-year rotation of cotton and grain sorghum, in which two rows are cropped and two rows are fallow. These systems not only afford protection from wind erosion but also promote effective use of soil moisture.

TROPICAL FARMING

The area of the world bounded roughly on the north by the Tropic of Cancer and on the south by the Tropic of Capricorn, a vast land that embraces large parts of Latin America, Africa, India, Australia, and Southeast Asia, contains climates less favourable to agriculture and human settlement than those of the temperate zones. Within this Equator-centred area occur the climates known as tropical, which are characterized by two general types: warm and wet, and warm with partly deficient rainfall. In either, the total precipitation is usually quite heavy, which leaches the tropical soils of nutrients. The area also has high temperatures with little variation the year round. The combination of high temperature and high rainfall causes organic matter to decompose quickly, leaving the soil deficient in humus. Vegetation flourishes in the tropics, along with weeds, insects, and disease organisms. Important climatic variations occur, depending upon land elevation.

Characteristics of tropical areas

Tropical crops include coconut, palm oil, rice, sugar, pineapple, sisal, cocoa, tea, coffee, jute, rubber, pepper, banana, and many others. In certain highland tropical areas, however, the crops common to temperate-climate agriculture can also be grown. The amount of tropical land well-suited to agriculture, however, is limited.

Plant-pest problems. The abundance of plant pests in the tropics, including weeds and disease, makes agriculture successful mainly in the plantation system, where needed control measures can be financed. The alternative is to move from deteriorated land to newer fields; this practice of shifting agriculture has also been common, because tropical soils lose their productive capacity so rapidly. The practice probably cannot be continued indefinitely, however, because of increasing population pressure.

The largest quantities of commercial tropical products originate in plantations, where skilled management is combined with sufficient capital to provide mechanized equipment. This is particularly true in the production of coffee, cocoa, rubber, coconut, banana, pineapple, sugarcane, and others. Much rice is produced in the Asian tropics and Indonesia, however, on small farms with intensive hand labour and simple tools, where the prime mover is likely to be the ox or the water buffalo, not the tractor.

Water management. Drainage, irrigation, and other special techniques of water management are important in tropical agriculture. An example is the cultivation of rice and sugarcane in the fertile coastal areas of Guyana. Originally through private enterprise and later by government efforts, large coastal areas were "empoldered" (diked) to keep back the sea in front and floods from the rivers in the rear. With a mean annual rainfall of 90 inches (2.3 metres), drainage is a critical factor; in fact, the system cannot discharge all possible floodwater, and so the crops must tolerate occasional drowning. With gravity drainage effective only at low tide, the drainage gates are opened on the ebbing tide and closed on the rising tide. Great difficulty is encountered in keeping the outlets unclogged by the heavy sediment discharge. Since rain does not always fall when it is needed, many fields are irrigated. Most of the rice soil is specially tilled after plowing in order to create a better seedbed under the water, using tractors operating in water four to six inches (ten to 15 centimetres) deep. After this special tillage, the seeds are broadcast in one to two inches (2.5

Guyana's system of coastal water control

to five centimetres) of water. Though maintenance and operation of such an intricate water-control system are not simple, Guyana rice production has been doubled through its use.

Mechanical problems. Mechanization faces many obstacles before wide adoption is possible in tropical regions. Difficult soils, stones, stumps, abundant labour, resistance from farmers, lack of incentives, lack of skills, lack of capital, low wages, high cost of machines, lack of dealer service, fragmented land ownership, all contribute to slow development of mechanization. Tropical soils differ markedly from those in the countries that manufacture land-preparation machinery, making adaptation of new design necessary. The encountering of stones, wood, trash, and termite mounds causes machines to break down. Depressing climatic conditions reduce the performance of the machine operators. Tropical farm regions are notoriously irregular or mountainous, impeding intensive machine culture. The best soils in Brazil require special erosion controls, reducing the potential for large-scale mechanization. One of the greatest overall impediments to mechanization is the fear that unemployment might result from it, a failure to understand that economic development and higher living standards depend partly on increasing the productivity of labour.

Mechanization in Trinidad

As an example of the problems encountered in mechanizing tropical crops, the harvesting experience of a large sugarcane plantation in Trinidad is illuminating. On flatland of some 30,000 acres (12,000 hectares), the cane is grown on heavy clay soil in a climate with 50 inches (1.3 metres) of rain during the seven-month wet season and ten inches during the five-month dry season. By 1960 the rising wage rate made harvest mechanization imperative. First, the traditional "bed" system, which functioned to remove floodwater, was changed to ridge planting; this made it possible for machines to operate and was a remarkable change in itself. Then it was decided to harvest with the cane combine, which tops, cuts, chops, and loads the chopped cane into transport vehicles. Although the combine is complicated and requires considerable power, it was deemed better than mechanical half-measures.

By 1969 the combines, however, were harvesting only 12.8 percent of the flatland crop, indicating that mechanization was far from complete. Three factors were responsible: first, the cane combines required extensive maintenance plus very expensive replacement parts. Second, it was difficult to mobilize a transport system to receive the output of the combines with any degree of economy. Third, the social problem of displaced workers had to be considered. The combines increased labour productivity sixfold over hand harvesting; thus, their introduction had to be slowed until surplus workers could be accommodated elsewhere. As of 1970, efforts to mechanize cane harvest over a ten-year period had achieved only limited success, indicating how complicated such a process really is.

Taking the largest view of possibilities for improving tropical agriculture, the most promising inputs of technology are improved crop varieties and increased use of fertilizers.

OTHER SPECIALIZED TECHNIQUES

A number of specialized agricultural techniques are discussed in the article HORTICULTURE. Two of these, hydroponics and greenhouses, are touched on here.

Hydroponics. The term hydroponics denotes soilless culture of plants. The possibilities of this technique have received considerable attention in recent years. In hydroponics, an outgrowth of laboratory techniques long used by scientists, plants are grown with their roots immersed in a water solution containing necessary minerals or rooted in a sand medium kept moistened by such a solution. Soilless culture of plants is similar in principle but larger in scale. A typical hydroponics technique has plants supported in a bed of peat, wood fibre, or similar material, on a wire screen with the roots dipping into the solution below. Aeration of the solution is provided. In another method, the plants are rooted in a medium of sand or gravel contained in a shallow tank into which the solution is pumped at intervals by automatic control. Between pumpings, the solution drains slowly down into a reservoir tank. Hydroponic techniques are practiced on a small scale both out-of-doors and in greenhouses.

Of the elements known to be necessary for plant growth, carbon, oxygen, and hydrogen are obtained by the plant from atmospheric gases or from soil water. The others are all obtained as mineral salts from the soil. The elements absorbed as salts—iron, manganese, boron, copper, zinc, and molybdenum—are required in minute quantities and are called the micronutrients. The principal elements that must be provided as dissolved salts in hydroponic techniques are nitrogen, phosphorus, sulfur, potassium, calcium, and magnesium. Numerous solutions have been devised to fulfill these requirements.

Crop yields of some plants can be obtained fully equal to those obtained on fertile soils. Wide-scale crop production by hydroponics, however, would be economic only for certain intensive types of agriculture or under special conditions. Some greenhouse crops, both vegetables and flowers, are grown by this method. In regions having no soil or extremely infertile soil but with favourable climate, hydroponic techniques have been very useful; for example, on some of the coral islands of the Pacific.

Greenhouses. The greenhouse is typically a structure whose roof and sides are transparent or translucent, permitting a sufficient quality and quantity of solar radiation to enter the structure for photosynthesis (see below). It allows the growing of crops independently of the outside climate, since its interior temperature and humidity can be controlled. Greenhouses vary in size and complexity from small home or hobby structures to large commercial units covering an acre or more of land. An even smaller greenhouse might be termed the hot bed, a glass-topped box containing fermenting organic matter; the fermentation process yields heat, allowing the gardener to start plants from seed in early spring for later transplanting.

Basic greenhouse construction

The basic construction of a greenhouse consists of a light but sturdy frame capable of resisting winds and other loads. Conventional foundations usually support vertical walls; the roof may be gabled, trussed, or arched. The conventional greenhouse is fitted with glass panes, but plastic-film or fibre-glass panels often supplant glass.

Maintenance of temperature within the greenhouse is difficult because of fluctuating outside conditions. When the sun shines brightly, little heat is needed, and the heating system must be controlled in some way to prevent injury to the crop. Hot water, steam, electric cable, or warm-air furnaces provide the heat, which is usually controlled by thermostat. Temperatures in greenhouses are regulated to suit the crop. Typical ranges are from 40° F (4° C) for lettuce, violets, carnations, and sweet peas to 70° F (21° C) for cucumbers, tomatoes, and orchids.

Cooling is often required during summer days in warm climates. Ventilation is the simplest technique, reducing inside temperature to near that of the outdoors. Additional cooling by refrigeration may be required; in dry regions, the evaporative cooler is efficient and also increases the relative humidity within the structure. Another form of environmental control consists of adding extra carbon dioxide to the air if the crop requires it for extra photosynthetic efficiency.

The commercial-greenhouse operator usually grows vegetables or ornamental plants. Such production makes more demands on the grower, because he must assume many of the tasks normally handled by nature in the open fields. He must regulate the temperature, ventilate, adjust the amount of entering sunlight, provide soil moisture, fertilize, and even facilitate pollination. During the off-season, the structure must be cleaned and fumigated, its soil restructured, and mechanical equipment checked. Mechanization of greenhouse operations has lagged far behind the pattern of agriculture in general. Disease is a particularly serious hazard in greenhouse farming, requiring constant attention and use of chemicals.

III. Harvesting and crop processing

HARVESTING MACHINERY

Reaping and threshing processes

Harvesting machinery is generally classified by crop: reapers for cutting cereal grains and threshers for separating the seed from the plant. The more modern combine cuts, threshes, and cleans the grain in one operation. Corn (maize) harvesting is performed by mechanical corn pickers that snap the ears from the stalk so that only the grain and cobs are harvested. Corn shelling may be done mechanically in the field, after or with picking. Stripper-type cotton harvesters, which strip the entire plant of both open and unopened bolls, work best late in the season after frost has killed the green vegetative growth. Hay and forage machines include mowers, crushers, windrowers, field choppers, balers, and some machines that press the hay into wafers or pellets.

Grass, legumes, corn, and other crops are often put into silos to keep them in a succulent and fermented state rather than stored dry as hay. To make silage, the crops must be cut up to permit tight packing in the silo, producing anaerobic fermentation and preventing formation of mold. Almost all silage crops are cut in the field with a forage harvester that cuts and chops the crop immediately or picks up and chops a windrow that has been cut and raked earlier.

Root crops such as potatoes are harvested with diggers and digger–pickers, which often pull up clods, stones, and vines with the crop. Though some machines carry men who manually sort out such extraneous material, this task is increasingly being performed mechanically. Modern sugar-beet harvesters lift the whole root from the ground, clean the earth from it, and deliver it to a bin or wagon. Sometimes the beet tops are removed before harvest of the roots and used for cattle feed. Peanuts are lifted, vines and all, and allowed to dry before removal of the pods.

Three types of tobacco harvest methods

Tobacco-harvesting aids may be classified in three principal ways, according to the harvesting and curing methods used, which depend on the type of tobacco and its use. Flue-cured tobacco, a large plant that may stand three to four feet (90–120 centimetres) high, is harvested with machines that carry several workers who ride the lower platforms of the machines, cut the leaves, and place them on conveyor belts, where the leaves are tied mechanically or by hand. Burley tobacco has usually been harvested by workers using a machete-type knife. After cutting, the large end of the stalk is fixed onto the sharpened end of a stick, which—when loaded with a number of stalks—is hung by hand in a tobacco barn for curing. Researchers are attempting to mechanize the cutting, impaling, and hanging of burley tobacco. Little has been done, however, toward the mechanization of the harvesting of the small aromatic tobacco leaves, which are grown in the shade, picked by hand, tied with a string, then hung for curing.

Tree-crop harvesting, accomplished by hand or with mechanical shakers, is described under FRUITS AND FRUIT FARMING. Vegetable crops such as asparagus, celery, lettuce, and cabbage are still harvested largely by hand, though scarcity and high cost of field labour has led to some mechanization in this area, notably with tomatoes.

CROP-PROCESSING MACHINERY

Machinery is widely used to prepare crops for convenient transportation, for safe storage, for the market, and for feeding to livestock. Advances in such machines have been rapid, particularly with new crops, increased yields, multiple-crop practices, and changing techniques.

In the most common method of crop drying the crop, usually grain, is spread on floors or mats and stirred frequently while exposed to the sun. Such systems, though extremely common in the underdeveloped countries, are very slow and dependent on the weather. Forced-air-drying systems allow the farmer much more freedom in choosing grain varieties and harvest time. Fairly simple in operation, these systems have been gaining popularity in the tropics. Heat is often added to increase air temperatures during the drying period.

Dryeration of corn

In dryeration, wet corn is placed in a batch or continuous dryer. After losing 10 to 12 percent moisture, the hot corn is transferred to the dryeration cooling bin, in which it is tempered for six to ten hours and then slowly cooled by ventilation for ten hours. This process reduces kernel damage and increases dryer output.

High moisture in stored hay not only causes rapid deterioration of its value as feed but often results in spontaneous combustion. When hay is first cut, it usually contains 70 percent or more moisture. It wilts and quickly dries to a moisture content of about 40 percent. At this stage, it can be dried to a safe storage condition, about 15 percent moisture, by blowing air through it, sometimes with supplemental heat.

Feed-processing mills, often referred to as feed grinders, are used principally for milling cereals for livestock feed, which aids digestion. The ground material is usually fairly coarse and at times may only be crushed. Modern mills frequently are designed to allow the farmer to grind the grain and to mix in various other ingredients in desired quantities.

Other types of crop-processing machinery include machines that separate desirable seed from weed seed, stems and leaves, and dirt; grading machinery to classify seed by width, length, or thickness; fruit graders and separators; and cotton gins, which separate cotton seeds from the fibres.

IV. Weather and agricultural technology

WEATHER INFORMATION

The interaction of weather and living systems is a basic aspect of agriculture. Although great strides in technology have resulted in massive production increases and improved quality, weather remains an important limiting factor. Though man is not yet able to change the weather, except on a very small scale, he is capable of adjusting agricultural practices to fit the climate. Thus, weather information is of utmost importance when combined with other factors, such as knowledge of crop or livestock response to weather factors; the farmer's capability to act on alternative decisions based on available weather information; existence of two-way communication by which specific weather forecasts and allied information can be requested and distributed; and the climatic probability of occurrence of influential weather elements and the ability of the meteorologist to predict their occurrence.

Other weather-research benefits. Apart from the many applications of weather forecasting to current problems, meteorological research may benefit agriculture in at least three other ways: (1) improved planning of wide-scale land usage depends partly on detailed knowledge of plant-climate interactions; radiation, evapotranspiration, diurnal temperature range, water balance, and other parameters are measured and analyzed before a plan realizing maximum economic benefit for a given area is prepared; (2) agronomic experiments are combined with climatological documentation to obtain the greatest scientific and technological return; (3) problems of irrigation, row spacing, timing of fertilizer application, variety selection, and transplanting can best be solved with the aid of climatic environmental data; cultural practices related to artificial modification of microclimates should be based on research knowledge rather than personal judgment.

Three classes of weather stations

Observing climatic elements. The climatic elements the observation of which is valuable for agricultural purposes can be approached on an idealized threefold scale: (1) microscale observations of small areas for research designed to elucidate basic physical processes (see also MICROCLIMATES); (2) mesoscale climatic networks designed for practicing farmers to improve their operations; and (3) macroscale regional networks intended for weather forecasting and for gathering basic climatic data (see also WEATHER FORECASTING; METEOROLOGICAL MEASUREMENT). Macroscale stations can be further divided into first-order and second-order stations, the number and type of observations different for each. Micrometeorology demands the most elaborate array of measuring devices, while a second-order macroscale station requires

the least; in fact, the latter station will measure only five elements: air temperature, rain, snow, humidity, and surface wind. A first-order macrostation will be equipped to measure 16 elements: global radiation, sunshine hours, clouds, net radiation, air temperature, soil temperature, rain, snow, hail, dew, fog, humidity, pan evaporation, pressure, upper air wind, and surface wind. Mesoscale measurements include ten elements and microscale 27 (three of which are derived from others).

The World Meteorological Organization and the various national weather services are concerned with establishment and improvement of macroscale regional climatic stations, both first-class and second-class. Spaced at least ten miles (16 kilometres) apart, their value for daily agricultural operations is limited, but they are useful for long-range planning and forecasting. Most parts of North America, Europe, and Australia have adequate networks of these stations, but wide gaps exist in the tropics, polar regions, and arid lands.

The degree day. One weather characteristic of agricultural value is the degree day. This concept holds that the growth of a plant is dependent on the total amount of heat to which it is subjected during its lifetime, accumulated as degree days. Common practice is to use 50° F (10° C) as a base. Thus, if the mean daily temperature for a particular day is 60° F (16° C), then ten degree days are accumulated for that day on the Fahrenheit scale. The total number of growing degree days required for maturity varies with crop variety as well as plant species. Also, the minimum threshold temperature (below which the plant is damaged or unable to grow) varies with plants; *e.g.*, 40° F (4° C) for peas, 50° F (10° C) for corn, and 55° F (13° C) for citrus fruits. Where studies have established the number of degree days required for maturity of a given crop, the planting dates can be scheduled for orderly harvest and processing. The system is helpful in selecting crop varieties appropriate to different areas; it also has value in scheduling spray programs and predicting insect emergence.

Weaknesses of the degree-day concept

The growing-degree-day concept has certain weaknesses: (1) it assumes that the relationship between growth and temperature is linear (actually it is not); (2) it makes no allowance for changing threshold temperatures with advancing crop development; (3) too much weight is given to temperatures above 80° F (27° C), which may be detrimental; and (4) no account is taken of the diurnal temperature range, which is often more significant than the mean daily value.

WEATHER EFFECTS

The essence of the weather–agriculture interaction for the farmer lies in wise adaptation of operations to the local climate and in techniques for manipulating or modifying the local environment (microclimate) to minimize weather stresses on plants and animals. Many of these techniques have been practiced for centuries: seeding and cultivation, irrigation, frost protection, animal shelters, windbreaks, and others are methods of altering the microclimate. The climatic factors and their relation to plant growth in terms of protective techniques are important.

Solar radiation. Solar radiation is the ultimate source for all physical and biological processes of the earth. Agriculture itself is a strategy for exploitation of solar energy, made possible by water and nutrients. During daytime hours, solar radiation is delivered both directly and by diffused sky reflection. The incoming radiation that is not reflected by the surface or reradiated to outer space is the net radiation, which is the energy available for maintaining the earth's surface temperature. At night the net radiation is negative; that is, energy is lost to outer space by long-wave radiation, and none is gained. The net radiation balance varies widely throughout the world, setting limits on basic agricultural possibilities.

Photosynthesis. Photosynthesis is the process by which higher plants manufacture dry matter through the aid of chlorophyll pigment, which uses solar energy to produce carbohydrates out of water and carbon dioxide. The overall efficiency of this critical process is somewhat low, and its mechanics are extremely complex. It is related to light intensity, wavelength, temperature, carbon dioxide concentration in the air, and the respiration rate of the plant. The distribution of solar energy within the plant community is affected by the leaf canopy's density, height, and capacity to transmit the energy; these therefore affect photosynthesis. The leaf-foliage density is characterized by the leaf-area index, the total leaf area of a plant over a given area of land. The optimum leaf-area index will vary between summer and winter and between temperate and tropical regions, but it represents a key factor in the search for better crop management based on improved photosynthesis. The efficiency of radiation utilization by field crops has been measured, showing that an ordinary crop converts less than 1 percent of available solar energy into organic matter.

Leaf canopy's effect on solar-energy distribution

Photoperiodism. Photoperiodism is another attribute of plants that may be changed or manipulated in the microclimate. The length of a day is a photoperiod, and the responses of the plant development to a photoperiod are called photoperiodism. Response to the photoperiod is different for different plants; long-day plants flower only under day lengths longer than 14 hours; in short-day plants, flowering is induced by photoperiods of less than ten hours; day-neutral plants form buds under any period of illumination. There are exceptions and variations in photoperiodic response; also, it is argued that the truly critical factor is actually the amount of exposure to darkness rather than to daylight. Temperature is intimately related to photoperiodism, tending to modify reactions to daylength. Photoperiodism is one determining factor in natural distribution of plants throughout the world.

The phenomenon has many practical applications. Selection of a plant or a variety for a given locality requires knowledge of its interaction with the photoclimate. Artificial illumination is used to control flowering seasons and to increase production of greenhouse crops. In plant breeding, such stimulation of flowering has greatly reduced the time span from germination to maturity, shortening the time necessary to develop new varieties. In sowing field crops, photoperiodism can be used to select the date of sowing to produce optimum harvest size. Crop yield is reduced both by planting in a season that will cause plants to flower early and by planting at a time that will cause very late flowering. In Sri Lanka (formerly Ceylon), certain rice varieties with a vegetative period of five to six months may extend their life to more than a year when planted in the wrong season, causing almost complete loss of yield. Cowpeas in Nigeria will flower early and produce many seeds only when planted in daylengths of 12 hours or less.

WEATHER CONDITIONS AND CONTROLS

Temperature. Regardless of how favourable light and moisture conditions may be, plant growth ceases when the air and leaf temperature drops below a certain minimum or exceeds a certain maximum value. Between these limits, there is an optimum temperature at which growth proceeds with greatest rapidity. These three temperature points are the cardinal temperatures for a given plant; the cardinal temperatures are known for most plant species, at least approximately. Cool-season crops (oats, rye, wheat, and barley) have low cardinal temperatures: minimum 32° to 41° F (0° to 5° C), optimum 77° to 88° F (25° to 31° C), and maximum 88° to 99° F (31° to 37° C). For hot-season crops, such as melons and sorghum, the span of cardinal temperatures is much higher. The cardinal temperatures may vary with stage of development. For example, cold treatment near 32° F (0° C) of germinated seeds before sowing can transform winter rye into the spring type; such treatment, called vernalization, has practical application in cold-climate plants.

Three cardinal temperatures

The range of diurnal temperature variation is also important; the best net photosynthesis is related to a large diurnal temperature range, or high daytime and low nighttime temperatures. Knowledge of the difference between leaf and air temperatures aids farmers in adopting

protective measures. In middle and high latitudes, frost often occurs before the air temperature drops to freezing; in summer, heat injury to plants might be much more serious than that suggested by the air temperature alone. Because of this factor, farmers in Taiwan shade the pineapple fruit to prevent heat damage.

Soil temperature sometimes is of greater ecological significance to plant life than air temperature. Germination of seed, root function, rate of plant growth, and occurrence and severity of plant diseases all are affected by soil temperature. Since an unfavourable soil temperature during the growing season can retard or ruin a crop, techniques have been developed for modifying the temperature. The two most important methods are (1) regulation of the energy exchange and (2) altering the thermal properties of the ground. Incoming energy can be regulated by an insulation layer on or near the ground surface, such as paper, straw, plastic, or trees; the outgoing radiation can be reduced by insulation materials or by generating smoke or fog in the air. Thermal properties of the ground can be modified by cultivation or irrigation, increasing the soils ability to absorb radiation, or by varying the rate of evaporation. Mulching is a common technique for soil temperature control. Carbon black or white material can change the soil's ability to absorb radiation. In the Soviet Union, for example, it was reported that 100 pounds of coal dust per acre (112 kilograms per hectare) caused a one-month advance in the maturity date of cotton.

Frost. Another aspect of temperature control is frost protection. Likelihood of damage from freezing temperature depends upon the plant species, the season, the manner of temperature change, the physiological state of the plant, and other factors. Orchards can be located so as to minimize the chances of frost damage.

Radiation and wind types of frost

Two types of frost are recognized: (1) radiation frost, which occurs on clear nights with little or no wind when the outgoing radiation is excessive and the air temperature is not necessarily at the freezing point, and (2) wind, or advection, frost, which occurs at any time, day or night, regardless of cloud cover, when wind moves air in from cold regions. Both types may occur simultaneously. Most frost-protection techniques can raise the temperature only a few degrees, while some are effective only against radiation frost.

Heating is probably the best known and most effective frost-protection measure. It is most effective on nights with a strong temperature inversion, a condition in which the air temperature increases markedly from the ground up to as high as 40 or 50 feet (12 or 15 metres). The depth of air to be heated is thus rather shallow, and the area over which a given temperature rise can be produced increases linearly with the strength of the inversion. Lacking a temperature inversion, heaters protect by radiating heat to the plants and the ground surface, and by emitting a layer of humid smoke that reduces the net outgoing loss from the ground.

In general, a large number of small heaters is most effective; large heaters set up convection currents that break up the warm ceiling and draw in cold air. For radiation-frost protection, the heaters are placed in "view" of the plants or trees, but for advective frost the heavier concentration is placed along the upwind border. Common fuels for the heaters include oil, coal, briquettes, and wood. Oil is most effective, because it can be ignited rapidly and extinguished easily. Heating is a costly technique; a few growers who tried it in England soon gave up the practice, and, even in places such as California, heating is becoming less common and is mostly restricted to a few high-value crops such as citrus fruits.

Wind machines for frost protection

The wind machine is popular for frost protection; although it affords less reliable results, its operating cost is only about 20 percent of that for heaters. These machines, which are like fans or propellers, break up the nocturnal temperature inversion by mechanically mixing the air, returning heat to the ground that was lifted during the day. The stronger the temperature inversion, the more effective is the wind machine, which is ineffective, however, against a daytime freeze or cold soil. Even under the best circumstances, ground-surface temperatures will rise very little; therefore, some operators install both heaters and wind machines, using the latter for strong-inversion nights and the former for wind-frost protection.

Flooding and sprinkling with water prevent excessive ground cooling by increasing the heat conductivity and heat capacity of the soil and releasing latent heat of fusion, or the heat given off when the water freezes. The temperature of the plant will not fall below the freezing point so long as the change of state from water to ice is taking place. Flooding has the disadvantage of retarding increase in soil warmth during the day; thus, it can be used effectively for only one or two nights. Sprinkling creates water particles in the air that reduce outgoing radiation, but plant temperature declines immediately on cessation of sprinkling, and the ice formation may cause damage to the crop. In general, successful protection by flooding and sprinkling demands much skill and judgment from the operator.

Brushing is a frost-protection technique in which shields of paper or aluminum foil are set up to reduce radiation loss to the sky; it has been used with fair success for tomato culture in California.

Massachusetts cranberry growers add a thin layer of sand to the soil periodically. The sandy surface warms up easily and cools slowly by radiation; it also reduces evaporation of its low water content. Sanding can raise the temperature of loam, clay, and organic soils, thus diminishing frost hazard. Windbreaks can also function as frost protection by reducing inflow of cold air and by shielding plants from the total night sky.

Spraying of harmless foams or gels on plants threatened by frost is a technique under investigation. The trapped air in the foam serves as insulative protection, while the foam can be designed to dissolve after any desired time interval. The technique has been explored for use on strawberries and other low-growing crops.

Irrigation. Irrigation is probably the most common form of agricultural microclimatic control practiced by man. The basic techniques of irrigation are discussed in the article IRRIGATION AND DRAINAGE; at this point it only remains to make mention of efforts to correct deficiencies in precipitation, the deficiencies that lead farmers to irrigate.

Rainmaking. Attempts to increase the amount of precipitation from clouds by seeding them with salt or silver iodide have been made for nearly three decades. Both aircraft and ground generators have been employed, but the techniques are typically beyond the means of an individual farmer. Results suggest that cloud modification is entirely possible, but the proof of increased rainfall at a level of statistical significance is a difficult problem. Success has been greatest under atmospheric conditions where natural rainfall is most probable. The prospect of modifying winter clouds to increase snowfall in mountain areas appears to be somewhat more promising, however.

Cloud modification versus rainfall

Most cloud-seeding efforts are expended in regions where precipitation is only marginal for agriculture. It is commonly assumed that at least 20 inches (50 centimetres) of rain per year, fairly well distributed, is required to maintain a stable farming community. Unfortunately, the years of large deficiencies in such areas are those with only limited opportunity for cloud seeding. Some observers believe that weather modification to increase precipitation may yet become practical and economically feasible; the legal, ethical, and ecological problems raised by the prospect will not be easily solved, however. The outlook for cloud seeding in the early 1970s appeared favourable for lightning, hail, and fog suppression or control.

Humidity. The value of high humidity in the greenhouse is well known, but knowledge of humidity–plant interaction under field conditions is comparatively slight. Other things being equal, the evapotranspiration rate decreases with increasing humidity; thus, rate of water use is higher at low levels of humidity. The benefits of irrigation are apparently greater when the humidity is high,

which simply means that the efficiency of water use increases with humidity.

Wind. Wind affects plant growth in at least three significant ways: transpiration, carbon dioxide intake, and mechanical breakage. Transpiration increases with wind speed, but the effect varies greatly among plant species; also, the effect is related to temperature and humidity of the air. In arid climates, dry and hot winds often cause rapid, harmful wilting. In winter, with frozen soil, the damaging effect of increased transpiration resulting from wind can be serious because the lost water cannot be readily replaced. By contrast, increasing wind promotes carbon dioxide intake within limits; this benefits the rate of photosynthesis. The effects of mechanical wind damage vary with species; some show a definite decrease in dry matter production with increasing wind, while others (usually short plants) are unaffected. Because of the long-recognized need, shelterbelts, massive plantings of trees that change the energy and moisture balance of the crop, are positioned to protect crops and to increase yields. A shelterbelt perpendicular to the prevailing wind reduces velocity on both sides. A medium-thick shelterbelt can reduce wind velocity by more than 10 percent to a distance of 20 times the tree height on the leeward side and three times the tree height to the windward. The shelterbelt length should be at least equal to that of the field protected. The sheltered area will suffer much less soil erosion and mechanical damage than unprotected areas. Other microclimatic effects of shelterbelts include: (1) small daytime temperature increases and nighttime decreases; (2) the occurrence of radiation frost in the leeside may be promoted; (3) rate of evaporation in the sheltered area is decreased, depending on wind velocity; (4) snow accumulates near the shelterbelt, causing increased moisture storage in dry farming.

Shelterbelts for crop protection

The overall effect of a shelterbelt is complicated but probably beneficial. There is much evidence that they increase efficiency of water use not only in subhumid and semi-arid regions but also in true deserts where oasis-type irrigation is practiced. The response to shelterbelts, however, depends on the species. Crops of low response to wind protection are the drought-hardy small grains and maize grown under dry farming conditions. Rice and forage crops such as alfalfa, lupine, and clover are moderately responsive. Crops that benefit most from wind protection are garden crops, such as lentils, potatoes, tomatoes, cucumbers, beets, strawberries, watermelons, deciduous and citrus fruits, and other tender crops, such as tobacco and tea. Some authorities assert that in strong wind areas shelterbelts will produce an average 20 percent yield increase, which is net gain of 15 percent when allowance is made for the land occupied by the belts themselves. Trees can be grown almost anywhere, even in the desert; tall plants such as corn, sorghums, or even elephant grass can also be employed in arid regions by including them in the irrigation schedule. It would appear that windbreaks are among the most practical means of beneficial weather modification in agriculture.

V. Pollution and agricultural technology

Practically all forms of technology exact a certain price in environmental damage; agriculture is no exception. Likewise, agriculture in turn is sometimes damaged by undesirable by-products of other technologies (see also POLLUTION CONTROL).

Air has physical properties and a chemical composition that are vital parameters of life for both plants and animals. Temperature, water vapour, movement, oxygen, and carbon dioxide in the atmosphere have a direct effect on food and fibre production. Air quality is changed by introduction of contaminants into it, and agricultural activities using such air may be affected adversely. Damage to plants by air pollutants is related to meteorological conditions, particularly temperature inversions in the atmosphere.

AIR POLLUTION

Air pollution damage to agriculture. For more than a century air pollution has affected agriculture. Burning coal and petroleum produce sulfur oxides. Fluorides result from smelting and glass and ceramic manufacture. Rising levels of ammonia, chlorine, ethylene, mercaptans, carbon monoxide, and nitrogen oxides are found in the air. Motor vehicles and growing population produce photochemical air pollution affecting not only the urban concentrations but also the contiguous rural areas. The mixture of pollutants from all sources, including agriculture, has released a host of contaminants into the air, such as aldehydes, hydrocarbons, organic acids, ozone, peroxyacetal nitrates, pesticides, and radionuclides. The effect of these pollutants on food, fibre, forage, and forest crops is variable, depending on concentration, geography, and weather conditions. Damage to crops by air pollution, of course, brings economic loss as well.

Types of air contaminants

The effects of air pollution on plants and animals may be measured by the following factors: (1) interference with enzyme systems; (2) change in cellular chemical constituents and physical structure; (3) retardation of growth and reduced production because of metabolic changes; (4) acute, immediate tissue degeneration. Pollutants that enter the air from sources other than agriculture and that produce plant response are classified as: (1) acid gases; (2) products of combustion; (3) products of reactions in the air; and (4) miscellaneous effluents.

Acid gases. Acid gases include fluorides, sulfur dioxide, and chlorine. Hydrogen fluoride is extremely toxic to plants; some plants are injured by contact with concentrations of less than one part per billion. The damage apparently occurs initially to the chlorophyll, producing a mottled chlorosis and later killing the cells. Plants vary in degree of tolerance to hydrogen fluoride; usually the plants that accumulate fluoride readily are the most tolerant. Corn is more susceptible than tomato. All plants are most susceptible to fluoride injury during periods of rapid growth.

Sulfur dioxide given off in combustion of oil and coal commonly causes necrosis (cell death) of the leaf (Figure 4). At certain concentrations, sulfur dioxide will affect

By courtesy of the U.S. Department of Agriculture

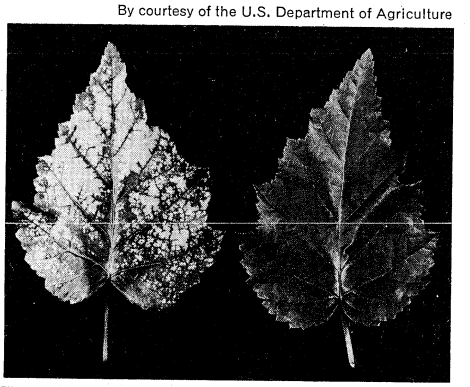


Figure 4: *Air pollution injury to vegetation.* (Left) The effect of sulfur dioxide on a white birch leaf. (Right) A normal leaf.

plants if the stomata (minute pores in the epidermis of a leaf or stem) are open. High light intensity, favourable growth temperatures, high relative humidity, and adequate water supply are conducive to open stomata. Plants that close their stomata at night can tolerate sulfur dioxide much better during that period. Conifers are more susceptible in spring and early summer, when the new needles are elongating. The sulfur dioxide absorbed by the leaf cells unites with water to form a toxic sulfite, but this is slowly oxidized to a relatively harmless sulfate. The toxicity of sulfur dioxide thus is a function of the rate at which it is absorbed; rapid absorption will cause greater injury. Chlorine damage to plants is somewhat rare; its typical symptoms are bleaching and necrosis.

Products of combustion. The primary products of combustion are ethylene, acetylene, propylene, and carbon monoxide. Of these, ethylene is known to affect plants adversely; while the others may also do so, it would require higher concentrations of them than typically occur in polluted air. For many years it was ob-

Effects of ethylene

served that illuminating gas (3 percent ethylene) leaking from pipelines caused damage to nearby vegetation. Now, with the use of natural gas, ethylene in the air is derived mostly from certain chemical industries and from automobile exhaust. Greenhouse flowers in metropolitan areas are typically damaged by ethylene. Such injury appears to be caused by excessive speeding up of the life process, thus bringing on damage. Ethylene was first identified as affecting plant life over large areas in the field by its effects on cotton and other plants near a polyethylene factory.

Ethylene, ozone, and peroxyacetyl nitrate are produced as reaction products in the air and are clearly implicated in plant injury. In addition, certain bisulfites and nitrogen dioxide are under suspicion; there are probably others. Ozone is a major air pollutant affecting agriculture. Damage has been identified in a number of field crops, including spinach, tobacco, fruits, vegetables, forest trees, and ornamentals. Symptoms of ozone toxicity appear as flecks, stipple, streaks, spots, tipburn, and premature yellowing of the foliage; these may be visible only on the upper leaf surface. Peroxyacetyl nitrate and its analogs produce symptoms called silver leaf and leaf banding, which have been observed in the Los Angeles area and elsewhere for many years.

The adverse effects of airborne radioactive contaminants on the agricultural economy at the present time are small.

Air pollution by agriculture. Contributions of agricultural technology to air pollution include pesticides, odours, smoke, dust, allergenic pollens, and trash. The widespread public concern about pesticides makes it imperative that pesticide technology be carefully controlled and that search for better methods be pursued vigorously.

Persistence in pesticides

Pesticides. The problem of persistence in pesticides can be highlighted by noting that this attribute exists in a range from moderately persistent (a lifetime of one to 18 months—2,4-D, atrazine); persistent (lifetime up to 20 years—DDT, aldrin, dieldrin, endrin, heptachlor, toxaphene); or permanent (lead, mercury, and arsenic). Presumably, the less persistent types should be more desirable, other things being equal; but those that degrade rapidly, such as the organophosphate insecticides, are extremely toxic and nonselective, which encourages rapid emergence of resistant insects and destroys their natural enemies. Thus, it is apparently not possible to adopt chemicals that function without some kind of drawback or disadvantage.

Whether pesticides are applied by spraying or by surface application, air is the usual medium through which the chemicals move to their intended and unintended targets. Reliable data on how pesticides behave in air, such as distance travelled, are lacking, because adequate monitoring is unavailable. Their chemical and physical nature, method of application, and the atmospheric conditions will influence their concentration and ultimate fate. There is no doubt that pesticides may be transported long distances on dust particles. The rate of removal from air is difficult to predict, but in the long run the chemicals return to earth.

Odours, pollen, and dust. Odours from animal concentrations are recognized as being highly undesirable to air quality. Where these operations exist contiguous to urbanized areas, public reaction is usually unfavourable. Disposal of animal waste on the land may worsen the odour problem; in addition, high wind may move dry increments into the air. Smoke is emitted by operations designed to dispose of crop residues, or by controlled burning of weeds and brush. Air quality is also affected by transmission of allergenic pollen such as ragweed pollen, which can be blown for hundreds of miles.

Improper land use and treatment can cause considerable deterioration in air quality. Practices that strip the soil of plant growth or crop residues for long periods contribute to wind erosion, particularly in dry-farming areas. Fortunately, the technology of preventing wind erosion is well understood and widely used. Trash related to agriculture is moved freely by wind and distributed in unwanted fashion. Hulls of rice and wheat and cotton-gin trash are examples of this kind of airborne nuisance.

In contrast to most other technologies, however, the agricultural variety offers a major beneficial contribution to air quality. The photosynthesis of green plants removes carbon dioxide from the air and adds oxygen to it, thus helping to maintain the life-giving balance between these gases.

SOIL AND WATER POLLUTION

Pollutants damaging to agriculture. Soil and water pollutants that may adversely affect agricultural operations include sediment, plant nutrients, inorganic salts and minerals, organic wastes, infectious agents, industrial and agricultural chemicals, and heat.

Sediments in dual pollution role

Sediment. Sediment is a resource out of place whose dual effect is to deplete the land from which it came and impair the quality of the water it enters. Aside from filling stream channels, irrigation canals, farm ponds, and irrigation reservoirs, sedimentation increases cost of water clarification. Suspended sediment impairs the dissolved-oxygen balance in water. The recreational value of farm ponds is diminished by sediment, while soil depleted farmland is reduced in value.

Plant nutrients. Nutrients of plants become resources out of place when they appear in groundwater and surface water; in fact, they become serious pollutants. Unwanted aquatic plants are nourished by plant nutrients derived from agricultural runoff, feedlots and barnyards, municipal and rural sewage, and industrial wastes. Aquatic plants clog irrigation and drainage structures, thus increasing maintenance cost and reducing capacity. Nitrates and nitrites in groundwater, which can poison human beings and livestock, result from both agricultural and industrial operations.

Inorganic salts and minerals. Inorganic salts and minerals that impair the quality of soil and water are derived from natural deposits, acid mine drainage, industrial processes, and drainage flow from irrigated areas. Salt accumulation on irrigated soils causes the most damage and loss in this category. A high proportion of sodium in irrigation water supply affects plant life adversely (see below *Salinity*). More than just a trace of boron is highly toxic; therefore, water used in municipal and industrial processes involving borax may not be usable for agriculture.

Organic wastes. Organic wastes emanating from municipal sewage, garbage, food-processing industries, pulp mills, and animal enterprises are attacked by aerobic bacteria. When this occurs in water, the oxygen content of the water is depleted or reduced to zero, at which point the anaerobic bacteria complete the process of reducing the wastes to inert material. This produces septic conditions that make the water unfit for recreational use, farmstead supply, or crop irrigation.

Diseases transmitted by water and soil

Infectious agents. Where not carried by wind, infectious agents are transmitted mainly by water and soil. Bacterial and virus diseases of crops are spread by machines that move contaminated soil. Insects are prime carriers of these diseases. Weed seeds are spread by irrigation water, as are nematodes. Animal diseases transmitted by water and soil include leptospirosis, salmonellosis, hog cholera, mastitis, foot and mouth disease, tuberculosis, brucellosis, histoplasmosis, Newcastle disease, anthrax, coccidiosis, and many others. Mosquitoes breeding in stagnant water can transmit encephalitis. Most crops and livestock in the world are susceptible to one or more highly infectious disease that may be transported by soil or water. The cost of losses from these diseases is staggering.

Chemicals. Organic chemicals in soil and water, such as detergents, insecticides, herbicides, fungicides, nematocides, rodenticides, growth regulators, and defoliants, can have adverse effects on agriculture. The application of persistent insecticides to potato lands has led to residues in sugar beets grown in the same soil the following year, for which there are no tolerances. Fish have been killed in farm ponds because of drainage of insecticide

pollutants. Use of heptachlor (no longer recommended) to control alfalfa weevil led to soil contamination and uptake by hay; dairy cows that ate the hay produced milk containing heptachlor.

Aerial and ground application of herbicides on nonagricultural lands (utility rights-of-way, roadsides, industrial sites) often cause damage to nontarget crops. Herbicide wastes may enter drainage or irrigation ditches and create trouble. The presence of chemical residues in agricultural commodities can cause serious problems ranging from confiscation to loss of public confidence. Practically all aspects of chemical usage are now regulated or restricted by government.

Heat. Introduced into water by industrial processes, heat can have a detrimental effect on fish and other creatures in the water; damage to recreational value can result. But, though heat is a water pollutant, its effect is minor with respect to agriculture.

Pollutants from agriculture. Some pollutants from agriculture have adverse effects on agriculture itself, as excesses of plant nutrients and salts from irrigation. These pollutants and others also affect the environment at large.

Eutrophication. Eutrophication occurs in a body of water when an increase of mineral and organic nutrients has reduced the dissolved oxygen, producing an environment that favours plant over animal life. The resulting algae and other water plants tend to choke other forms of life in the oxygen competition, especially where carbon and phosphorus are plentiful. Doubtless, much phosphorus in streams and lakes is delivered from agriculture, but primarily through soil erosion rather than runoff. Though the principal source of phosphorus is apparently municipal sewage-treatment plants, direct runoff from feedlots may also contain large amounts. The solution to the problem of phosphorus in surface water lies in using good soil-conservation practices and in minimizing runoff from animal concentrations and manure.

Problem of pollution from nitrates

In contrast, identification of nitrate sources in water supplies has suffered from conflicting evidence. Where nitrate is found in water, some have concluded it came from chemical fertilizers, while others have suggested it came from natural soil nitrification or nitrification of sewage effluent or animal wastes. The problem has serious aspects, because nitrate can cause serious illness in human beings. One difficulty in identifying nitrogen sources lies in the fact that it is present in soils for reasons other than fertilization; the growth of legumes, for example.

Salinity. Salinity is a major problem in irrigation agriculture. Through evapotranspiration, salts in the irrigation water become more concentrated in the drainage effluent. It is therefore claimed that water quality is seriously impaired by irrigation agriculture. Irrigation water always contains some salt, most of which is excluded by the plant roots; since the evaporated water is pure, the soil accumulates the residual salt, which is added to what the arid soils already have in abundance. This accumulation of salt must be removed if plants are to be grown at all, and it is removed by leaching with excess water.

Survival of an irrigated area will depend, therefore, on a favourable salt balance: salt leaving the area must equal or exceed that received in the water supply. The irrigation farmer is not actually "producing" a contaminant but is transferring one in a more concentrated form. Future intensified use of limited irrigation water may add to the severity of this problem. Where the return flow is readily recoverable, as from tile drains or pumped wells, it could be purified by a desalination process, returning the purified fraction to the watercourses and disposing of the concentrated-salt fraction in such a way that usable groundwater is not affected.

Agricultural processing wastes. The wastes from processing of agricultural products represent another pollution hazard. These include runoff or effluent from sawmilling, pulp manufacture, fruit and vegetable canning, cleaning of dairies, slaughtering of meat animals, tanning, manufacturing of cornstarch and soy protein, sugar refining, distilling, wool processing, and many others. The runoff from agricultural enterprises can contain disease organisms and other infectious agents. Insects associated with agriculture can transmit diseases. Plant diseases move from agriculture to lawns, gardens, parks, and golf courses.

Monitoring pesticides. Monitoring of pesticides in water has been carried on in various areas since World War II. Some monitor networks, backed by analysis laboratories, are quite extensive. The accumulated data show how and when certain pesticides move from target areas into other parts of the environment. Ponds and catch basins sometimes show measurable amounts of pesticide residues from water leaving fields. Although most organic insecticides are hydrophobic and almost insoluble in water, they can become attached to materials suspended in water; but, after the materials settle, the insecticide residues usually become negligible. This confirms the earlier supposition that the movement of chemicals from target areas is greatest when silt and organic loads are high in runoff water.

Transport pattern of pesticides

Levels of pesticides in soils are constantly changing. So many variables and processes are involved, that rates of accumulation of even the most persistent insecticides are quite variable and difficult to determine. Soil monitoring programs are underway, however, and are providing much-needed information. The problem of accumulation in soils arises because the tiny organisms in soil are not capable of degrading many pesticides at rates sufficiently high to prevent soil and also water pollution. Thus, the persistent types, such as DDT and other chlorinated hydrocarbons, remain available for absorption by higher animals (including human beings) and for causing harm to nontarget organisms.

VI. The future of agricultural technology

AGRICULTURAL EXPERIMENTATION AND RESEARCH

Research and development in both the private and public sectors have played an increasingly important role in shaping agricultural technology. Application of scientific knowledge gained by research will be the key factor in the future technology of agriculture; it is indeed remarkable how much of the knowledge is already available and waiting for widespread application and local adaptation.

The most recent scientific advances have occurred in four areas: (1) the mechanization of crop culture and livestock management, a development that has profoundly changed the varieties of crops, their water and fertilizer requirements, and the need for labour; (2) the use of chemicals for control of insects, weeds, and regulation of plant and animal growth; (3) development of new varieties of crops and animals through genetic experimentation; and (4) the new management systems, systems engineering, and operations research made possible by the computer.

Four areas of agricultural progress

These advances have brought dramatic increases in yields of both crops and animal products and have greatly expanded the capacity of individual farmers to produce food and fibre.

Crop research and development. *Insect control.* In the United States alone, it is estimated that weeds, insects, and plant disease cause an annual 20 percent reduction in crop production. The equivalent of 67,000,000 acres (27,000,000 hectares) is devoted to feeding insects, weeds, and plant pathogens. Obviously, there is a strong economic incentive for research here.

The search for control of unwanted insects has produced knowledge of the chemical structure of the insect juvenile hormone, which has been synthesized. All insects respond to extremely small topical application of the hormone (about one gram per acre) by failure to mature, or pupate, thus bringing them under control with a minimum chance to develop the resistance that has plagued the use of certain chemicals.

Soil reclamation. A technique for reclamation of semi-arid sandy soils consists in laying down an asphalt moisture barrier ⅛ inch (three millimetres) thick, two feet (60 centimetres) below the surface. The barrier im-



Figure 5: The use of plastic tunnels to protect horticultural crops in Malta.
By courtesy of the Food and Agriculture Organization, Rome

proves the water-retaining capacity of such soil, reduces the need for irrigation water, and has improved yields under experimental conditions. Asphalt paddies in sandy soils in Taiwan produced an elevenfold increase in rice production with one-eighth the usual amount of water.

Carbon dioxide enrichment. Carbon dioxide enrichment of greenhouse atmosphere has increased yields of certain food crops up to 100 percent. It seems reasonable to expect similar results in field crops, although wide-scale field trials have not been conducted. Laboratory tests show that corn, grain sorghum, soybeans, sugar beets, barley, and rice all respond favourably to carbon dioxide enrichment. With increased plant populations, better leaf configuration, and improved fertilization and irrigation, carbon dioxide may become the limiting factor in productivity of field crops, and this factor may be supplied by farmers in the future because carbonates are so plentiful.

Protected agriculture. Vast sums are being expended on protecting plants from unfavourable climate. Worldwide, the area under glass, fibre glass, and plastic is increasing at an annual rate of 10 percent, with extensive developments in Japan and Europe (Figure 5). Such protected cultivation can lead to tremendous production; for example, yields of tomatoes exceeding 120 tons per acre (270,000 kilograms per hectare) per year have been achieved under such conditions. Hydroponic culture, soil heating, and artificial lighting are made possible under protected cultivation, achieving virtual independence of climate and on soils previously unsuited to agriculture. Relief of heat stress can be provided by evaporative cooling and irrigation can be practiced very efficiently. Other forms of protected cultivation include frost protection by use of sprinkler irrigation and solid-petroleum-block fuel systems in orchards and vineyards.

Greenhouses on Mexico's desert coast

Another example of protected cultivation has been developed on a seacoast of Mexico with abundant, year-round sunlight but no freshwater. Plastic-film greenhouses were constructed at extremely low cost, and plants were grown in them in an atmosphere of about 100 percent relative humidity, enriched with carbon dioxide. The high humidity reduced plant water requirements to only 1 to 5 percent of normal, and the carbon dioxide enhanced growth rates. Water is obtained by desalinizing seawater as part of electricity generation, a process that also yields carbon dioxide. This kind of technology is adaptable to much of the earth's uninhabited, unproductive desert shoreline. It could reach feasibility and sustain large populations when nuclear reactors are employed to generate electricity, distill the water, and recover valuable minerals from the ocean.

Foliar feeding. Phenomenal increases in plant productivity are often achieved when specific nutrients are applied to the foliage. Foliar feeding with two sprays of zinc sulfate at the rate of one pound of zinc to the acre (1,120 grams per hectare) has increased yields of beans in one area from zero in the control to 23 bushels. In another, a 3 percent spray solution of ferrous sulfate resulted in grain sorghum yield of 4,000 pounds per acre (4,500 kilograms per hectare) compared with 250 pounds per acre (280 kilograms per hectare) for nonsprayed crops. The critical stages in plant development when added growth factors increase productivity occur during the early seedling stage, at flowering, and at early seed or fruit development. Half of the phosphorus requirement of corn or tomato occurs by the time only 20 percent of growth has taken place. The early stage responds to "starter" fertilizer. The second stage is when foliar feeding has great potential; then leaf areas are large, and the canopy is nearly complete. Foliar feeding is also promoted by new cultural practices such as narrow rows, equidistant spacing, high plant populations, irrigation water with added fertilizer, and ultra-low-volume aerial spraying with undiluted formulations. Foliar feeding of nitrogen on cereal crops will assume importance in countries such as India, where an acute shortage of fertilizer demands high utilization efficiency.

Growth regulation. Chemical regulators are becoming important in extending the limits of crop productivity by manipulating the genetic structure of the plants. Cyclocel ([2-chloroethyl]trimethylammonium chloride) has been found to be the chemical equivalent of the genetically developed, short-straw wheat varieties resistant to lodging (being beaten to the ground by weather). Wheat, tomato, and others are very sensitive to this chemical; treated wheat results in a short, thick-stemmed plant, dark-green leaves, high tillering (side shoots), better filled heads, and resistance to lodging. Cyclocel is used extensively in Europe, where heavy fertilization, high harvest-time rainfall, and intensive cultivation produce tall, leafy stalks and lodging losses as high as 50 percent. About two pounds per acre has been effective. The chemical is undesirable for barley and rice. Soybeans treated with TIBA (2,3,5-triiodobenzoic acid) produce small leaves, short petioles, and short, strong stems resistant to lodging.

A high-potential substance for regulating growth of tree fruits is Alar (N-dimethylaminosuccinamic acid). Young trees treated with Alar will bear fruit after four

Alar as a growth regulator for tree fruits

years instead of the usual seven to ten. Alar can delay flowering for frost protection, and midsummer applications can delay preharvest fruit drop, improve colour, and prolong storage life. It accelerates cherry ripening and improves cultural factors in grapes. Gibberellin has been used for several years as a growth substance on grapes and citrus fruits, mainly to improve market acceptability of the fruits.

Mechanization. Mechanized planting and harvesting are the backbone of high per-man productivity. Research on planting improvements has led to development of equipment to create an ideal soil environment for seed to promote rapid, vigorous germination and the establishment of adequate stands. The mechanical harvesting of grain, forage, beans, potatoes, and sugar beets has been accomplished. Much work needs to be done, however, to achieve full mechanization of vegetable and fruit harvesting. In developed nations with high labour costs, the crops that will survive commercially are those the production and harvesting of which can be mechanized. Mechanical harvest of tomatoes was delayed until varieties were developed so that all fruit ripened simultaneously. Newly planted fruit orchards are being designed for machine harvest by special tree training. Seeds for fields of pickling cucumbers intended for mechanical harvest are drilled in narrow rows, which more than double the yield; such mechanization is accompanied by bulk hydrohandling and cooling. The double-trellis system of training grapes for ease of mechanical harvest has increased yields up to 100 percent. These developments indicate that mechanization of fruit and vegetable harvest demands research toward new kinds of plants and also creates new potentials in productivity.

Eliminating cultivation and planting in rows

Weed control. It now seems possible that weeds can be virtually eliminated from the major row crops and from most tree fruits. Application of selective herbicides by means of timed sprinkler irrigation tends to eliminate the need for planting crops in rows, for there is no longer any need for the usual double-row weeding cultivation. This type of weed control permits production of corn, sugar beets, and other crops in solid rows with consequent increase in plant population and, in turn, requires the development of new machine designs.

It has been discovered that certain chemicals used normally as herbicides also show promise for increasing protein production. The original observations were that simazin, 2-chloro-4,6-bis(ethylamino)-*s*-triazine, applied around fruit trees not only controlled weeds but also enhanced shoot growth and foliage colour. Subsequent laboratory tests revealed that at subtoxic levels simazin increased the accumulation of protein in rye by 20 to 80 percent. Seed from treated pea plants contained 40 percent more protein with no difference in amino-acid distribution between the protein of treated and nontreated plants. The laboratory results were confirmed in field trials, where maximal increases of total crude protein were obtained by use of less than one-half pound per acre (560 grams per hectare) of simazin applied as postemergence spray. Similar results were noted in experiments with ryegrass forage, beans, rice foliage, alfalfa forage, and oats. Subsequent field trials with wheat have shown substantial increases in grain protein following application of one-quarter to one-half pound of simazin per acre. Chemical herbicides other than simazin have also influenced protein yield; other triazines, for example, are effective as well as some of the substituted uracils. Confirmation of these results has been reported in India with rice and moong; experiments in Mexico are still under way. It appears that greatest benefits of Simazin application will be obtained in areas where soils are low in nitrogen and where nitrogen fertilizers are scarce.

Livestock research and development. In the developed countries the trend toward feeding ruminant livestock with feed grains and forages grown on arable land is causing concern as an inefficient utilization of land in the face of the increasing need for grain as human food. More research effort, therefore, is likely to be directed toward increasing the yield of animal products from rangeland and permanent pasture. Large-scale nitrogen fertilization of pastures and other forages will improve their yield and nutritive value, and the practice will become justified as new forages are developed and the cost of nitrogen is lowered; in fact, world fertilizer consumption is expected to double by 1980. An alternative would consist in adding urea and ammonia salts to carbohydrates and forages as the sole source of nitrogen. That would increase the bacterial synthesis of protein in the rumen of cattle and sheep. Much urea is now fed to livestock; if its price trends low enough, it could be fed with forage whose protein content would thus be of less importance than its digestibility. Another possibility lies in the chemical treatment of high-lignin forages to increase digestibility. These techniques offer some hope for reduced consumption of grain by ruminants.

Large-scale fertilization of pastures

Dairy cows. The full potential of the dairy cow has not yet been realized. Worldwide, milk production per cow has increased by 3 or 4 percent per year for more than ten years—largely from genetic improvements, better nutrition, and improved management. The effects of physiological and genetic limitations will probably not be felt until the year 2000. Meanwhile, breeding research and adoption of controlled environment housing will continue or even accelerate the rate of increase in milk production.

Beef cattle and swine. Productivity of beef cattle and swine is related to reproductive capacity, rate of growth, and the lean-meat percentage of the carcass. Of these factors, low reproduction rate is the most important. The percentage of mated beef cows who raise calves to maturity is 75 to 80 percent in the United States; 90 percent for The Netherlands and Denmark; but only 30 to 50 percent in developing countries. In the future, selection programs designed to increase the percentage of multiple births in beef cattle and swine will help to increase production efficiency. Reduction in losses both before and after birth can be realized by additional research. Progress in breeding and selection of swine for lean-meat carcasses has been achieved, a process that should be done also with beef cattle, because market surveys have shown preference for lean Holstein beef rather than the usual English breeds. The preference of beef feeders for crossbreeds is already well-established; heterosis (hybrid vigour) often results in a 5 to 20 percent increase over the best parental breed. Research on climate control at high temperature should result in improved production efficiency.

Efficiency in poultry production

Poultry. The efficiency of poultry production has undergone marked improvement. The potential limit for the laying hen is one two-ounce (56-gram) egg from each 2½-pound (1,125-gram) hen daily for one year, based on three pounds (1,350 grams) of feed per dozen eggs. The average hen, however, lays only about 230 eggs per year, weighs four pounds (1,800 grams), and requires four pounds of feed to produce a dozen eggs. One point of potential gain lies in causing the hen to lay eggs without shells, although public acceptance might be limited. The ultimate for feed conversion in the broiler is one pound of meat for every pound of feed utilized in a five-week growing period. The present conversion ratio is 2.2 pounds feed for every pound of meat, utilized over a seven- to nine-week growing period. Increased productivity in poultry of the future will result from new genetic combinations and further improvements in environmental and management factors. One promising area for future research may be in exploration of performance as related to environmental conditions.

The future for livestock. The biological limits in livestock productivity have not been achieved in any class of farm animal. Fundamental research in biochemistry, endocrinology, environmental physiology, genetics, parasitology, microbiology, and disease has provided knowledge that will make progress possible. For example, swine have reached market weights of over 200 pounds (90 kilograms) in 120 days with little more than two pounds of feed per pound of gain. Some hogs have yielded loin eyes of nine square inches (58 square centimetres) and 50 percent of the carcass weight in ham and

loin. Sows have farrowed more than 30 pigs in one litter. Knowledge of the endocrine system and its hormone production now makes it possible to synchronize estrus (heat), cause superovulation, and ultimately to control body composition. Fibre digestion efficiency in ruminants depends on the type of bacteria and protozoa in the rumen. By controlling these microbes, the quality of protein synthesized can be improved. There is reason to hope that in the fairly near future hogs will be marketed 100 days from birth and beef animals within six months; that each hen will produce two eggs a day; and that the beef cow will produce two calves per year.

Sponsors of research. In most countries the burden of agricultural research and development is borne by both government and private industry. In the United States, for example, many early advancements were made by innovative farmers themselves, but the land-grant college system authorized in 1862 for each state tended to concentrate research in the hands of professionally trained people. The state experiment stations, which began in 1887, followed by extension services in 1914, were developments that rounded out organization for production of research-based technology. Those research efforts, teamed with those of the United States Department of Agriculture, created the food-production system that allows 5 percent of the population to feed the other 95 percent at low cost and on a constantly diminishing land area. The public institutions have received massive research assistance from private firms, particularly in the technologies of farm machinery, farm chemicals, animal nutrition, and seeds. The intimate public–private coordination and cooperation is probably unique, and the resulting benefits to the American people have stimulated many other countries, particularly the developing nations, to establish similar organizations. International bodies such as the Food and Agriculture Organization of the United Nations and the International Rice Research Institute in the Philippines have made enormous contributions to agricultural technology since World War II. These groups have been particularly successful in improving productivity in the developing nations.

AGRICULTURAL TECHNOLOGY AND THE WORLD FOOD CRISIS

Whether all men can have enough food depends on what man does about both food and population. It has been said that the solution to the world food crisis lies in increasing the per capita food production of the developing nations while reducing their rate of population increase, instead of increasing productivity in developed areas.

The principal strategy for increased production is technological. Technology must be created and nourished, however, by a host of institutions, public and private. Research, education, capital formation, industrial development, governmental attitudes and services, land-tenure patterns, and many other factors must be favourably integrated or changed before technology can realize its full potential. If populations are not controlled in one way or another, technology may not suffice to relieve the world food crisis.

Asia's green revolution

Introduction of new methods usually demands concurrent improvements in others. An example is the new rice and wheat varieties resulting in the so-called green revolution of Asia. Obtaining full advantage of these biological improvements demands better fertilizer and irrigation and more mechanization. In turn, the transport and storage of the increased yield demands facilities (vehicles and structures) that may not have existed before in a particular area.

Expansion of croplands. A prime question related to future food supplies is whether the land surface can be more completely utilized. It is estimated, on the basis of available water, that one-third of the earth's total land area is capable of being cropped. In the early 1970s only about one-fifth was being so used, and 10 percent of that was irrigated. Of course, the best land already was under cultivation, with the vacant areas mainly in tropical South America and Africa and in desert lands of Australia and Africa. Assuming that water cannot be delivered to the desert areas in significant quantities in the near future, this leaves the tropics as the only remaining area of potential cultivation.

Wide areas of vacant tropical lands are well watered and suitable for farming. A third of all arable and potentially arable land lies in the humid tropics. Most tropical soils are leached of plant nutrients, and some are lateritic (a product of rock decay); nutrients could be supplied to the leached soils, but the laterite hardens irreversibly in use, creating limited production possibilities. The traditional agriculture, as has been mentioned, is bush fallow; *i.e.* a plot is used for a year or two after clearing and then allowed to fallow for a decade or more. The bush-fallow system supports only small populations.

The best tropical lands lie in volcanic regions at moderate elevations. These are found in Costa Rica, Malagasy Republic, Kenya, the Indonesian island of Java, and elsewhere. Low volcanic areas and alluvial regions of the tropics are also highly cultivated, such as the valleys of the Nile, Ganges, Irrawaddy, and Mekong. The production of such areas can be enhanced by technology.

Other tropical regions seem almost hopeless because of infertile soils and rainfall exceeding 100 inches (2.5 metres) per year with no dry season. Borneo is nearly uninhabited because of these conditions, and likewise the Congo and Amazon basins. It is asserted that drainage and generous fertilization would make these lands productive, but the cost of the plant nutrients would be prohibitive. Even so, the culture of rubber or oil palm in Borneo, Malaysia, Liberia, Ghana, and the Congo Basin can provide support for some populations.

Technology theoretically could do much to increase cultivation in the vacant tropical lands. The first essential is fertilizer, but it must be transported long distances or even imported. Nutrients by themselves are insufficient; the soil and its capability must be understood. Protection from insects and disease is vital. Machines, seeds, petroleum, and credit are also required. Lacking technology and fertility, the vacant tropical lands do not appear to offer much potential for appreciable food production increase.

Utilization of forest and grazing land. Cropland resources can be gained to some extent from areas now used as forest and grazing land, although worldwide the prospect of deriving much arable land from grassland is not particularly good. One reason is that the grazing lands are less desirable for crops. Too, the world's grasslands are valuable in production of animal products. Forests are also less likely to occupy land suited for crop production, but often they do. In some countries, the forest products are essential to the economy, as is true of the teak in Java and mahogany in the Philippines. Shifting patterns of wood technology, however, can change the forest-type economy; the competition from synthetic fibres, for example, may cause some forest land to be shifted to cropping. Generally, however, the ratio of forest to other land use is determined more by the character of the land itself than by the needs of man.

Forest products versus crops

In the less-populated developed countries considerable land suitable for cultivation is now forest or used for grazing. The United States and Canada are examples, and probably the Soviet Union as well.

Other possible ways of obtaining arable land would include reclamation of swamps and deserts and polder building to replace shallow ocean with land.

Other requirements. Though it surely will be possible to open some new areas to cultivation, in the long run this way of meeting the food crisis will probably not suffice. Increased production from currently available land will likely prove more effective, at least for a while; this can be done almost immediately in the more advanced countries. The main requirements, other than land, are fertilizer and water.

Fertilizer. Fertilizers are obtained from nonrenewable material. The manufacturing technology depends heavily on energy sources for nitrogen fixation and, eventually, for phosphate treatment. Mining and transport also are considerations. The raw material supply—of potassium compounds, sulfur and sulfates, phosphate rock, lime-

stone, natural gas or coal, and air—appears to be adequate for the foreseeable future. The problem of high cost, however, would appear if the total needs of the underdeveloped areas were to be met, even though the cost of fertilizer is a small part of the total crop cost.

Water supply. The value of adequate water at the right time to crop production is clear; whether the region is arid or humid, highest productivity is linked to timely irrigation. In humid regions sources of water are near at hand, but in dry regions the source of water is of first importance. Large-scale distillation of seawater, using nuclear power, may become highly significant in some areas. In others, long-distance transport of water from areas of surplus to the dry areas could step up production enormously. These modes of gaining water for agriculture will appear when the need becomes acute.

Other possible food sources. *New protein products.* Expansion of protein production from unconventional sources could help improve food supply. Petroleum and natural gas provide a source of protein. Protein made from leaves is becoming available in India; it is even produced from wool in New Zealand. A British company announced development of a protein derived from the action of fungus on starch. Meat analogues made from spun soy protein are widely available in North America. Torula yeast grown on animal or vegetable wastes can supply protein. The recycling of animal wastes as feed supplement also is a possibility. Synthetic amino acids made from abundant raw materials (lysine, methionine, threonine) can provide protein supplement to grains. Urea is now commonly used as a nitrogen source in feeding ruminants; eventually, it may afford another way of making proteins from nonproteinaceous materials. These and other schemes are indeed exciting examples of new technology, but most of them demand some degree of industrial capacity for significant production—capacity not yet available to many developing countries.

Meat from soy protein

Food from developed countries. It is probably true that agricultural production in developed countries could double or triple in about five years by removal of government restraint and wide adoption of now-current technology. American cotton growers could easily supply the world demand, for example, if this were to become desirable. The enormous surplus of food resulting from such a policy could then be made available to those who need it. But the farmer could not increase production without incentives of some kind. Transport, storage, and distribution costs would still have to be paid; world markets would be totally disrupted; and the countries receiving the bounty would lose incentive to provide for their own nutrition or to limit their population. This course of action does not appear to be practicable, except in emergency situations. On balance, the industrialized countries can best aid the others by helping them to adopt the agricultural technology and institutions that offer the best hope of becoming self-sufficient.

Animal products. There are those who believe that man will eventually have to eliminate animal products from his diet because animals consume grain that human beings too can eat. Meanwhile, the technology of animal production moves ahead, because rising income almost invariably increases demand for the various animal products. Breeding and adaptation to the exotic climates will increase the value of animals to the developing countries, and the unique ability of the ruminant to utilize grassland will play an increasing role.

Technology and population growth. In the recent past, technology has enabled agricultural output to keep slightly ahead of population growth in the developing countries. Their output increased 2.8 percent annually, while their population gained 2.3 percent. But per capita agricultural production rose only 0.1 percent compared to 1.5 percent in the industrialized countries. Thus, during the 1950s and 1960s, production was still inadequate to meet the needs of greater numbers of people with rising incomes. Even in nations where productivity per farmworker rose substantially, often the gain was not enough to improve rural income and at the same time transfer capital from agriculture to other sectors of the economy.

Farm output and population gain

During the 1970s, farm production in the 50-odd developing countries will have to increase about 4 percent each year to feed a population that is growing at a rate of 2.5 percent, if per capita consumption is to be raised and diets upgraded. Crop area is expected to increase only about 1 percent per year; therefore, crop output per acre must increase 3 percent to attain the 4 percent required gain—about twice the rate attained in the 1950–70 period. Those are overall figures for the developing countries; of course, there will be individual variations.

Another aspect of the problem is that farm population is rising at 1.7 percent per year in those countries, despite migration to the cities. If total farm output is to advance by 4 percent, output per worker would have to rise 2.3 percent (4 percent minus 1.7 percent). This order of advance in productivity is much higher than that experienced since 1950. Even so, this increase would provide only modest gains in real per capita income for farm people; at present, farmers' incomes average 25 to 50 percent of nonfarmers'.

To achieve such food-production goals, the developing countries must expand markets for their farm products, develop land and water resources, buy more inputs from the nonfarm sector, and use more technology more efficiently. Barring the appearance of spectacular new technology, the basic technological ingredients will be plant breeding, fertilizers, and crop protection.

BIBLIOGRAPHY. Books on crop farming include: C.B. RICHEY (ed.), *Agricultural Engineers' Handbook* (1961); R.E. STEWART, "Introduction to Systems Engineering in Agriculture," *Trans. Am. Soc. Agric. Engrs.*, vol. 10 (1967); GRAHAM F. DONALDSON, *Farm Machinery Capacity*, Royal Commission on Farm Machinery Study No. 10, Canada (November 1970); GLENN O. SCHWAB *et al.*, *Soil and Water Conservation Engineering*, 2nd ed. (1966), recommended for general reading; UNITED STATES DEPARTMENT OF AGRICULTURE (USDA), *Soil: Yearbook of Agriculture, 1957* (1957), a general approach; *Power to Produce: Yearbook of Agriculture, 1960* (1960); *Protecting Our Food: Yearbook of Agriculture, 1966* (1966); *Contours of Change: Yearbook of Agriculture, 1970* (1970); WILLIAM R. GILL and GLEN E. VANDEN BERG, *Soil Dynamics in Tillage and Traction*, Agriculture Handbook No. 316 (1967); HARRIS PEARSON SMITH, *Farm Machinery and Equipment*, 5th ed. (1964), a general book; NYLE C. BRADY (ed.), *Agriculture and the Quality of Our Environment* (1967); FACULTY OF ENGINEERING, UNIVERSITY OF THE WEST INDIES, TRINIDAD AND TOBAGO, *Proceedings, First International Seminar on Agricultural Engineering* (March 1970); *Van Nostrand's Scientific Encyclopedia*, 4th ed. (1968), article on hydroponics.

Sources on weather and agricultural technology include: JEN-HU CHANG, *Climate and Agriculture* (1968), a general approach; USDA, AGRICULTURAL CLIMATOLOGY SERVICE OFFICE, *Weekly Weather and Crop Bulletin;* ROBERT H. SHAW (ed.), *Symposium on Ground Level Climatology* (1967); and ALLEN L. HAMMOND, "Weather Modification: A Technology Coming of Age," *Science*, 172:548–549 (1971).

For information on agricultural experimentation and research, see S.H. WITTWER, "Progress of Research and Technology on the Food Supply in the United States," *Mich. Agric. Exp. Stn. J.*, article no. 4585 (1969); and for information on pollution and agricultural technology, see CECIL H. WADLEIGH, "Wastes in Relation to Agriculture and Forestry," *Misc. Publs. U.S. Dep. Agric. 1065* (March 1968); USDA, SECRETARY OF AGRICULTURE, *A Report to the President: Control of Agriculture-Related Pollution* (January 1969); U.S. COUNCIL ON ENVIRONMENTAL QUALITY, *Environmental Quality*, first annual report (August 1970); AMERICAN CHEMICAL SOCIETY, *Cleaning Our Environment, The Chemical Basis for Action: A Report* (1969); USDA, *A National Program of Research for Environmental Quality: Pollution in Relation to Agriculture and Forestry* (1968); THOMAS R. DETWYLER, *Man's Impact on Environment* (1971); STANTON MILLER and CAROL KNAPP, "Pest Control: A Special Report," *Envir. Sci. Tech.*, vol. 5, no. 5 (1971); and ALBERT H. MOSEMAN (ed.), *Agricultural Sciences for the Developing Nations* (1964).

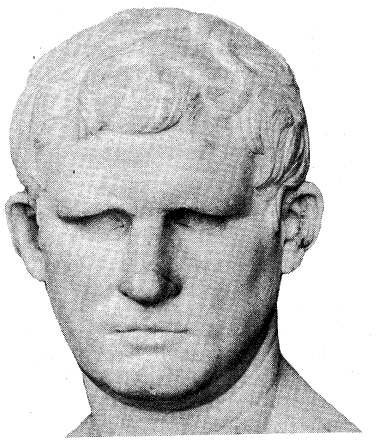
Information on agricultural technology and the world food crisis may be found in the NATIONAL ACADEMY OF SCIENCES, COMMITTEE ON RESOURCES AND MAN, *Resources and Man: A Study and Recommendations* (1969); IOWA STATE UNIVERSITY CENTER FOR AGRICULTURAL AND ECONOMIC DEVELOPMENT, *Alternatives for Balancing World Food Production and Needs* (1967); and the USDA, *The Farm Index* (monthly), see esp. "The Pace of Agriculture in Developing Nations: Challenges Loom in the 1970's," 9:17–20 (1970).

(R.E.S.)

Agrippa, Marcus Vipsanius

Marcus Vipsanius Agrippa was an influential friend and deputy of the first Roman emperor, Augustus (reigned 27 BC–AD 14). He was responsible for several crucial military successes in Augustus' rise to power; in the newly established empire, he attained a rank that clearly marked him as second only to the Emperor in authority. Of modest birth but not a modest man, Agrippa was disliked by the Roman aristocracy. In his own interest he scrupulously maintained a subordinate role in relation to Augustus, but he felt himself inferior to no one else.

Cliche Musees Nationaux, Paris



Agrippa, marble portrait bust, early 1st century BC. In the Louvre, Paris.

Agrippa was nearly an exact contemporary of Augustus. He was born probably in early 63 BC into the obscure family of Vipsanii. Virtually nothing is known of his early life until he is found as the companion of Octavian (the future emperor Augustus) at Apollonia, in Illyria, at the time of Julius Caesar's murder in 44. Octavian, the adopted son of Caesar, returned with Agrippa to Italy to make his political claim as Caesar's heir. In 43 Agrippa is thought to have held the office of tribune of the plebs; presumably in this capacity he prosecuted the tyrannicide Cassius, then absent in the East. In the struggle for power after Caesar's death, Agrippa served as one of Octavian's key commanders. In 41–40 he fought against Mark Antony's brother Lucius. In 40 he held the post of praetor urbanus (magistrate mainly in charge of administration of justice at Rome) and was a major figure in negotiating a settlement between Octavian and Antony at Brundisium. During the next two years he was away on campaigns in Aquitania and on the Rhine. When he returned to Italy, he conspicuously refused to celebrate a triumph for his successes in the north, but in 37 he held the office of consul. In the spring of 37 Octavian and Antony came to an agreement at Tarentum, and it must have been then that Antony arranged the marriage of Agrippa to the daughter of Titus Atticus, a wealthy friend of Cicero.

Octavian's general

Octavian's efforts to resist at sea the son of the republican general Gnaeus Pompey, Sextus Pompeius, had not met with success. Agrippa therefore took charge of the operations. He constructed a fine harbour at Puteoli in the Bay of Naples and then won, in 36, two decisive naval victories (at Mylae and Naulochus), which ended the threat from Pompeius. For this achievement Agrippa accepted the award of a golden crown. In 35–34 Octavian waged a vigorous campaign in Dalmatia, and in this Agrippa had a distinguished military role. In 33 Agrippa served as curule aedile (magistrate of public buildings and works) at Rome, even though it was a much lower post than the consulate that he had already held. He used the opportunity to win favour for Octavian by spending his own funds lavishly on building baths, cleaning sewers, and improving the water supply. When Octavian and Antony finally came into direct conflict at the Battle of Actium in 31, Agrippa commanded the fleet and was chiefly responsible for Octavian's victory.

During Octavian's absence from Rome after Actium, Agrippa managed affairs in the city together with Maecenas, the great patron of poets. In 29–28 Agrippa and Octavian jointly conducted a census and carried out a purge of the Senate; in 28 and 27 Agrippa held the consulate again, both times with Octavian (from 27, Augustus) as his colleague. In 23, a year of constitutional crisis, Augustus fell ill and presented his signet ring to Agrippa, who seemed thus to be designated the Emperor's successor. Agrippa took Augustus' daughter Julia as his wife after divorcing a niece of Augustus (Marcella the Elder), who had replaced Atticus' daughter as his wife some four or five years previously.

Agrippa went immediately to Mytilene on the island of Lesbos, from which he administered affairs in the East. The nature of Agrippa's constitutional power (*imperium*) at this time is controversial. It has been argued whether the Senate in 23 gave him an *imperium* greater than that of any other provincial governor in the East (*imperium majus*). After Augustus' death Roman historians claimed that Agrippa's sojourn at Mytilene was a kind of exile as a result of Augustus' preference for his own nephew Marcellus. This appears implausible. Agrippa was soon back in Rome to act on behalf of the Emperor, who himself left for the East in 22. Before Augustus' return, in 19, Agrippa had left for Gaul and Spain. In Spain he finally subdued the recalcitrant Cantabrians.

Returning to Rome in 18, Agrippa received the power of a tribune (*tribunicia potestas*), which Augustus also possessed. Perhaps, too, he received an *imperium majus*, if he had not been granted it in 23. He participated in Augustus' celebration of the Secular Games at Rome in 17, after which he returned to the East as vicegerent of the Emperor. In 15 he accepted an invitation from Herod I the Great to visit Judaea; while in the East, he established colonies of veterans at Berytus and Heliopolis, in Lebanon. He next settled an uprising in the Bosporan kingdom on the Black Sea and set up the cultivated dynast Polemo as king. Herod led a fleet to support Agrippa in the Bosporan affair, and, when it was over, the two travelled together along the coast of western Asia Minor.

Vicegerent in the East

In 13 Agrippa's *tribunicia potestas* was renewed, and at this time without doubt he received (or had renewed) a grant of *imperium majus*. Troubles in Pannonia required his presence, but the rigours of the winter of 13–12 caused a fatal illness; he died in March of 12 BC. Augustus delivered a funeral oration in honour of his colleague; a fragment of that oration, in Greek translation, has recently come to light.

Agrippa deserved the honours Augustus heaped upon him. It is conceivable that without Agrippa, Octavian would never have become emperor. Rome remembered him for his generosity in attending to aqueducts, sewers, and baths; and in the mid-20s he completed the celebrated Pantheon. One of Agrippa's five children by Julia, Agrippina the Elder, was the mother of one emperor (Caligula) and the grandmother of another (Nero). Of Agrippa's writings, an autobiography is lost, but an extensive geographical commentary that he wrote influenced the extant works of the geographer Strabo and of Pliny the Elder.

BIBLIOGRAPHY. The best full-length study of Agrippa is MEYER REINHOLD, *Marcus Agrippa* (1933). There is also an important account of him in SIR RONALD SYME'S classic work, *The Roman Revolution* (1939). On problems arising from the new fragment of Augustus' funeral oration on Agrippa, see E.W. GRAY, "The Imperium of M. Agrippa," *Zeitschrift für Papyrologie und Epigraphik*, 6:227–238 (1970).

(G.W.Bo.)

Aḥmad ibn Ḥanbal

Aḥmad ibn Ḥanbal, theologian, jurist, and expert on the Traditions of Muḥammad, the founder of Islām, is honoured as a steadfast martyr for his faith. His teachings form the basis of one of the four orthodox schools of law of the larger of Islām's two major branches, the Sunnī. His school, the Ḥanbalī, is considered more strictly traditionalist than the other three: the Ḥanafī, Mālikī, and Shāfi'ī. It is also the smallest.

Ibn Ḥanbal was born in Baghdad in 780 and died there in 855. Of pure Arab stock, he belonged to the tribe of Shaybān through both parents. He was still an infant when his father died at 30. When Ibn Ḥanbal was 15 he began to study the Traditions (Ḥadīth) of the Prophet Muḥammad. Seeking to learn from the great masters of his day, he travelled to the cities of Kūfah and Basra in Iraq; Mecca, Hejaz, and Medina in Arabia; and to the lands of Yemen and Syria. He made five pilgrimages to the holy city of Mecca, three times on foot. Ibn Ḥanbal led a life of asceticism and self-denial, winning many disciples. He had eight children, of whom two were well known and closely associated with his intellectual work: Ṣālih (d. 880) and ʿAbd Allāh (d. 903).

The central fact of Ibn Ḥanbal's life is the suffering to which he was subjected during the inquisition, known as *al-miḥnah*, ordered by the caliph al-Ma'mūn. But for this great trial, and the unflagging courage he displayed in the face of his persecutors, Ibn Ḥanbal would most likely have been remembered solely for his work on the Traditions. As it is, he remains to this day, in addition to his recognized stature as an expert on Traditions, one of the most venerated fathers of Islām, a staunch upholder of Muslim orthodoxy.

The inquisition was inaugurated in 833, when the Caliph made obligatory upon all Muslims the belief that the Qur'ān was created, a doctrine of the Muʿtazilites, a rationalist Islāmic school that claimed that reason was equal to revelation as a means to religious truth. The Caliph had already made public profession of this belief in 827. Heretofore, the sacred book had been regarded as the uncreated, eternal word of God. The inquisition was conducted in Baghdad, seat of the ʿAbbāsid caliphate, as well as in the provinces. It lasted from 833 to 848, a period involving the reign of four caliphs, ending during the caliphate of al-Mutawakkil, who returned to the traditionalist view.

Torture and imprisonment

At the risk of his life, Ibn Ḥanbal refused to subscribe to the Muʿtazilī doctrine. He was put in chains, beaten, and imprisoned for about two years. After his release he did not resume his lectures until the inquisition was publicly proclaimed at an end. Some orthodox theologians, to survive the ordeal, had recanted, and later claimed the privilege of dissimulation, *taqīyah*, as a justification for their behaviour. This is a dispensation granted in the Qur'ān to those who wish to avail themselves of it when forced to profess a false faith, while denying it in their hearts. Other theologians, following the example of Ibn Ḥanbal, refused to repudiate their beliefs.

In 833 Ibn Ḥanbal and another theologian, Muḥammad ibn Nūḥ, who had also refused to recant, were cited to appear for trial before Caliph al-Ma'mūn, who was in Tarsus (now in modern Turkey) at the time. They were sent off in chains from Baghdad; but shortly after beginning their journey, the Caliph died, and on their trip back to the capital, Ibn Nūḥ died.

Ibn Ḥanbal was ordered to appear before the new caliph al-Muʿtaṣim. He was on trial for three days, and on the third day, after the learned men disputed with him, there followed a private conference with the Caliph who asked Ibn Ḥanbal to yield at least a little so that he might grant him his freedom. Ibn Ḥanbal made the same reply he had been making from the beginning of the inquisition; he would yield when given some ground for modifying his faith derived from the sources he regarded as authoritative, namely the Qur'ān and the traditions of Muḥammad. Losing patience, the Caliph ordered that he be taken away and flogged. Throughout the flogging the Caliph persisted in his attempts to obtain a recantation, but to no avail. Ibn Ḥanbal's unflinching spirit was beginning to have its effect upon the Caliph; but the latter's advisers warned that if he desisted from punishing him, he would be accused of having opposed the doctrine of his predecessor al-Ma'mūn, and the victory of Ibn Ḥanbal would have dire consequences on the reign of the caliphs. But the Caliph's treatment of Ibn Ḥanbal had to be suspended, nevertheless, because of the mounting anger of the populace gathering outside the palace and preparing to attack it. Ibn Ḥanbal is reported to have been beaten by 150 floggers, each in turn striking him twice and moving aside. The scars from his wounds remained with him to the end of his life.

The inquisition continued under the next caliph, al-Wāthiq, but Ibn Ḥanbal was no longer molested, in spite of attempts on the part of his opponents to persuade the caliph to persecute him. The new caliph, like his predecessor, was most likely influenced by the threat of a popular uprising should he lay violent hands on a man popularly held to be a saint. The momentum of the inquisition carried it two years into the reign of al-Mutawakkil, who finally put an end to it in 848.

Ibn Ḥanbal earned the greatest reputation of all the persons involved in the inquisition, and the everlasting gratitude of the Muslim people. He is credited with having held his ground in the face of all odds, saving Muslims from becoming unbelievers. At his funeral the procession was estimated at more than 800,000 mourners.

Collection of the Traditions of Muḥammad

The most important of Ibn Ḥanbal's works is his collection of the Traditions of the Prophet Muḥammad. This collection was heretofore believed to have been compiled by the author's son (ʿAbd Allāh), but there is now evidence that the work was compiled and arranged by Ibn Ḥanbal himself. These Traditions were considered by Ibn Ḥanbal as a sound basis for argument in law and religion.

Historical scholarship regarding Ibn Ḥanbal and his school has suffered from a lack of sufficient documentation, among other things. There are, therefore, some opinions regarding Ibn Ḥanbal that bear closer scrutiny in the light of new documents and recent studies. Too much stress has been laid on the influence on him of the teachings of Shāfiʿī, the founder of the Shāfiʿī school, whom Ibn Hanbal apparently met only once. He had a high respect for Shāfiʿī but also for the other great jurists who belonged to other schools of law, without, for that matter, relinquishing his own independent opinions. He was against codification of the law, maintaining that canonists had to be free to derive the solutions for questions of law from scriptural sources, namely the Qu'rān and the *sunnah* (the body of Islāmic custom and practice based on Muḥammad's words and deeds). It was to this end that he compiled his great *Musnad*, wherein he registered all the traditions considered in his day acceptable as bases for the solution of questions, along with the Qu'rān itself.

Compilation of *Musnad*

The fact that the Ḥanbalī school was organized at all was due to the impact of Ibn Ḥanbal on his time. The other orthodox schools were already prospering in Baghdad when the Ḥanbalī school sprang up in their midst, drawing its membership from theirs. The lateness of the hour accounts for the relatively small membership attained by the Ḥanbalī school compared with the older schools. It is, however, not by the number of its members that the importance of the school and its originator should be judged but rather by their impact on the development of Islāmic religious history. In the Middle Ages the school acted as a spearhead of traditionalist orthodoxy in its struggle against rationalism. One of Ibn Ḥanbal's greatest followers, Ibn Taymīyah (1263–1328), was claimed by both the Wahhābīyah, a reform movement founded in the mid-18th century, and the modern Salafīyah movement, which arose in Egypt and advocated the continued supremacy of Islāmic law but with fresh interpretations to meet the community's changing needs. Ibn Ḥanbal himself is one of the fathers of Islām whose names have constantly been invoked against the forces of rationalism down through the ages.

BIBLIOGRAPHY. For bibliography and information pertaining to Ibn Ḥanbal's works, see FUAT SEZGIN, *Geschichte des arabischen Schrifttums*, vol. 1, pp. 502–509 (1967); for his life and works generally, see W.M. PATTON, *Ahmed ibn Hanbal and the Mihna* (1897); and H. LAOUST, "Le Hanbalisme sous le califat de Bagdad," in *Revue des Études Islamiques*, pp. 67–128 (1959); and the *Encyclopædia of Islam*, vol. 1, pp. 272a–277b (1960). On Sufism in the Ḥanbalī school of thought, see G. MAKDISI, "L'Isnad initiatique soufi de Muwaffaq ad-Din Ibn Qudama," in *L'Herne: Louis Massignon*, pp. 88–96 (1970).

(G.M.)

Ahmad Khan, Sir Sayyid

Sir Sayyid Ahmad Khan was for long known primarily as the founder of the Muhammadan Anglo-Oriental College (later the Alīgarh Muslim University), Alīgarh, Uttar Pradesh, India, but is now recognized to have been the principal motivating force behind the general Muslim revival in India in the latter half of the 19th century. Even Abul Kalam Azad, a Muslim leader critical of Sayyid's political policy, agreed that Sayyid had left his mark on all spheres of intellectual activity—religion, education, social life, language, and literature.

Sayyid Ahmad was born in Delhi on Oct. 17, 1817. His family, though progressive, was highly regarded by the dying Mughal dynasty. His father, who received an allowance from the Mughal administration, became something of a religious recluse; his maternal grandfather had twice served as prime minister of the Mughal emperor of his time and had also held positions of trust under the East India Company. Sayyid's brother established one of the first printing presses at Delhi and started one of the earliest newspapers in Urdu, the principal language of the Muslims of northern India.

The death of Sayyid's father left the family in financial difficulties and, after a limited education, Sayyid had to work for his livelihood. Starting as a clerk with the East India Company in 1838, he qualified three years later as a subjudge and served in the judicial department at various places.

Career as author

Sayyid Ahmad had a versatile personality and his position in the judicial department left him time to be active in many fields. His career as an author (in Urdu) started at the age of 23 with religious tracts. In 1847 he brought out a noteworthy book, *Āthār aṣṣanādīd* ("Monuments of the Great"), on the antiquities of Delhi. Even more important was his pamphlet, "The Causes of the Indian Revolt." During the Indian Mutiny of 1857 he had taken the side of the British, but in this booklet he ably and fearlessly laid bare the weaknesses and errors of the British administration that had led to dissatisfaction and a country-wide explosion. Widely read by British officials, it had considerable influence on British policy.

His interest in religion was also active and lifelong. He began a sympathetic interpretation of the Bible, wrote *Essays on the Life of Mohammed* (translated into English by his son), and found time to write several volumes of a modernist commentary on the Qur'ān. In these works he sought to harmonize the Islāmic faith with the scientific and politically progressive ideas of his time.

Career in education

The supreme interest of Sayyid's life was, however, education—in its widest sense. He began by establishing schools, at Muradabad (1858) and Ghāzīpur (1863). A more ambitious undertaking was the foundation of the Scientific Society, which published translations of many educational texts and issued a bilingual journal—in Urdu and English.

These institutions were for the use of all citizens and were jointly operated by the Hindus and the Muslims. In the late 1860s there occurred developments that were to alter the course of his activities. In 1867 he was transferred to Benares, a city on the Ganges with great religious significance for the Hindus. This was the year in which an elaborate religio-political festival, the Hindu *melā*, was first organized at Calcutta, signifying a noteworthy attempt at separate communal organization of the Hindus. At about the same time a movement started at Benares to replace Urdu, the language cultivated by the Muslims, with Hindi. This movement and the attempts to substitute Hindi for Urdu in the publications of the Scientific Society convinced Sayyid that the paths of the Hindus and the Muslims must diverge. Thus, when during a visit to England (1869–70) he prepared plans for a great educational institution, they were for "a Muslim Cambridge." On his return he set up a committee for the purpose, and also started an influential journal, *Tahdhīb al-Akhlāq* ("Social Reform"), for the "uplift and reform of the Muslims." A Muslim school was established at Alīgarh in May 1875, and after his retirement in 1876, Sayyid devoted himself to enlarging it into a college. In January 1877 the foundation stone of the college was laid by the Viceroy. In spite of conservative opposition to Sayyid's projects, the college made rapid progress. In 1886 Sayyid organized the All-India Muhammadan Educational Conference, which met annually at different places to promote education and to provide the Muslims with a common platform. Until the founding of the Muslim League in 1906, it was the principal national centre of Indian Islām. Sayyid was made a Knight Commander of the Star of India in 1888.

For the college, and even more for the conference, Sayyid was able to attract the elite of the Muslim intelligentsia, who not only helped his educational program but also transformed Urdu literature: poetry, historiography, fiction, and prose. Their efforts brought about a general awakening among Indian Muslims and gave them a new confidence and cohesion.

Political views

Sayyid advised the Muslims against joining active politics and to concentrate instead on education. Later, when some Muslims joined the Indian National Congress, he came out strongly against that organization and its objectives, which included the establishment of parliamentary democracy in India. He argued that in a country where communal divisions were all-important and education and political organization were confined to a few classes, parliamentary democracy would work only inequitably. Muslims, generally, followed his advice and abstained from politics until several years later when they had established their own political organization.

Sayyid Ahmad was a relic of the age of the great Mughals. Opinions differ regarding his politics but most impartial observers praise his solid achievements. Under his leadership the disintegration of Muslim India was halted and within a generation it was again prominent and influential. He died at Alīgarh on March 27, 1898.

BIBLIOGRAPHY. ALTAF HUSAIN HALI, *Hayat-i-Javid* (1901), the only comprehensive account of Sayyid Ahmad's life and work; G.F.I. GRAHAM, *The Life and Work of Syed Ahmed Khan*, rev. ed. (1909), contains little biographical information beyond 1887, but is still a useful account. See also S.K. BHATNAGAR, *History of the M.A.O. College, Aligarh* (1969); J.M.S. BALJON, *The Reforms and Religious Ideas of Sir Sayyid Ahmad Khan* (1949); and HADI HUSSAIN, *Syed Ahmed Khan: Pioneer of Muslim Resurgence* (1970).

(S.M.I.)

Aircraft

Within the context of this article, aircraft are considered to be vehicles operating primarily in the earth's atmosphere but at appreciable distances above the ground. Excluded, therefore, are such vehicles as spacecraft and air-cushion machines, which are covered in SPACE EXPLORATION and AIR-CUSHION MACHINES. Military aircraft are treated in AIRCRAFT, MILITARY.

All aircraft fall into two general categories—lighter-than-air or heavier-than-air. Several distinct types are recognized within each group. Each may perform a variety of missions calling for modifications for special usage.

Lighter-than-air craft rise and float because they displace a volume of air the weight of which is equal to or greater than the total weight of the aircraft. Such craft include balloons and airships.

A balloon is a light, nonporous bag (usually spherical) filled with hot air or a light gas (hydrogen or helium) to which a basket (or car) may be attached to hold pilot and passenger. The wind provides the only propulsion.

An airship is a mechanically driven balloon (usually of elongated shape) capable of being steered, to which a suitable car (or cabin) is attached. Airships are nonrigid when the shape of the gas container (envelope) is maintained solely by internal gas pressure; semirigid when the envelope is directly attached to a fore-and-aft keel structure; and rigid when a number of lifting gas cells are contained within an external fabric-covered structural skeleton the form of which is independent of each gas cell's content or pressure.

Heavier-than-air craft derive their flight capability (lift) from the dynamic reaction of air flowing around suitably

shaped surfaces (wings or airfoils). Such craft include gliders and sailplanes, conventional airplanes, short-takeoff and -landing (STOL) airplanes, and vertical-takeoff and -landing (VTOL) aircraft.

Gliders and sailplanes are motorless aircraft of light weight and efficient aerodynamic configuration capable of controlled flight in atmospheric updrafts.

Conventional airplanes are aircraft of normal configuration; *i.e.*, with fixed wings, control surfaces, power plants, landing gear, etc., assembled on an elongated fuselage housing pilot, passengers, and cargo. They may be landplanes, seaplanes, or amphibians.

STOL airplanes are of a normal in-flight configuration in which the magnitude and direction of thrust and lift may be modified to permit unusually steep takeoff or landing paths for use in limited airport areas.

VTOL aircraft are unconventional aircraft, usually based on rotating wing systems (helicopters or autogiros) or on rotatable-jet-thrust systems, capable of vertical lift-off or landing in areas of little more than their own overall dimensions.

This article is divided into the following sections:

I. Lighter-than-air craft
  Balloons
  Airships
    Configurations
    Missions and special uses
II. Heavier-than-air craft
  General considerations
    Aerodynamic principles
    Propulsion systems
    Control systems
    Instrumentation
    Flight simulators
    Materials and structures
  Fixed-wing aircraft
    Configurations
    General aviation
    Commercial transport
    Military aircraft
    Gliders and sailplanes
  Rotary-wing aircraft
    The autogiro
    The helicopter
  Experimental and research aircraft
    Supersonic transport airplanes
    STOL aircraft and convertiplanes
    Hypersonic aircraft

I. Lighter-than-air craft

Although balloon and airship flight is explained by the theorem (first postulated by Archimedes in the 3rd century BC) that a body will float when its own weight is equal to or less than the weight of the fluid it displaces, some 2000 years elapsed before anyone thought of atmospheric air as a fluid responsive to physical laws. More years went by before it occurred to anyone (except one or two students of philosophy) that Archimedes' principle offered a solution to the problem of manned flight.

BALLOONS

The invention of the balloon came about by accident, when two brothers, Joseph and Étienne Montgolfier, sons of a paper maker of Annonay, France, while watching a fire, realized that smoke always flowed upward. They concluded that smoke had some mysterious property that they called "levity." The observation led them to explore the possibility of applying the property to balloon flight. They experimented with larger bags and smokier fires. Shortly thereafter, the first men to travel in the air took off from Paris in a Montgolfier hot-air balloon. (See FLIGHT, HISTORY OF.)

Free ballooning has changed but little since the first flight in 1783. Hot air was generally abandoned as a lifting medium when so-called coal gas and hydrogen became available in the 1800s. Such gases are explosive, however, and, after many fatalities from fires, lightning strikes, and static discharges, these gases were eventually outlawed for both balloons and airships. Nonexplosive helium, although somewhat heavier than hydrogen, eventually became an accepted replacement.

In the mid-1960s hot-air ballooning was revived, as a sport. The classic free-balloon configuration has been retained; *i.e.*, a spherical (or pear-shaped) gas bag (envelope) beneath which a basket for the crew of one or two pilots is suspended. Initial inflation and maintenance of hot-air supply in flight is from one or more blowtorch burners suspended between the open end of the envelope and fed from pressurized gas (propane) carried in bottles in the basket. By controlling the rate of combustion at the burners, the temperature of the air in the bag can be controlled, thus controlling the rate of ascent or descent.

Use of blowtorch burners

Free-balloon pilots have control over only two movements—up and down. This control is exercised by adding or discharging ballast (usually in the form of sandbags) or by changing the volume (or density) of the lifting gas by allowing some to escape or by adding heat.

Once off the ground, balloons have no directional-control capability and so must drift with the wind. Pilots, therefore, try to maintain a flight altitude at which winds will carry them in a preplanned direction. The plan implies advance knowledge of and reliable predictions about meteorological conditions. In the early years such weather knowledge was largely a matter of luck, and disasters were frequent. Even with satellite weather surveillance and continuous radio communication, an attempt to cross the Atlantic by free balloon in 1970 ended in tragedy when the expedition encountered unexpected foul weather and was forced down off Newfoundland in an accident in which three lives were lost.

Historically, free balloons have participated in all kinds of aircraft missions except the commercial transport of passengers or cargo. The first aerial photographs were made from balloons. Geographical exploration by balloon, however, has seldom yielded important dividends. An attempt to drift over the North Pole in 1897 ended in a tragedy from which, however, undamaged photographic plates were recovered and successfully developed 33 years later.

Military use of balloons (free or tethered) dates back to the Battle of Fleurus in Belgium in 1794. Both sides in the American Civil War made limited use of tethered balloons as observation platforms for battlefield surveillance, a practice that became commonplace in World War I. In World War II, to discourage bombing attacks, balloon barrages were used in protective cordons around cities.

World War II balloon barrages

Free balloons, both manned and unmanned, have proved invaluable in atmospheric research and weather prediction. Early scientists made painstaking records of upper-air temperatures, air pressure, wind velocities, and other upper-air phenomena from balloons. These pioneers sometimes died from lack of oxygen or were frozen to death in the process, but it was from their work that the structure and physical characteristics of the earth's atmosphere were first plotted.

The most elaborate upper-atmospheric research (before the advent of the rocket and high-altitude-research airplanes) was undertaken by Soviet balloonists in 1933–34 and by United States Army and civilian teams in 1933–35. The Russians reached an altitude of 63,200 feet (19,260 metres). The United States "Explorer II," a spherical balloon carrying an insulated and pressurized gondola with a crew of two, made and recorded atmospheric observations up to an altitude of 72,395 feet (22,066 metres). Later such research was done largely by using unmanned devices; *i.e.*, by radiosondes carried to high altitudes by balloons that continuously transmit their instrument readings to the ground by radio.

AIRSHIPS

An airship (dirigible) differs from a free balloon in that it is steerable under power at the control of its pilot. The terms airship and dirigible are roughly synonymous.

Very shortly after the first balloon was demonstrated, its limitations (lack of propulsion system and directional control) became evident. The American statesman and philosopher Benjamin Franklin, after witnessing the first manned flight of the Montgolfier balloon in Paris, noted, "Since men may be supported in the air, nothing is want-

ing but some light and handy instrument to give and to direct motion." More than a century went by before such an "instrument" was devised.

Many ideas were tried and found to be unworkable. Seagoing experience suggested sails, paddles, and oars; but none had the slightest beneficial effect. The Archimedean concept of the screw propeller, successfully applied to marine propulsion in the mid-19th century, seemed to offer promise, but a suitable source of primary power was lacking. Manpower was insufficient. Steam engines were too heavy. The first practical solution came with the development of lightweight internal-combustion engines late in the 19th century.

Screw-propeller concept

A major change in overall configuration was also necessary—the spherical gas bag of the balloon had to be transformed into an elongated shape that could pass efficiently through the air. Again, this development came about as a result not of theoretical considerations or laboratory measurements but of empirically based concepts, such as the observation of fish. Fishlike fins and tails appear in many early airship sketches.

An airship derives lift from two sources: (1) by displacement of air as a balloon (static lift) and (2) from the reaction of airflow over its envelope and control surfaces when under way (dynamic lift). When the total weight of the machine and its contents is slightly less than the weight of the displaced air, the ship tends to rise and is said to be light. When the weights are the same (at any particular altitude), it is in equilibrium. It is said to be heavy when the weight somewhat exceeds the static lift.

The static-lift requirement varies in flight, depending on the weight of consumables (fuel, oil, water, ballast) aboard, the weight of accumulated snow or rain, the temperature of the lifting gas, and the pressure of the ambient air. Within limits, such variations can be compensated for by adjusting the flight attitude by the use of the aerodynamic controls. Thus, if the ship is light, it can be flown slightly nose-down to produce a negative dynamic lift. If it is heavy, a nose-up attitude produces an upward dynamic reaction to assist the deficient static lift.

Airships, because of their relatively light construction and large bulk, are difficult to handle on the ground in windy weather and are extremely vulnerable to violent meteorological phenomena encountered in flight. The cost and complexity of large ships and the required ground-handling equipment, coupled with a record of in-flight disasters due to unpredictable weather, eventually forced their total abandonment. A few enthusiasts still predict a useful future for airships, but in the 1970s no new projects were in sight or were seriously contemplated.

Configurations. From the first successful flight (1852) to about 1910, virtually all airships were of the nonrigid type. The general configuration consisted of an elongated envelope with pointed ends from which a light, openwork box-girder frame was suspended by cables. An engine and single propeller were mounted at the forward end of the frame, with rudder and elevator control surfaces at the rear. A rudimentary seat for the aeronaut was usually in the centre. Sometimes a catwalk was provided so that the pilot could have access to the engine or could shift his weight fore and aft to change longitudinal balance.

As aerodynamic theory and technology accumulated, refinements appeared. By the end of World War I, airships attained reasonably good streamlined shapes, and, for greater efficiency, control surfaces were attached directly to the stern sections of the envelopes. Openwork cars were replaced by enclosed fuselages housing power plants, passengers, and crew. Later, control cars and envelopes were integrated by attaching the cabins directly to the undersurface of the envelope, eliminating suspension cables. Engines and propellers were mounted in outboard housings (nacelles) attached to the cabin structure.

Semirigid airships were never built in significant numbers. A few experimental types appeared in the U.S. and in France before World War I, and the model reached its highest state of development in Italy in the early 1920s. Most notable was the "Norge," flown across the North Pole in 1926. The type had disappeared completely by 1928.

The rigid (commonly called Zeppelin) airship had the most spectacular and most useful history of all lighter-than-air craft. Between the launching of the LZ-1 (Luftschiff Zeppelin No. 1) from Lake Constance in southern Germany in 1900 and the fiery end of LZ-129 (the "Hindenburg") at Lakehurst, New Jersey, in 1937, the Zeppelins and their derivatives established the first commercial passenger services in the world and ranged over the globe on exploratory missions.

Except for structural and aerodynamic refinements incorporated over the years, the first and the last rigid airships had a strong family resemblance. Their hull framework was made up of longitudinally built-up aluminum-alloy lattice girders, spaced around transverse rings of the same material and construction. All were wire braced. A few all-wood designs were tried in Germany and in England but were soon abandoned. The rings divided the hull longitudinally into a number of spaces, each of which contained a separate gas-tight cell. Gas containers were originally made of light fabric lined with goldbeater's skin, a specially prepared animal membrane, but were ultimately fabricated from impervious synthetic materials. Outer-hull covering was weatherproofed fabric. Control surfaces were also fabric covered.

Until the early 1920s, control cabins and engine nacelles were usually mounted on rigid struts or brackets or both, external to the hull structure. In later ships, particularly those designed and built in the United States, the crew and passenger accommodations were integrated into the hull structure. In some cases in which nonflammable helium was used as the lifting gas, engines were actually installed inside the hull structure and drove external propellers through extension shafting.

Missions and special uses. Airships have rarely been used for private travel, though about 1899 the playboy-aeronaut Alberto Santos-Dumont made social calls on friends in Paris by flying in from Saint-Cloud aboard one of his one-man airships. In the pre-World War I days, airship flights were favourite attractions at fairs and expositions, and later nonrigid airships were often used to carry advertising messages. The significant missions performed by airships, however, were predominantly military, with commercial passenger carrying secondary. The first military aircraft acquired by the United States (1905), a rudimentary nonrigid airship of the general configuration described above, proved of little practical use and was soon discarded in favour of the airplane.

During World War I, fleets of French and British blimps (nonrigid) performed yeoman service on convoy and antisubmarine patrol duty over the European approaches. A few military semirigids were tried but proved too vulnerable to be effective. Germany's fleet of rigid Zeppelins, however, provided the world's first aerial terror weapon. A dozen bombing raids successfully penetrated the defenses of London before anti-aircraft guns and the planes of the Royal Air Force rendered the attacks too costly to be continued.

After World War I the United States Navy took a temporary interest in the possibilities of the rigid airship for long-range scouting at sea and as an airplane carrier. The loss of three large ships due to bad weather and structural failures substantially diluted enthusiasm, and, although a number of in-flight launchings and recoveries of small airplanes from airships were successfully carried out, the much greater utility of the seaborne aircraft carrier caused the experiment to be abandoned.

Before 1914, the Zeppelin Company had established the first scheduled commercial airline in the world, with a fleet of five passenger-carrying airships shuttling among principal German cities. This service was discontinued during World War I, but, when restrictions on construction of nonmilitary aircraft were relaxed (1925), work on commercially oriented airships was revived. By August 1929 the passenger-carrying "Graf Zeppelin" had made two transatlantic round trips and a flight around the world carrying 16 paying customers and a load of mail. In 1936 the much larger "Hindenburg," fitted with

Trans-atlantic Zeppelin trips

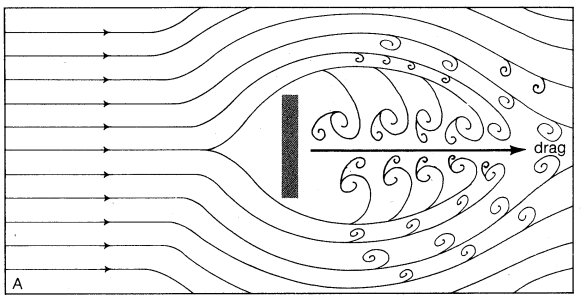
accommodations for 50 passengers (comparable in luxury and convenience with those offered by Atlantic steamships), made ten uneventful round-trip crossings between Germany and the United States. At the end of the first crossing the following year (May 6, 1937), for a reason never fully determined but presumed to be static-electrical discharge, the hydrogen-filled ship exploded and was totally destroyed while making a landing at Lakehurst. This disaster terminated all plans for transatlantic services and, in fact, terminated interest in further airship development anywhere in the world.

II. Heavier-than-air craft

GENERAL CONSIDERATIONS

Aerodynamic principles. A stream of air blowing squarely against a flat plate exerts a dynamic pressure on the plate more or less proportional to its velocity. The faster the wind, the higher the pressure. This phenomenon may be easily demonstrated by holding a hand out of the window of a moving automobile. By inclining the hand with respect to the wind, upward or downward forces are experienced, roughly proportional to the angle of inclination and the speed of the vehicle.

Flat plates (or human hands) are poor aerodynamic shapes. They are not streamlined; that is, air does not flow smoothly around them. Because of this, irregular pressures and vortices (eddies) create a turbulent region behind them (Figure 1A), which absorbs considerable



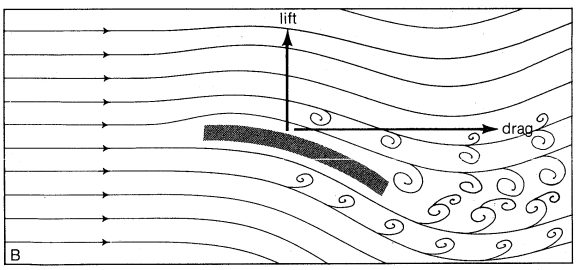


Figure 1: *Effect of airflow against a plate.*
(A) Airflow from left against flat vertical plate creates turbulence and severe drag. (B) Airflow over cambered plate creates drag, but also some lift. By modifying section shape (see Figure 2), the lift-to-drag ratio can be greatly improved.

energy and results in high resistance, or drag. Also, the vertical reactions (lift) are irregular, unstable, and relatively small. The ratio of lift to drag is low.

Wing shape. Lift can be improved by putting a small degree of camber (curvature) in a thin, flat plate presented edgewise to the wind (Figure 1B). The flow of air at small inclinations becomes smoother. Up to a certain angle of attack, the turbulence is small, the lift increases, and the lift-to-drag ratio improves. Beyond the critical angle, however, the smooth flow breaks down, and the lift-to-drag ratio deteriorates rapidly.

In considering composite aerodynamic bodies (*e.g.*, an airplane with body, wings, tail surfaces, and other protuberances), drag comes, first, from the friction arising from motion through the air and the resulting turbulence and, second, from the horizontal component of the lifting force developed by the wing. Total drag is the sum of the two.

Aerodynamic research in wind tunnels early indicated that, by rounding and thickening the forward third of a curved plate and providing a smooth, curved taper to a sharp rear (trailing) edge, lift goes up and drag goes down. The lift-to-drag ratio is considerably increased. Such shapes are called airfoils. Many thousands of variations in thickness and in contour have been tested and the changes in lift and drag with angle of attack recorded. These tabulations provide the design data for airplane wings, propeller blades, helicopter rotor blades, and other streamlined bodies. There is no one best shape for all conditions. Each application requires detailed analysis to select the optimum shape to produce the desired aerodynamic performance.

It was first assumed that lift results from direct air pressure against the undersurface of a wing, as with kites. For this reason, fabric covering was applied to the lower wing surface only, leaving the structural framing exposed on top. Research soon revealed, however, that approximately two-thirds of the upward reaction comes from negative air pressure (suction) over the top of a wing and about one-third from the undersurface (Figure 2). Simple covering over the top surfaces greatly im-

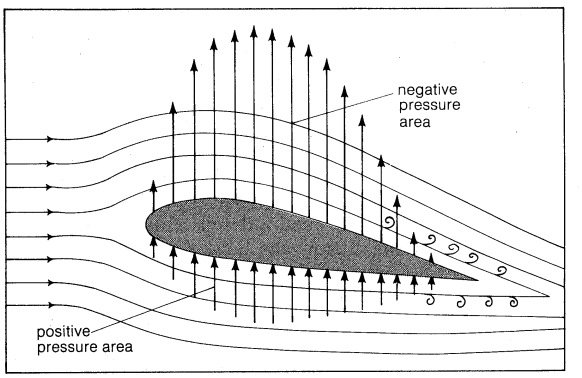


Figure 2: Relative lift developed by the upper and lower surface of an airfoil.

Advantage of monoplanes

proved flight efficiency. Monoplanes have been almost universally adopted over multiplane configurations (*i.e.*, biplanes, triplanes, etc.) because airflow interference between adjacent planes reduces overall efficiency.

The planform of airfoils as aircraft wings also affects their overall characteristics. The pressure distribution with respect to span (length of wing) varies with aspect ratio (Figure 3), defined as the ratio of the span to the chord, the chord being the length of a straight line drawn from the leading to the trailing edge. The pressure distribution also varies with wing taper, with tip shape, and from interference effects between wing and fuselage or from other protuberances or attachments. There are

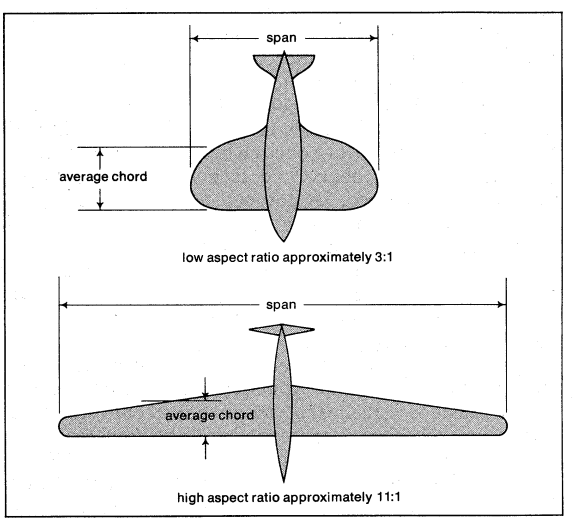


Figure 3: Aspect ratios of wings.

many variables and many variations. Generally speaking, short-span, very tapered, low aspect-ratio (4–5 to 1) wings are associated with fast, highly manoeuvrable aircraft; long-span, high aspect ratios of 12–15 to 1 appear in long-range, relatively slow, high-altitude craft and in very efficient sailplanes.

Counteracting fore-and-aft instability. Most airfoils of interest to aircraft designers are inherently unstable fore and aft; *i.e.*, in pitch. The resultant upward pressures at normal angles of attack act at a centre of lift somewhat behind the centre of gravity of the section. Because these forces are offset from the centre of gravity, a negative moment (turning effect) is produced and tends to force the nose of the aircraft down. To compensate, a stabilizing force must be applied somewhere to create an equal and opposite moment. Canard (duck-like) configurations with a small lifting surface forward of the wing have been tried, but the normal arrangement is to provide a small horizontal surface (stabilizer) well aft of the main wing system (*i.e.*, at the tail of the aircraft) designed so that its lift reaction is downward to counteract the negative pitching moment of the wing (Figure 4). Pitch control is obtained by means of movable flaps (elevators) hinged to the trailing edge of the stabilizer.

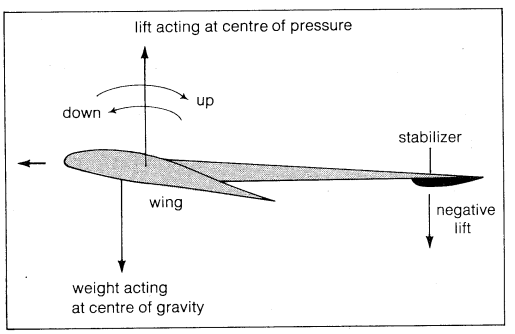


Figure 4: Stabilizer offsetting pitching action of the airfoil.

Tailless designs have been tried in an attempt to reduce overall drag, but they have never been widely adopted. It is possible to compensate for the negative-pitching characteristic by altering wing planform (such as by twist or greatly increased sweepback), but the overall disadvantages generally outweigh the improvement.

Refinements in wing shape. Regardless of wing arrangements or planform, at some maximum angle of attack the smooth flow of air across the upper surface of any airfoil breaks down, and turbulence sets in. In most cases, the transition from smooth to turbulent flow is abrupt. Drag goes up rapidly, and lift is suddenly lost. The wing is said to stall (Figure 5A). (In aerodynamic terms, this phenomenon is in no way related to engine malfunction.) Under such conditions an airplane goes out of control, drops rapidly (usually with a disconcerting spinning motion), and, unless normal airflow over its wings and control surfaces can be restored, will crash. This reaction is what is called tailspin, which resulted in many fatal accidents until its cause was understood and effective recovery techniques applied.

Wing stall

Wing stall can be delayed and flight-speed range extended by the application of wing slots. A narrow spanwise slot of a properly designed contour near the leading edge of an airfoil (Figure 5B) allows a sheet of high-pressure air from the bottom surface to flow up and over the top surface. This upward flow retards flow separation in the boundary layer and delays the onset of turbulence. A greater angle of attack can thus be maintained before stalling occurs and the lift-to-drag ratio increased.

Lift can also be increased by increased airfoil camber, or curvature. Because completely flexible wings (as in birds) are mechanically infeasible, hinged trailing-edge flaps provide a relatively crude but reasonably effective means of camber control (Figure 6A). Lift increases with increasing flap deflection, but so do turbulence and drag. Longitudinal stability also deteriorates.

The most effective arrangement is a combination of leading-edge slots and trailing-edge flaps. By interconnecting the activating mechanisms, it is possible to match varying effective camber with optimum slot openings to maintain relatively high lift over a range of flight speeds. As an added refinement, the flaps themselves may be slotted to smooth out local airflow (Figure 6B). Two points are obvious: (1) the analogy to the arrangement of certain feathers of a bird's wing and (2) the complicated structural and mechanical problems involved in application to aircraft.

Occasionally it is necessary quickly and effectively to reduce wing lift in order to lose altitude rapidly under control without excessive changes in attitude or to prevent aircraft on the ground from being inadvertently lifted off and damaged in high winds. For such purposes so-called spoilers are sometimes installed, either in the form of split trailing-edge flaps (one surface moving upward, the other downward, as in Figure 6C), or as narrow, spanwise slats or flaps that can be raised vertically from the top surface of the wing near the leading edge (Figure 6D) to deliberately generate turbulence and destroy lift.

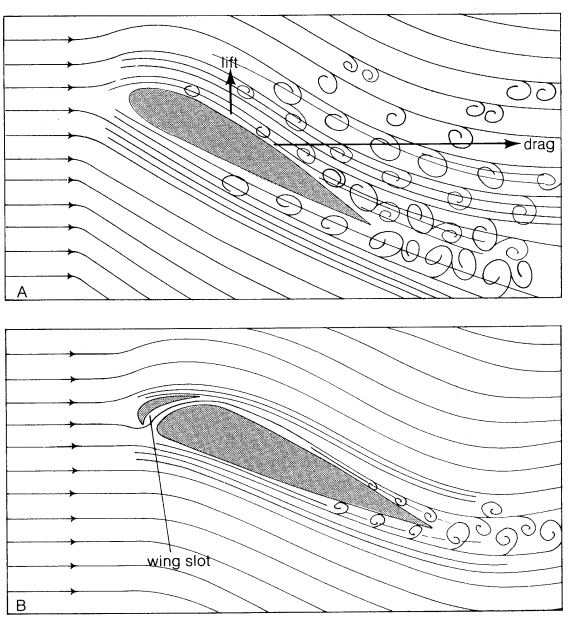
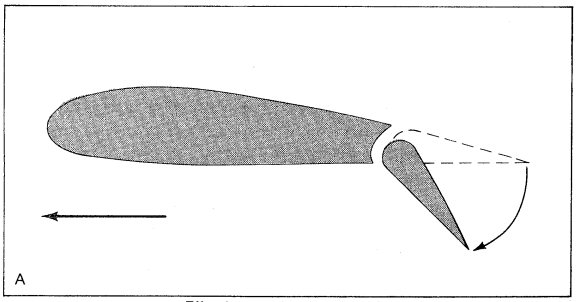


Figure 5: (A) Wing stall at high angle of attack. (B) Wing stall delayed by use of a wing slot.

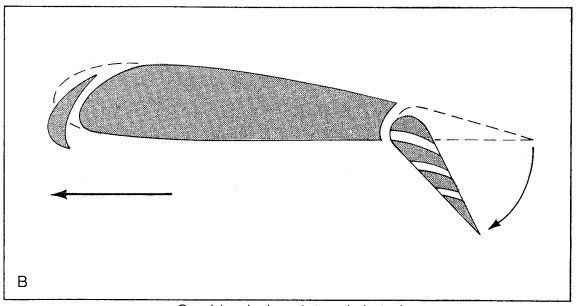
Icing problems. Lift may be inadvertently and, sometimes, irrevocably lost from natural causes. One of the hazards to aircraft flying in any but clear, dry weather arises from loss of lift because of accumulations of snow and ice on wings and control surfaces. Only a small amount may roughen surfaces enough to modify section shape and seriously impair normal flight characteristics. After exposure on the ground in winter weather, wing and tail surfaces are thoroughly cleaned before takeoff. Even a light, powdery coating of snow may be enough to cause trouble.

In flight, an airplane may ice up very rapidly in sleet, rain, or even heavy clouds, especially if its surfaces are already at subfreezing temperatures. Layers of hard ice tend to build up along leading edges of wings and other surfaces to spoil proper aerodynamic performance. At the same time, the weight of the aircraft is materially increased. Many airplanes have gone out of control and have been lost under such conditions.

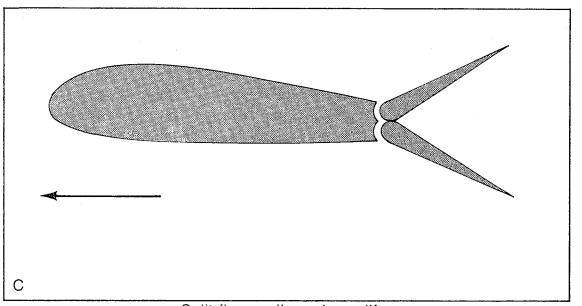
Early attempts to prevent icing by chemical pastes or coatings proved of no value. The more effective method has been provision of an inflatable rubber boot cemented spanwise along leading edges. By periodically inflating and deflating this boot with compressed air, ice accumulations are cracked loose and blow away as rapidly as they form. The system, however, involves considerable weight and mechanical complexity. In most aircraft of



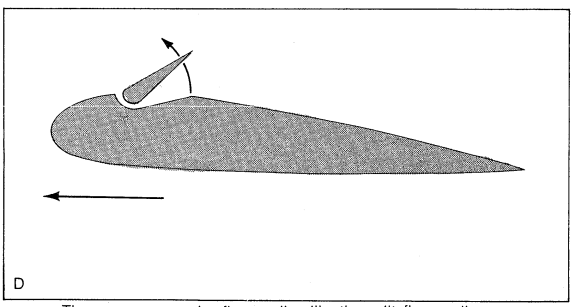
Effective camber increased by use of trailing edge flap.



Combined wing slot and slotted flap for maximum camber effect.



Split-flap spoiler reduces lift and greatly increases drag.



The narrow, spanwise flap spoiler, like the split-flap spoiler, is used to reduce wing lift.

Figure 6: Various flaps and wing slots.

any size intended to operate in all weather conditions, wings are kept ice-free by application of heat. Hot air from engine exhaust heaters or from independent heaters is ducted inside metal leading edges and discharged at wing tips. A relatively small amount of heat applied continuously is sufficient for most flight conditions.

Propulsion systems. All conventional airplanes take off and climb because the propulsive system (engine- or turbine-driven propeller or turbine-generated jet) produces a forward thrust substantially in excess of the combined ground and total air-resistance effects. The resultant forward motion induces airflow over wings, which (because of cross-section shape and inclination) generates upward forces in excess of the total weight of the machine. Once in level, forward flight at uniform speed, the horizontal forces (thrust and drag) are equal and opposite, and the vertical forces (lift and weight) are also in equilibrium.

Equilibrium of forces

Unconventional aircraft, such as short-takeoff and -landing, or vertical-takeoff and -landing vehicles, make partial or complete use of vertical-thrust devices (jets or rotating wings) to get them off the ground quickly into steep or vertical climb.

Once off the ground and at an altitude to clear all obstacles safely, STOL airplanes (and some VTOL types) look like and behave like conventional airplanes. The ability of these aircraft to make steep takeoffs or landings on short runways (*e.g.*, in congested areas) is derived from an ability to deflect downward a part of the airstream from propellers or jets. This deflection is accomplished by the use of slots and flaps or by partial vertical rotation of power plants and even rotation of entire wing systems.

VTOL aircraft rise vertically from rest without use of runways of any kind. This rise has been achieved experimentally by deflecting powerful vertical jets downward by a system of vanes in the jet exhaust or by rotating complete turbine-jet engines to a vertical position. The most successful and mostly widely used VTOL machine, however, is the helicopter, which derives its direct lift (and also its forward propulsion) from a system of rotating wings (see below *Rotary-wing aircraft*).

For information on the history, design, and characteristics of the propulsion systems that are used in aircraft, see the articles FLIGHT, HISTORY OF; GASOLINE ENGINE; JET ENGINE.

All aircraft-propulsion systems, excluding solid or liquid rocket propulsion, depend upon accelerating a mass of ambient air to a velocity greater than the forward or upward speed of the aircraft, then discharging it rearward or downward. By increasing the momentum (mass times velocity) of the airstream, an equal but opposite momentum is applied to the aircraft. The resultant forward force is equal to the mass of air moved, multiplied by its acceleration.

The methods used to accelerate air for propulsion are piston-engine or turbine-driven screw propellers (or helicopter rotors) or gas-turbine-generated jets.

Propellers. The theory and the application of the propeller for marine use was well known before 1900. Because of the great density difference (about 1:800) between air and water, however, an air propeller must move a much larger volume of air to produce a given reaction. This difference means relatively larger diameters and much higher rotational speeds. Two-, three-, and four-bladed propellers are in common use. Most small aircraft use two blades; most large military and civilian multi-engined types use four-bladed propellers. Five-bladed units have been tried in an effort to reduce noise, but results were not sufficiently promising to offset structural and vibrational difficulties.

Multiple-bladed propellers

Each propeller blade is, in fact, an airfoil of high aspect ratio and varying planform. Every cross-section element (strip) produces its own increment of lift and drag, which, when integrated from blade root to tip, result in useful forward thrust plus resistance to rotation. The latter must be overcome by the application of power from an engine.

Because the blade is rotating, the direction and velocity of air meeting each blade element varies widely from hub to tip. For maximum overall efficiency the angle of attack (pitch) of each element must be adjusted to produce the best local lift-to-drag ratio. For this reason, each propeller blade is twisted to make the geometrical pitch approximate the calculated optimum pitch at any point along the blade.

Solid, fixed, pitched aircraft propellers, carved from wood or machined from solid forgings, operate at maximum efficiency only at one specific air speed, air density, rotational speed, and engine power. Departure from the design conditions reduces efficiency and impairs effectiveness. Every airplane flight, however, involves wide deviations from most of the preselected design criteria. As a result, performance is normally compromised. The obvious solution, overlooking the increased mechanical problems involved, is to manufacture hubs and blades separately so that some adjustment is possible after as-

sembly. This accommodation has a further advantage in that high-strength steel can be used in the hub, where stresses are very high, with lighter forged aluminum for blades. The simple, variable-pitch propeller, which is adjustable only on the ground, gives an opportunity to optimize any selected flight situation without a complete propeller change. Many such propellers are in use on small airplanes for personal or general flying.

The next step, which first appeared in the early 1930s, was the two-position propeller, variable in flight under pilot control. A low-pitch setting is available for takeoff and climb with a higher pitch position for cruise and maximum speed. This improved airplane and engine performance, particularly for military and large transport types.

Continuously variable pitch

The ultimate configuration, which in various forms is in universal use on all large and high-performance commercial and military propeller-driven airplanes, is the continuously variable pitch propeller. Automatically or with pilot control, blade pitch can be incrementally changed in flight over a range of some 50 degrees to accommodate varying ambient and operating conditions. The pitch-changing mechanism is either hydraulic or electrical and is controlled electronically. To maintain optimum propeller-rotational speed, an engine-driven governor can be set to maintain the desired number of revolutions per minute of the engine.

Important additional features have been engineered into variable-pitch propellers; first, the ability to feather (change the angle of) a propeller on an engine stopped in flight. By turning blades edgewise to the airstream, windmilling (spinning in the airflow) of a dead engine is eliminated, and propeller drag is greatly reduced. In addition, blades may be completely reversed, from positive to negative angles, to produce negative thrust to help reduce ground roll after landing.

Piston engines and gas turbines. Steam or electric power to drive aircraft propellers has long since been abandoned as impractical. All modern propeller-driven aircraft are powered either with gasoline-fuelled piston engines or by kerosene-fuelled geared turbines. All are air-cooled.

Aircraft piston engines fall into two principal categories, depending on cylinder arrangement: (1) nine-cylinder, single-row or 18-cylinder, double-row radial engines used in business aircraft, commercial transport, and large military transport types; and (2) four- to six-cylinder, horizontally opposed flat types used in thousands of single or twin-engined light planes. In most flat engines and many of the smaller radials the propeller is connected directly to the engine crankshaft. The larger radials generally drive propellers through a train of gears.

All piston engines installed in conventional airplanes are completely cowled (housed in a removable metal covering that completely encloses the engine except for ventilating arrangements). This arrangement not only permits blending (fairing) engines into wing or fuselage contours to reduce overall airplane drag but also provides for internal ducting and baffling to produce optimum flow of air over and around all cylinders for efficient cooling. Adjustable cowl flaps are sometimes provided to expand or to restrict air flow to maintain uniform engine temperatures under varying flight conditions.

Turboprops (*i.e.*, propellers mechanically driven by gas-turbine engines) are in use in a few single-engined airplanes, in a number of two- and four-engined machines ranging from executive transports to medium commercial transports, and in certain large, specialized military aircraft. Their basic component is a gas turbine of the same type used to produce pure jet propulsion (see *Jet propulsion* below), consisting of air compressor, combustion chambers, and turbines. A turboprop engine, however, is designed to absorb most of the energy in the turbine. Because this arrangement provides much more power than is required to run the compressor, the excess is used to drive a propeller through a coaxial shaft and reduction gearing.

Machines of this type have good power-to-weight ratios. They operate most efficiently at low ambient temperatures and at high altitudes. They have been produced in units of more than 6,000 horsepower. For effective speed control, however, they must be fitted with fully automatic variable-pitch propellers. Because of their relatively small diameters, they may be easily shaped (faired) into aircraft structures to limit aerodynamic interference and drag.

Jet propulsion. The most effective way to produce large thrust reactions for high-speed flight is by turbojet engines, which, for a given thrust requirement at cruising speed, are much lighter than any piston-engine–propeller combination. Although, at very high forward speeds, propeller efficiency drops off rapidly, the efficiency of a jet remains high at high speeds.

Engine weight and thrust

Large thrust implies an ability to accelerate a large quantity of air to a high velocity. In a jet engine, ambient air taken in at the engine inlet is compressed 10 to 15 times in a compressor and discharged into a combustion section into which fuel is injected and burned to produce temperatures of the order of 1,400° to 1,900° F (760° to 1,040° C). The hot gases then pass through a multistage turbine, which drives the air compressor through a coaxial shaft, thence into a discharge nozzle at about two to three times the ambient air pressure, producing thrust from the high-velocity stream of hot gas ejected rearward.

To obtain considerably increased thrust for short periods of time, fuel may be injected into the hot jet stream after it leaves the turbine. The use period is short because the rate of consumption of supplementary fuel is very high. This so-called afterburning (reheat in British terminology) feeds on excess oxygen, which comes unconsumed through the combustion section of the turbine. Thrust for takeoff or for speed bursts in flight or for both may be increased by 50 percent or more by this procedure. Jet noise is also tremendously increased.

An intermediate form of engine, between the pure jet and the turboprop, came into use for long range transports in the late 1960s and early 1970s. The bypass, or fan-jet, engine includes a large low-pressure stage (or fan) ahead of the high-pressure compressor inlet. A large volume of low-pressure air is allowed to flow around (bypass) the compressor and turbine and to mix with the jet stream to increase the mass of accelerated air. This arrangement raises efficiency, reduces fuel consumption, and increases cruising range.

Reverse thrust to help decelerate large jet-powered aircraft after landing is obtained by introducing turning vanes or clamshell scoops into (or behind) the jet discharge nozzles to reverse the jet flow. This arrangement produces a negative thrust reaction, which acts against the forward momentum of the plane to reduce the work load on mechanical wheel brakes and to materially shorten the stopping distance after landing.

Deflection of jet flow

Control systems. Airflow over movable control surfaces (elevators, ailerons, flaps, and rudder) supplies the forces necessary to change flight attitudes in conventional aircraft. Elevators and rudder are normally designed as part of the tail assembly (empennage). The former control vertical motions of the plane (climb or descent), and the rudder provides directional control (yaw) as in a surface ship. Ailerons are opposite-acting trailing-edge flaps mounted near wing tips. They control the lateral attitude (roll) of the aircraft. Inboard trailing-edge flaps lowered synchronously greatly increase lift and drag and are used principally to assist takeoff and to reduce speed in landing.

Air speed for a given flight path and control configuration is a function of thrust and is controlled by engine-throttle setting. Maximum available thrust is usually applied during takeoff and climb but is reduced in free flight as drag decreases. Reverse thrust, developed either by changing propeller-blade pitch from positive to negative angles or by introducing reversing vanes into jet-engine discharge nozzles, is used to retard speed on the ground and to shorten roll distance (in conjunction with wheel brakes) after touchdown. For airplanes that land

at very high speeds, small (drogue) parachutes are sometimes deployed from the tail of the fuselage after landing to assist in shortening the roll-out.

Stick and rudder

The pilot's basic equipment (which evolved very early and is still in use in small civilian airplanes and in most single-seat military fighters) consists of a control lever, called a stick, and rudder bar (or pedals) for flight control and ignition switch and throttle for engine control. Fore-and-aft movement of the stick moves elevators up or down to change pitch attitude. Lateral stick movements change aileron positions to produce roll. The stick is mounted on a so-called universal-type joint so that intermediate angular positions provide the coordinated pitch and roll control necessary to make properly banked turns. In larger airplanes (military or civilian types), a handwheel mounted on a movable column (or yoke) replaces the simple stick. Fore-and-aft motion of the column activates elevators as before; rotation of the wheel controls roll. Foot pedals connected to the rudder provide directional control in the air. They may also be interconnected with a steerable nose (or tail) wheel and with wheel brakes to assist in manoeuvring on the ground.

Most nonconventional aircraft (except true helicopters) also employ aerodynamic controls of this character. True helicopters (see below *Rotary-wing aircraft*) obtain their lift, propulsion, stability, and control from the manipulation of blade angles and tilt of their rotor systems. The general appearance and arrangement of the pilot's cockpit is similar to that of a conventional airplane of comparable size. Manipulation of the controls, however, is quite different and requires special training.

As aircraft grew in size, control principles remained the same, but the forces necessary to manipulate them soon exceeded the average pilot's physical capabilities. This situation necessitated the use of some sort of hydraulic or electric servo mechanism to multiply the pilot's effort. This system is inserted between the pilot's control movements and the aerodynamic control surfaces and operates in a way similar to automobile power steering and power brakes. An electronic feedback mechanism is provided to give the pilot the "feel" of flight (*i.e.*, to give the pilot an immediate sense of aircraft response to his intentions).

With the addition of separate controls for landing gear, slots, flaps, spoilers, cabin and wing temperature and pressure, engine and propeller power settings, and fuel management (apart from navigational and communications requirements), the complexity of control systems in large aircraft has long since exceeded the handling capacity of a single person. Three- and four-man flight crews are now required in many cases. Pilot and copilot handle the flight controls while a flight engineer supervises and manages the auxiliary systems. Radio operators and navigators are sometimes added for long flights.

Flight crews

Instrumentation. Aircraft instruments fall into four general groups: (1) those that indicate flight speed, attitude, and altitude; (2) those required to monitor powerplant performance; (3) navigational instruments; and (4) communication-control equipment.

Only minimal instrumentation is provided in gliders and sailplanes. For primary downhill gliding, no instruments are necessary, except possibly a crude airspeed indicator. In sailplanes for cross-country soaring, however, the instrument complement should include a compass, an accurate air-speed meter, a turn and bank indicator, a sensitive altimeter, and a sensitive instrument to show rates of ascent or descent.

The installation of an engine (or engines) of any kind immediately expands the instrument inventory. In addition to the minimal equipment required for flight, other instrumentation, such as fuel-tank gauges, fuel-flow meters, oil-pressure and temperature gauges, engine-speed indicators, and manifold-pressure gauges, appear. For more sophisticated installations, separate cylinder-head temperature gauges, fuel-flow meters, and propeller-pitch indicators are added. For jet engines, thrust and jet-exhaust temperatures must be instrument indicated.

The total number of engine-control instruments increases with the number of engines installed and with the distribution and complexity of the fuel system and of other ancillary equipment on board, such as cabin pressure and temperature, landing-gear position, flap, trim tabs, and spoiler or thrust-reverse positions. The larger and more complicated the aircraft, the greater the number of specialized instruments required.

In addition to the supervision of aircraft and powerplant operations, complicated electronic devices are in everyday use for navigation and communication. Except for casual or sport flying in good weather in the vicinity of an airport, most business, commercial, and military flying is carried out on preplanned routes and fixed time schedules. This system presupposes an ability to identify in-flight positions accurately and to pinpoint destinations, regardless of weather or other unfavourable conditions. Further, it is necessary to land safely at destination whether or not runways may be seen up to the moment of touchdown.

A wide variety of systems and specialized equipment exists to carry out such operations. All are not yet completely reliable, but great progress has been made in improving the safety of all flight operations. The ultimate goal is routine, completely automated flight.

Detailed description of the electronics systems involved is beyond the scope of this discussion. The important point is that the amount and complexity of on-board instrumentation has multiplied tremendously. Care must be taken in the design and construction of aircraft to provide proper space for its use and maintenance.

Flight simulators. One of the most extensively instrumented aeronautical devices and one of the greatest contributors to flight safety never gets off the ground. First employed as an amusement-park device, the flight simulator taught the essentials of air navigation and blind flight to thousands of military pilots during World War II. It has since (in vastly more elaborate form) been widely adopted not only as a means of continually checking out and upgrading the competence of airline and military pilots under all normal and emergency conditions but also as a useful tool for aeronautical engineers to prove out the calculated in-flight behaviour of their aircraft without risk.

Concurrent with specifications for new aircraft destined for important commercial or military use, an order is usually placed for one or more complete flight-deck or cockpit simulators. These reproduce in complete detail all arrangements, instrumentation, and controls exactly as found in the aircraft. Through banks of computers and electronic circuitry, instruments respond to control movements as they would for all conditions of flight. Any desired flight plan can be set up and followed in detail, from engine start, takeoff, predetermined flight pattern, and landing through approach to engine shutdown. During the "flight," radio and visual navigation procedures can be followed and voice communication maintained with "the ground." Also, in many installations, the feel of actual flight is experienced by appropriate sound effects and physical movements; *e.g.*, engine noise corresponding with throttle settings or shaking and vibration as experienced in turbulence of various kinds. Visual landings may be made by watching film projections on a screen through cockpit windows.

Among the important features of simulator training are the ability to set up and to solve emergency situations of any kind that may be encountered in flight, without risk to flight crews or aircraft, and the group training of flight-deck personnel for coordinated action under all foreseeable conditions.

Materials and structures. The earlier type of aircraft construction, consisting of wooden beams, spars, ribs, and struts with fabric covering and tied together with steel wire and cables, virtually disappeared during the 1930s. Major exceptions were the expendable troop-carrying gliders of World War II and current small (generally single-engined) planes for personal and private use. For the latter, however, the primary structures are generally made of welded-steel tubes with plywood or fabric coverings.

The availability of strong, corrosion-resistant alloys of aluminum, magnesium, stainless steel, and titanium, in thin sheets, forgings and complicated extruded or rolled shapes, changed the approach to aircraft structural design. Monocoque construction (*i.e.*, thin shells, shaped and stiffened by internal stringers and ribs) is now universal for wings, nacelles and pods, fuselages, and tail surfaces. By careful design, extremely strong lightweight structures, combining smooth, aerodynamically acceptable contours of widely varying shape with ample internal space for passengers, ancillary machinery, cargo, and fuel, can be readily achieved.

In modern practice, subassemblies and major assemblies are fastened together by such methods as welding, bolting, and riveting, all of predictable strength and holding characteristics. Testing of sample components and assemblies in their final configuration is a routine requirement.

The advent of supersonic flight has imposed temperature as well as strength limitations on materials. The frictional resistance of air at high speeds generates heat that is transmitted to the aircraft structure. Below Mach 2 (twice the speed of sound, or approximately 1,200 miles per hour at medium altitude), the heat rise can be maintained at temperatures below which aluminum alloys begin to lose their strength. Before Mach 3 is reached, however, this temperature becomes critical, and materials with greater heat resistance must be substituted. Certain stainless steels and alloys of titanium and molybdenum qualify in this regard. Though their manufacture and fabrication are difficult and expensive, their use is mandatory for aircraft expected to fly at Mach 3 or beyond.

Materials problems above Mach 2

FIXED-WING AIRCRAFT

The essential components of an airplane are a wing system to sustain it in flight, tail surfaces to stabilize the wing, movable surfaces (ailerons, elevators, and rudders) to control the attitude of the machine in flight, and a power plant to provide the thrust to push the machine through the air. An enclosed body (fuselage) to house the crew, passengers, and cargo must be provided, as well as controls and instruments for the pilots and navigators. Provision also must be made to support the machine when it is at rest on the ground (or on water) and during takeoff and landing.

Many hundreds of thousands of engineering man-hours go into specifications, design, construction, testing, and evaluation of a new airplane. After the basic configuration has been agreed upon, thousands of hours of detailed study must be spent upon the preliminary design before prototype construction can begin. Carefully built, geometrically and dynamically accurate scale models must be tested in wind tunnels to check the flying characteristics and to furnish information on the air-load distribution under different flight conditions. Vibration and flutter sources from fluctuating airflow on wing, body, and tail surfaces must be identified and evaluated.

Because designers are always pushing the state of the art to extreme limits and must work to the highest attainable strength-to-weight ratios in the use of materials, calculations must be exact. Margins for error are small. Rule-of-thumb methods cannot be applied to airplane structural design. Each part must be examined in detail and be designed for maximum efficiency. The problem is more difficult because of uncertainty in determining the magnitude and location of air loads on airplane structures. Often full-scale structural components must be built and tested to destruction to be certain that they will stand up under the dynamic and vibrational loads expected in flight. Such tests may be repeated many times before the final design is accepted and the first prototype built. In this process, modern high-speed computers have proven invaluable. The assembled airframe is then subjected to many complicated load and vibrational tests to be certain that it will meet overall strength specifications.

Testing components and airframes

After the power plants and all other auxiliary equipment are thoroughly tested for functional operation on the ground, the complete machine is ready for flight testing. A long series of well-engineered, well-instrumented test flights is usually required to guarantee that the original specifications for performance and safety have been met. For small, relatively simple machines the test stage may take days or weeks. Large and complicated transports or bombers may require a year or more of shakedown before being released for production.

Configurations. *Wings.* Most modern airplanes are monoplanes. Their lift is derived from one horizontal, fixed-wing system; they are termed high-wing when the wing is attached at the top of the fuselage, midwing when it extends from or near the centre of fuselage section, and low-wing when the wing structure fastens to the fuselage structure at the bottom. Many of the smaller and lighter single-engine airplanes for private-owner use are of the high-wing configuration, as are most modern military aircraft, including bombers and the larger personnel and cargo carriers. Some high-performance fighters tend toward midwing variations. Virtually all multiengined aircraft for executive and commercial transport service are low-winged monoplanes. The multiplane with two or more horizontal wing systems, one above the other, connected by a series of struts and wires, has virtually disappeared. The few biplanes still in service are used mainly for sport flying or for agricultural crop dusting.

Monoplanes may be characterized by the planform (top view) of their wings, which may be rectangular, tapered, swept-back, or delta. A so-called variable-geometry concept incorporating adjustable sweepback is a compromise to obtain good high- and low-speed characteristics from the same wing. For takeoff and landing, hinged wing panels are extended at right angles to the fuselage in order to utilize the advantages of long-span and high-aspect ratio. In flight, for maximum speed and manoeuvrability, the panels are swung rearward to make (with the tail surfaces) a delta configuration. A number of high-performance attack aircraft have been built to this configuration.

Fins and stabilizers. The normal location for stabilizers, fins, rudders, and elevators is well behind the wing, mounted on the tapered tail section of the fuselage. The standard arrangement (until the advent of jet engines) was a cruciform assembly, with vertical fins and rudders and horizontal stabilizer and elevators forming a cross, the axis of which is approximately coincident with the fuselage axis. In some large, high-wing airplanes with relatively short, large-diameter fuselages, airflow over the horizontal tail surfaces was disturbed to the extent that control effectiveness under some flight conditions was compromised. Also, in designs in which jet pods were to be mounted at the after end of the fuselage, it was necessary to relocate the horizontal surfaces to keep them clear of the jet blasts. The solution was the so-called T-tail configuration, in which the horizontal surfaces are mounted at the top of the vertical fin structure.

T-tail configuration

Where three jet engines are used, the fin structure may incorporate the air intake for the centre engine or the entire engine pod. The consequent increase in weight of the tail structure has changed the appearance of many airplanes. To maintain a viable relationship between centre of lift of the wings and centre of gravity of the entire machine, it has been necessary to move the wings aft and to extend the nose section of the fuselage far forward for proper balance.

Engines. For small machines, a single engine and propeller combination is usually mounted in the forward end of the fuselage. When two engines are used, they are normally mounted in nacelles joined into leading edges of wings.

In single turbojet trainers or fighters for military or naval use, the engine is usually mounted inside the fuselage behind the pilot's cockpit, with air intakes either in the nose or as scoops along the sides or in the wing roots. The turbine exhausts through an afterburner into a tailpipe, which extends slightly beyond the after end of the fuselage.

For two- or four-engined planes (either piston engine or turbine–propeller combinations or turbojets), the power

plants are normally distributed outboard along the wings, either in nacelles projecting from wing and edges or in suspended pods. Jet pods are also attached to the rear of the fuselage, as indicated above. A combination of both systems (wing pods and tail pod) may be employed.

Landing gear. Landing-gear configuration depends on whether the aircraft is designed to operate from land airports or from water surfaces or from both. Very few current commercial operations are water based, however, except for those of the so-called bush pilots, who fly small seaplanes (usually twin-float) hundreds of thousands of miles per year from lake to lake in the sub-Arctic regions of Canada and Alaska to keep the miners, hunters, and prospectors transported and supplied. All transcontinental and intercontinental airlines use land-based airplanes.

Float seaplanes and flying boats have been generally abandoned by the navies of the world. Fighters, bombers, and utility transports take off and land on aircraft-carrier decks. Larger transports and cargo planes operate from airport runways ashore. Amphibians, which can land and take off from the water as well as land, have become rare, as have seaplanes and flying boats. All these require considerable strength and bulk to be seaworthy. The hull shape must be modified in a number of ways to permit takeoff capability and on-the-water stability; these modifications add to the drag and militate against aerodynamic efficiency. Corrosion rates are high, and aluminum and magnesium cannot be used as structural materials. Protection against corrosion is difficult, requires constant inspection, and is expensive. Further, suitable airports now exist in virtually every country, and landplanes can reach most of them safely from any point of departure.

Corrosion problems

Pneumatic-tired, wheeled landing gear has been used in most airplanes since the earliest days. Spring-loaded, hydraulically damped shock absorbers take up landing impact loads. For aircraft weighing thousands of pounds these devices become complicated and very heavy. Wheel brakes are generally hydraulically operated.

Until the late 1930s the common arrangement called for a pair of wheels well forward of the airplane centre of gravity with a skid, or tail wheel, aft. The airplane rested on the ground in a distinctly nose-up attitude. During takeoff, as speed increased, the tail was raised by the use of the elevators until the fuselage was rolling in a level position with respect to the ground. When flying speed was reached, the tail was depressed by pulling back on the stick, and the increased lift from the greater angle of attack on the wings lifted the machine off.

In landing the procedure was reversed. From an approach glide, the machine was levelled off a few feet off the ground, then, by use of the elevators (as speed decreased), the machine was stalled and dropped to a three-point landing on wheels and tail skid, the drag of the latter being used to reduce speed and shorten the ground roll after landing. This manoeuvre has always been considered the most difficult to learn. It requires good judgment and long practice. Many small, privately owned, sport and general-aviation types are still designed to perform in this way. In many cases the landing gear is exposed and contributes to the overall drag of the airplane in flight.

In the late 1930s, landing-gear design underwent significant changes. The undercarriage plan was reversed. For all large aircraft and most small ones, the main wheels were moved aft to a position well behind the aircraft centre of gravity and a nose wheel (steerable) installed well forward. The fuselage remains parallel with the ground at rest, during takeoff run, and on landing. The aircraft is never deliberately stalled. This arrangement, originally labelled tricycle landing gear, greatly simplifies ground handling, loading, and fuelling. The term is still applied, although very large aircraft may have double wheels or multiwheeled trucks at each of the three support points.

In another change, wheels and struts were made to retract into the fuselage, nacelles, or wings. This increases weight and mechanical complexity but pays off in greatly improved flight performance.

Landing gear required for carrier-based planes

One special case requires comment; *i.e.*, takeoff and landing aboard aircraft carriers at sea. Although carriers are fast enough so that, under most conditions, a wind speed of 25–30 knots (29–35 miles per hour) can be maintained along the flight deck, modern heavily loaded aircraft do not have enough runway length to take off under their own power nor enough room to decelerate to a stop after touchdown.

For carrier-based planes, landing-gear configurations are similar to those for land-based planes but must be strengthened to take the extra loads imposed by acceleration (from zero to flying speed in some 200 feet [60 metres]) by catapult. Also, fuselages must be modified and reinforced to permit installation of a tail hook to engage arresting gear cables strung crosswise on deck in the landing area.

Another necessary modification for carrier-based airplanes is the folding wing. To make it possible to accommodate a desirable number of machines in limited hangar-deck space and to permit handling on the elevators connecting flight deck and hangar deck, wings must be foldable. They are commonly hinged at each side of the fuselage and folded upward to reduce span. The hinges and operating mechanism add considerable weight and structural complexity.

General aviation. Airplanes used primarily for pleasure or for personal convenience range from small, single-seated, single-engined types for sport flying to twin-engined airplanes capable of carrying eight to ten people in all types of weather along established airways.

Between these two extremes is a wide range of aircraft for the private owner. The majority are single-engined, cabin-type monoplanes, usually for two to four people, including the pilot. They are generally powered with air-cooled reciprocating engines, driving fixed-pitch propellers. A few aircraft have appeared with gas turbines and variable-pitch propellers. Landing gear is simple, strong, and, in some types, retractable in flight. Ease of control and ability to get in and out of small fields are characteristic.

With the availability of good airports in many parts of the world, many commercial firms have acquired and operate their own airplanes or fleets of airplanes. Such machines are used to transport personnel or goods from plant to plant or country to country. These airplanes operate under the same general conditions and operating requirements as airline airplanes.

Specialized forms of airplanes have been developed for certain commercial purposes. One is the crop duster, designed to be flown at very low altitudes over growing crops to distribute insecticides or other chemicals. These are usually single-engine craft of relatively small dimensions, flown by a single pilot. Primary requirements are good control at slow speeds and low altitudes and good manoeuvrability.

Crop dusters and racing planes

Many farmers and cattle raisers in countries in which large land holdings are commonplace use small airplanes for general survey and inspection purposes. These aircraft can also carry light equipment and supplies to outlying farm areas.

During the 1960s closed-course airplane racing was revived. Few machines have been designed especially for such purposes. For the most part, they have been modified private-owner types or modified surplus fighter aircraft from World War II.

Commercial transport. For passenger service, aircraft must provide comfortable seating, proper air-conditioning (at all altitudes), and adequate services and facilities, which may include such extras as in-flight motion pictures. For short-haul service, facilities can be reduced, but all safety services and equipment are retained. Extensive navigation and communications equipment must be supplemented by emergency gear for rapid and safe evacuation of the airplane on the ground or for emergency landings.

Large airplanes for overseas or transcontinental use require considerable crew space. In some cases as many as

15 crew members are carried, including pilots, navigator, radio operator, and cabin attendants. For some short-haul services two pilots and one cabin attendant make up the crew complement.

All commercial transport aircraft operate most efficiently when flying at very high altitudes (25,000–40,000 feet [7,600–12,000 metres]). At such heights oxygen content and atmospheric pressure are far below the limits of human tolerance, and it is necessary to supercharge (pressurize) all personnel spaces to somewhere near sea-level conditions. Most airplanes maintain cabin pressures at the equivalent of an altitude of about 8,000 feet (2,400 metres) above sea level, regardless of flight altitude. All passenger airliners carry emergency oxygen so that, if the pressurizing system should fail, passengers can be supplied until the airplane can descend to a lower altitude.

Cabin pressure

Most passenger airplanes, in addition to passengers' personal luggage, carry a certain amount of cargo, usually mail or freight. Cargo and luggage commonly are carried in compartments at the bottom of the fuselage, easily accessible from the ground through doors and hatches.

In airplanes intended only for cargo carrying, the pilots' compartments are usually of the same design and arrangement as for passenger airplanes, but the interior equipment of the fuselage is reduced to a minimum. Suitable anchor points are provided in the deck and interior walls to accommodate cargo of a variety of shapes and sizes. In modern practice, container units shaped to fit into cargo compartments are preloaded on the ground to avoid time loss in piecemeal loading of aircraft. Containers are loaded and unloaded by mechanical means.

Military aircraft. The requirements for military aircraft call for design criteria differing in many respects from those for civilian aircraft. Such planes are described in more detail in AIRCRAFT, MILITARY. In the present article only their major characteristics are dealt with.

Military aircraft fall into seven general categories: fighters, attack, bombers, reconnaissance, transport, trainers, and special-mission.

Fighters. The primary mission of fighter airplanes (for either air force or navy) is to secure control of essential airspaces. Fighters must be capable of the highest possible performance in order to be able to outmanoeuvre opposing fighters, and they must be armed with specialized weapons capable of destroying enemy aircraft.

A single-seated, high-performance fighter of the early 1970s might weigh as much as and be vastly more complicated than many of the multi-engined bombers of World War II. A modern fighter airplane must be an almost completely automatic weapons system that will detect, identify, close with, and open fire upon an enemy with very little help from the man in the cockpit. The pilot must rely upon electronic devices to find his enemy and upon electronic computers carried in his plane to lock onto his target and fire his guns and rockets at the proper moment.

In appearance a modern supersonic fighter bears a close relationship to a guided missile. The fuselage is usually long, narrow, and pointed, with the pilot's cockpit well forward in the nose. Wings are of short span, generally well-swept-back, frequently of delta configuration. Because of the somewhat peculiar weight distribution in the fuselage, with pilot and armament well forward and the heavy jet engine aft, the wing systems are generally located well to the rear of the fuselage, giving a characteristic long, needle-nosed look. The wings themselves are thin, and their surfaces are smooth and free from protuberances.

Very little space is available for the comfort of the pilot. Every available cubic inch is utilized for the storage of fuel, ammunition, guns, computers, power plant, and miscellaneous equipment.

The pilot's cockpit is heated and pressurized for high-altitude operation, but he must also wear pressurized clothing to enable him to withstand the high accelerations to which he is subject during high-speed manoeuvres, together with a protective helmet. Automatic ejection-type seats that incorporate parachutes are installed to provide emergency means of escape.

Attack airplanes. These aircraft closely resemble fighters but are designed to operate at lower altitudes and to carry rockets or other forms of missiles for use against ground targets. They may also be used as dive bombers. These planes are generally somewhat larger than the high-altitude fighters, with one or more engines mounted either in the fuselage or outboard on the wings. They are sometimes designed for extremely low-level attack, approaching their targets by following the profile of the ground at altitudes of 50 to 100 feet (15 to 30 metres).

Bombers. These aircraft are designed to make high-altitude attacks against ground targets, using nuclear or conventional weapons.

With the advent of powerful jet engines and the introduction of nuclear weapons of relatively small physical dimensions and weight, it became possible to design long-range bombardment aircraft with performance capabilities approaching those of the best fighter types. Mach 2- and Mach 3-speed bombers became theoretically possible. With nuclear-power plants, their range would be almost unlimited. The long-range, high-altitude bomber is the basic offensive weapon of strategic air forces. By refuelling in the air from flying tankers, such aircraft have flown around the world nonstop.

Reconnaissance. The classic method of aerial reconnaissance was high-altitude photography, using techniques that were developed to a high degree during World War II. Later reconnaissance techniques, using radar, infrared, or electronic radiations from ground targets, made it possible to identify military activities over very great distances. Modern high-speed airplanes of almost every type may be fitted with photographic or other sensing apparatus for surveillance of enemy operations. Except for certain specialized operations, the external configuration of the airplane is generally unaffected. Internally, cameras and associated equipment simply replace the guns and ammunition of the ordinary attack or fighter airplanes.

Transport. Military transport airplanes resemble those used on civil airlines, but the basic requirements differ greatly. Military transports may be called upon to operate in any part of the world, in a wide range of climatic conditions, along unmarked routes, and often from poorly prepared fields. The character of the cargo handled also is different from that carried in ordinary commercial services. Passenger accommodations must be readily removable, in whole or in part, to accommodate cargo. For combat-troop-carrier operations personal comfort is a minor consideration.

For general-cargo carrying, the facilities and equipment are similar to those for commercial-cargo airplanes. When airdrop of military equipment (by parachute) is contemplated, airplanes are arranged so that large cargo packages or heavy equipment items (up to and including field artillery and light tanks) may be jettisoned from the rear of the airplane over the designated drop zone. For dropping paratroops, extra doors are generally provided in the rear of transport airplanes so that the parachutists and their equipment can be evacuated from the airplane in the shortest possible time.

Airdrop operations

Very large military transport airplanes are capable of handling a number of trucks, tanks, or other motor vehicles—or a complete intercontinental ballistic missile—in a single load. The cargo compartment of such airplanes may be of the order of 90 feet (27 metres) in length and 15 to 20 feet (5 to 6 metres) in diameter. Access to the cargo compartment is through large, double doors in the nose or near the tail of the airplane. Hoists and ramps are provided for easy loading of large and bulky pieces of equipment.

A wide range of small civil-type airplanes are used by both the ground and air forces. Light, single-engined, two-seated airplanes may be used in battle areas for rapid transport of staff officers. Four- six-, or eight-seated airplanes (single- or twin-engined) are also used for transporting military personnel or equipment.

Special-mission. One highly specialized type of logistical-support airplane is the tanker used to refuel other aircraft in flight. It must be compatible with the airplane it is intended to serve; *i.e.*, it must be capable of flying at a range of speeds and altitudes at which the airplanes to be refuelled fly. Planes to be refuelled approach the tanker from below and from the rear, where trailing tubes or a refuelling hose can be picked up. With the planes flying in close formation, the required amount of fuel is transferred from one airplane to the other. One bomber or as many as three or four fighter airplanes may be refuelled at the same time by a single tanker.

Primary-training airplanes are generally of simplified construction with a minimum of complicated equipment. Transition trainers are considerably more complicated. They are fast and highly manoeuvrable and can be fitted with a variety of equipment found also in combat types. Because the use of high-speed, high-performance military airplanes for training purposes is expensive, difficult, and dangerous, a great deal of pilot and crew training is done on flight simulators.

Any operational airplane, from a fighter to the largest bomber, may be fitted with radio-operated controls and flown pilotless (as a drone) for use as a target by other airplanes using live ammunition or for other purposes, such as reconnaissance. Sometimes drone airplanes have been flown through nuclear-bomb clouds or in proximity to nuclear explosions to register the results of such explosions on the aircraft and on equipment carried by it. More frequently, however, small, high-speed, relatively inexpensive, unmanned airplanes are designed, built, and flown as targets for air-to-air or ground-to-air gunnery or rocket practice. Many of these airplanes are powered by small turbojet engines.

Gliders and sailplanes. The simplest heavier-than-air craft are gliders and sailplanes (soaring craft). The fundamental aerodynamics (lift and drag) and control considerations are the same as for other aircraft, but propulsion systems are lacking. Once airborne, gravity supplies the necessary force component for forward motion.

Launching a glider or a sailplane generally requires assistance. The time-honoured method, used by aviation pioneers at the turn of the century, was to drag the machine to the top of a convenient hill and have it pushed off into space by helpers. In the pre-World War II training camps in Germany and at United States gliding centres, long lengths of rubber shock cord, rigged like a giant slingshot, were used. Later refinements (still in use) include towing by automobile and towing to high altitudes behind a powered airplane. Gliders towed behind powered airplanes were used as troop carriers in World War II. Germany used this technique in the invasion of Crete, and thousands of United States troops were landed by gliders behind the beachheads in Normandy on D-Day.

Gliders in World War II

Primary-training gliders of the type used to indoctrinate flight students before World War II were of simple construction, little more than wings and tail fastened to an open wooden framework that carried a seat for the pilot and the necessary controls. Flight capability was little more than a downhill glide from the launching point to the ground. This type has virtually disappeared.

Secondary types are of more refined design, of lighter construction, with larger wings and airplane-like fuselage and enclosed pilot cockpit (sometimes for two people). Under favourable conditions they have considerable soaring capability and can stay in the air for relatively long periods of time.

Sailplanes reached a high state of development in Europe and in America in the late 1930s, and in the 1950s and 1960s there was a great revival of interest.

Aerodynamic and structural refinements are carried to extreme limits. The highest possible lift-to-drag ratios are sought by airfoil selection, low-profile drag, the smoothest possible surfaces, and extreme aspect ratios (some approximately 15 to 1). Weight is reduced by the most sophisticated design methods, using the lightest available materials: fabric, plywoods, and aluminum alloys. Wing loadings, in terms of weight of aircraft and pilot per square foot of wing area, are extremely low. The result is very flat gliding angles (30 or 40 to 1; that is, 30 or 40 feet of horizontal travel for every foot of vertical travel) and very low sinking speeds. In many places in the world it is possible to find areas where natural updrafts (*e.g.*, in mountains, along sea cliffs, or over hot deserts) rise at a rate faster than the sinking speed of the sailplane. A competent pilot can soar for hours and reach extremely high altitudes by riding such upward air currents. In the vicinity of thunderstorms, updrafts of 150 miles per hour or more occur. Very high altitudes may be reached this way, but the violent turbulence presents a potentially great risk of structural damage to the aircraft. The experiences and observations of sailplane pilots have added materially to meteorological science.

Attaining high altitudes

ROTARY-WING AIRCRAFT

The concept of flight on rotating wings (*i.e.*, obtaining direct lift by means of a horizontal rotor or propeller) dates back to Leonardo da Vinci. About 1840, a British scientist, Sir George Cayley, sketched a twin-rotor, helicoptor-like device in his notebook. Finally, in 1907, Paul Cornu, in France, succeeded in lifting himself a few feet from the ground for a few seconds. Through the 1920s and 1930s many attempts were made; none succeeded. It was not until 1938, in Germany, that the first fully controlled helicopter flights of long duration were made.

The autogiro. The first successful rotary-wing aircraft, the autogiro, appeared in the early 1920s. It was conceived by its inventor, Juan de la Cierva of Spain, as a means of landing an aircraft safely (after the then-all-too-prevalent engine failure) by making the generation of lift independent of aircraft forward speed.

Although somewhat similar in appearance, an autogiro differs from a helicopter in that its rotor system is not power-driven but rotates as a windmill under aerodynamic forces. The rotor system simply replaces the conventional wing as a means of producing lift. Thrust for forward speed is produced by a normal airplane engine and propeller combination, and flying controls are similar to those of fixed-wing aircraft.

An autogiro cannot fly unless its blades are already in motion. It must first taxi on the ground until the rotor is turning at specified speed. In later designs, the engine was used to bring the rotor up to speed, but power was disconnected from the rotor prior to takeoff. Takeoff and climb can be very steep but not vertical, except when heading into a strong wind. Near-vertical landings can be made by throttling down the engine and allowing the machine to settle with the rotor system in free rotation (windmilling).

During the late 1920s and early 1930s autogiros were built in considerable numbers both in Europe and in the United States, but by the 1960s they had virtually disappeared, replaced by the more useful helicopter.

The helicopter. The helicopter differs in many respects from more conventional aircraft; it has a powered rotor, permitting such manoeuvres as vertical flight and essentially stationary hovering.

Principles of operation. The principle of helicopter flight is very simple, but the practice proved considerably more difficult. Satisfactory control while hovering posed the greatest difficulty. As with an airplane propeller, thrust (in this case, lift) is developed by imparting velocity to a body of air flowing through the rotor. The lift is proportional to the change in momentum (mass times velocity) of the air. To limit the power required and to facilitate control in flight, all rotary-wing aircraft employ rotors of large diameter, turning at relatively slow speeds.

Rotor systems are normally composed of two to six blades connected to a single hub. Each blade is, in effect, an airfoil of very high aspect ratio (length to chord; *i.e.*, long and slender), which responds to aerodynamic forces developed by the airflow around it. For a given rate of revolution, however, airflow varies widely in velocity and direction from hub to tip; therefore, the lift and drag vary along each blade for each segment of a revolution. There is no steady-state condition along the blades, as is

normal for an airplane wing in forward flight at a given speed.

When airflow induced by rotation is modified by edgewise translation (straight-line movement) of the rotor system, another complication develops. On the side where a blade is advancing into the airstream, the rate of flow over any blade element is the sum of the rotational and translational velocities. The lift is high. On the opposite side of the rotor, the blade is moving downwind, and its relative air speed is the difference between the two velocities. At some point on the blade the airflow and the resultant lift may be zero. In any case, it is clear that, unless some form of compensation is designed into the system, the total effect is an imbalance of the lift and drag forces acting on opposite blades, which would seriously affect rotor stability.

Imbalance of forces

The common method of compensation for lift variation is to make the blades quite flexible and sometimes to hinge them at the hub. Thus, in a single revolution each blade accommodates itself (within preset mechanical limits) to the loads imposed upon it, flapping up or down as the lift varies and speeding up or slowing down in the plane of rotation in accordance with variations in drag and torque (turning force). Articulation at the hub also relieves the severe bending stresses and stresses resulting from repeated application of loads (fatigue stresses) that normally would occur with rigid fastening. The mechanism is complicated, and its weight offsets some of the gain from the use of lighter primary structure in the blades. The availability of materials of improved fatigue resistance, however, makes feasible rigid hub assemblies that rely on blade flexibility to absorb the stresses.

Power to turn the rotor system may be derived from the aerodynamic reaction of airflow induced by motion through the air, as in an autogiro, or from power applied to the rotor shaft by a piston engine or a geared gas turbine. When power is applied to a helicopter rotor shaft, the turning reaction tends to rotate the vehicle as well as rotor but in an opposite direction. For single-rotor helicopters, an opposing turning moment (rotational force) is produced by a small, vertical tail rotor driven by gears and shafting from the main engine or turbine. By controlling the pitch of the tail rotor blades, thus increasing or decreasing its horizontal thrust, directional control of the vehicle is secured.

A balance in turning forces may also be secured by mounting two lifting rotors driven in opposite direction by the same power plant. A tandem arrangement is frequently used for large transport or cargo types in which long fuselages are needed. A smaller and more compact design utilizes twin rotors intermeshed as in an egg-beater.

Helicopters cannot depend on the usual devices provided for airplanes for stability and control. In hovering and at low forward (or backward) speeds, little or no airflow exists to react with aerodynamic stabilizers, movable control surfaces, etc. Control comes from pitch changes in the rotor system, either collective (in which all blades are adjusted to the same pitch) or cyclic (in which the pitch of each blade varies throughout the 360 degrees of a revolution) or a combination of both. All flight manoeuvres, up, down, forward, backward, sidewise, or in any desired combination, can be accomplished by pitch changes and power setting. Collective pitch is usually controlled by a horizontal lever at the pilot's elbow ("up" to rise; "down" for descent). Cyclic pitch is generally controlled from a stick in front of the pilot, as in a fixed-wing airplane. Movement of the stick adjusts flight attitude and direction, fore and aft, laterally, or in any combination.

Pitch changes

A helicopter lacks the static stability and control damping (the tendency of controls to return to a neutral position) characteristic of fixed-wing aircraft. It is, therefore, inherently more difficult to learn to control. Sophisticated instrumentation has been developed to ease the problem for both the trainee and the experienced pilot.

The design of landing gear for helicopters is relatively simple. Because they touchdown normally at zero forward and vertical velocities, the dynamic loads imposed by airplanes during landings are eliminated. Simple wheels or skids or both, without complicated shock-absorbing devices, are generally all that is needed. Also, because of the relatively high drag of the overall system, there is little to be gained by retracting landing gear in flight.

Configurations and missions. As with other forms of aircraft, the size and shape of helicopters relate to the missions that they are intended to perform. Their range and carrying capacity are of an order of magnitude less than that of present-day airplanes. The smallest are spiderlike devices built to carry one man. They consist of little more than a bare framework on tripod legs, fitted with a minimal seat for the pilot, a bracket to support a small gasoline engine, topped by a short pylon to carry a two-bladed rotor. A short tube attached to the pylon and projecting some six feet (two metres) aft carries the necessary vertical stabilizing rotor. Such machines have little civilian use, except for recreation, and only limited military application.

The most common helicopter type in civilian use is the two-, three-, or four-seat machine with single rotor, fitted with a bubble-like canopy or cabin over the pilot and passenger compartment in the nose. The tail rotor is usually supported on an open structural framework or on a light, tapered monocoque beam. Normal power plants are four- or six-cyclinder, flat, air-cooled gasoline engines, but small turboprop engines have been used in some larger machines. Landing gear is usually a pair of simple ski-like skids.

Because of relatively high first cost and maintenance expense, few helicopters are privately owned. Business and commercial applications for the smaller machines are many, and the list is growing. Because of their manoeuvrability and hovering capabilities they lend themselves well to crop dusting, insect control, and conservation operations. Municipal and highway authorities, police and fire departments, ambulance operators, commercial photographers, and news agencies find increasing use for the two- to four-seat machines. Helicopter platforms are being installed in increasing numbers on or adjacent to municipal buildings, hospitals, and business centres.

Some of the larger single-rotor machines, designed originally for military use, are handling many unusual assignments. Oil companies operating wells far out in coastal waters or deep in inaccessible regions use fleets of helicopters to service and supply the drilling sites. Cross-country pipelines can be easily inspected and serviced by means of helicopters. Power companies transport cable, heavy hardware, even poles and steel towers by helicopter to sites that cannot be reached by road. Heavy machinery can be positioned on the tops of high buildings. The scope of such activities will undoubtedly be greatly extended.

Military helicopters are frequently employed for civilian disaster relief and in support of nonmilitary scientific activities. Thousands of tons of food and medical supplies were flown by helicopter into the devastated areas of East Pakistan (later Bangladesh) following the hurricane and tidal floods in the autumn of 1970.

Passenger operations

For passenger transport in urban areas, particularly to shuttle passengers between airports and city centres, the larger, twin-rotor helicopters are finding acceptance. They are powered by two interconnected turboprop engines. A crew of three and up to 25 passengers and their luggage can be carried in accommodations similar to those available in the smaller commercial planes. These machines operate from small spaces located on high buildings or at convenient waterfront sites and deposit their passengers alongside airline terminals for transfer to transport planes. They also serve to interconnect suburban and downtown centres on regular schedules. Such services have been offered for several years in leading European and U.S. cities.

The largest aerial load carrier is the so-called sky crane, a type in use by the United States and the Soviet Union. The Sikorsky S-64 sky crane was in use by the U.S. Army in Southeast Asia in 1970–71 and by a few civilian equip-

ment contractors. This huge and ungainly machine, with a six-bladed main rotor 72 feet (22 metres) in diameter, powered by two 4,500 horsepower turbines, can lift over 22,000 pounds (about 10,000 kilograms).

EXPERIMENTAL AND RESEARCH AIRCRAFT

Supersonic transport airplanes. Although military aircraft have had supersonic flight capability for many years, and high-altitude reconnaissance aircraft have logged many thousands of hours at about Mach 3, the technical feasibility and economic advantages of the supersonic transport (SST) had not yet been demonstrated in the early 1970s.

First SST flights

Soviet and Anglo-French SST experimental prototypes (Tu-144 and Concorde, respectively) were first flown in 1969–70. They are each Mach 2 aircraft (1,400–1,500 miles per hour), powered by four turbojet engines. Primary structural material in both cases is aluminum alloy. Passenger capacity is about 120 for each. Production models may be in commercial service by the mid-1970s.

Design studies for a U.S. Mach 3 SST have been underway for a decade. An original concept involved use of a variable geometry ("swing") wing that was abandoned because of weight and complexity. The most recent configuration called for a fixed delta wing with conventional tail surfaces. Four 60,000-pound-thrust jet engines provide power. Specifications required titanium alloys and stainless steel for structural use to resist friction-generated heat at Mach 3.

In mid-1971, however, the U.S. program came to an abrupt halt when Congress cut off funding for the project. Problems were not of technical feasibility, but involved economic and environmental considerations; *i.e.*, (a) whether the high investment cost was justifiable, and (b) whether higher noise levels and the possibilities of atmospheric pollution could be tolerated. Such questions have yet to be resolved.

Jet-engine noise

Sonic boom is produced by any aircraft in supersonic flight. It consists of two shock waves some 200–500 feet (60–150 metres) apart. Its strength (as measured in peak air-pressure rise) depends upon aircraft weight, length, speed, and altitude. Supersonic flights over inhabited areas are prohibited in many countries.

Jet-engine noise in the vicinity of large airports has long been cause for complaint. Legal noise limits have already been imposed in many areas. Claims have been made that the larger engines of an SST will greatly exceed allowable limits. SST supporters assert, however, that maximum power would be applied only for the breakthrough from subsonic to supersonic flight and beyond, always at high altitudes; that SSTs will not fly supersonically within miles of its terminal airport. Continuous research in jet-engine design and in aerodynamic configurations of engine installations may yet substantially reduce the inherent exhaust-noise levels.

Pollution. The possibility of pollution of the earth's stratosphere (between 40,000 and 90,000 feet [12,200 and 27,400 metres]), in which SSTs will operate, is of concern. In this region the air density is low, and the air temperature is approximately constant. Because of these factors there is very little mixing effect to induce dissipation of pollutants. There are arguments that water vapour, solid particles, and gases discharged by jet engines may tend to accumulate in the stratosphere and may eventually have injurious physical and chemical effects on the atmospheric environment near the earth's surface. Few data presently exist on the characteristics of the stratosphere, so that the effects of pollutants can be only a matter for speculation.

STOL aircraft and convertiplanes. Two of the more important nonconventional types are the STOL and VTOL aircraft. Many arrangements for achieving vertical flight or short landings and takeoffs have been designed or are under investigation.

The most successful is the helicopter (see above *Rotary-wing aircraft*), whose rotor system produces the necessary lift. To reduce the high-speed vibration and drag of the rotor, stub wings are sometimes added to produce lift at forward speeds, thereby reducing the load on the rotor. The result, in various configurations, is the so-called convertiplane. In hovering, however, the downwash of the rotor produces large down loads on wings and compromises hovering performance.

To avoid this difficulty, several alternates have been tried. These have ranged from tilting the entire aircraft through 90 degrees after vertical takeoff from a tail-sitting position to tilting the power plants or the power plants and wings in combination. The same effects can be obtained by mounting jet engines in a normal, horizontal position and deflecting the jet blast downward to effect vertical thrust for takeoff. When sufficient altitude has been gained, the vanes are retracted, and the aircraft moves in level flight. The most spectacular example in this class is the British Harrier military plane, which can rise vertically from an area little greater than its overall dimensions, then achieve supersonic speed in level flight. The process can be reversed for vertical landing.

All direct-lift machines have certain problems in common. The first is the detrimental effect of the high-energy slipstream or jet striking the ground. Loose material thrown about constitutes a hazard both to the machine and to personnel in the vicinity. Also, because the entire lifting force depends upon the engines while hovering or at low speeds, power failure could prove catastrophic. The greatest possible reliability and adequate emergency backup systems are necessary to ensure the safety of the machine and its occupants.

Because there is little or no flow over aerodynamic surfaces, entirely different types of controls must be utilized while flying at low speeds. Many different systems were under study in the early 1970s, but no single system had demonstrated clear superiority.

Hypersonic aircraft. There is no clear line of demarcation between supersonic and hypersonic aircraft. In most usage, the supersonic range is roughly Mach 1 to Mach 4, hypersonic beyond.

Beginning in 1948, the United States National Advisory Committee for Aeronautics (now NASA) programmed a series of research vehicles (X-1, X-2, etc.) to penetrate what was then called the sound barrier and to explore supersonic flight. By 1952 it had been demonstrated that neither the sound barrier nor the "thermal thicket" (*i.e.*, heating effects from supersonic flight) posed insuperable problems. Sights were then raised and research inaugurated looking toward flight in the Mach 4 to Mach 10 region.

The X-15 airplane, first flown in 1959, was the result. It represents probably the ultimate in performance for conventional aircraft for the foreseeable future. It is conventional in external configuration: a long-bodied, swept-wing monoplane with normal tail surface. It is highly unconventional as to power plant and internal arrangements in that it is fitted with a liquid-fuelled rocket engine delivering a very high thrust for short duration.

To utilize the short thrust cycle (about five minutes) effectively without wasting fuel for takeoff and climb, X-15 was launched by dropping it from a B-52 bomber at altitudes around 40,000 feet (12,200 metres). From there, its rocket engine carried it to altitudes of nearly 70 miles (115 kilometres) and developed speeds of over 4,000 miles (6,400 kilometres) per hour. After its fuel was exhausted, it returned to its base as a glider, landing on a pair of tail skids and a nose wheel. Three of these machines were built and extensively flown. The data obtained contributed to upper-air-research knowledge and helped bridge the gap between atmospheric flight and space flight.

One further step is contemplated—the space shuttle. The objective is to develop a vehicle capable of aerodynamic takeoffs and landings (within the earth's atmosphere) that could go into near-orbital space flight at speeds beyond the present hypersonic range of Mach 4 to 10. It may not be entirely unreasonable to contemplate passenger-carrying flights halfway around the world in such vehicles in well under an hour by the end of the century. For the foreseeable future, however, such vehicles will be used principally for communication with and resupply of orbiting space stations.

BIBLIOGRAPHY. CHARLES DOLLFUS and HENRI BOUCHE, *L'Histoire de l'Aeronautique*, rev. ed. (1942), a comprehensive survey of the development of all forms of aircraft from the prehistoric through the early '30s; *Jane's All the World's Aircraft* (annual), a worldwide updating of progress in aircraft development in all countries, available in most technical libraries; P.H. WILKINSON, *Aircraft Engines of the World* (annual); C.F. TAYLOR, *Aircraft Propulsion* (1962), a review of the evolution of aircraft piston engines; THEODORE VON KARMAN, *Aerodynamics* (1954), the basic principles by the outstanding authority; C.D. PERKINS and R.E. HAGE, *Airplane Performance, Stability and Control* (1949); THEODORE P. WRIGHT, *Articles and Addresses, vol. 1, Aeronautical Engineering* (1961), technical papers and discussions by one of America's great engineers; U.S. NATIONAL AERONAUTICS AND SPACE ADMINISTRATION, *Fifty Years of Aeronautical Research* (1968), a detailed and comprehensive account of aeronautical research and development from 1917 to 1967; E.C. GODDARD and G.E. PENDRAY (eds.), *The Papers of Robert F. Goddard* (1970), a complete record of the notes and experiments of the Father of American Rocketry; E.M. EMME, *A History of Space Flight* (1965), spacecraft development from prehistoric times through Gemini; ROYAL AERONAUTICAL SOCIETY OF GREAT BRITAIN, *The Future of Aeronautics* (1971), a compendium of forecasts by a group of international experts; *Aerospace Industries Association of America: Report* (annual), statistical and technical developments of all forms and usage of aircraft.

(S.P.J.)

Aircraft, Military

In addition to fighters, bombers, and reconnaissance craft, aircraft used for military purposes include numerous specialized types, such as trainers, communications aircraft, transports, and carrier-based aircraft. Lighter-than-air balloons and airships, used extensively up to the end of World War II, have since been discarded. Unpowered gliders, employed for the transport of airborne troops in World War II, have also disappeared from the inventory of modern air forces (see also AIRCRAFT; FLIGHT, HISTORY OF).

EARLY HISTORY

The concept of military aviation is almost as old as human interest in flying. Drawings of the legendary Persian king Keykāvūs show his clutching a bow and arrows as he travels through the air on a throne to which four eagles are harnessed. When the first practical aircraft were produced, in the form of hot-air and hydrogen balloons, in 1783, they were adopted quickly for nonoffensive military duties. In 1793 the French Convention authorized formation of a company of "Aérostiers" for military observation from tethered balloons, an operation that came into being on April 2, 1794. Two months later, the first military reconnaissance from such a balloon was made before the city of Maubeuge. Until the "Aérostiers" were disbanded in 1799, their reports contributed to the success of French armies in many battles and sieges. Similar reconnaissance balloons were used later by other armies, notably by both armies during the American Civil War and by the British in Africa from 1884 to 1901. Large man-lifting kites were evaluated by the British Army in the first decade of the 20th century but were soon abandoned.

Reconnaissance balloons

Napoleon studied the possibility of using large balloons as troop transports for an invasion of England, and in 1849 the Austrians launched primitive unmanned "flying bombs" against Venice. Consisting of small, hot-air balloons made of paper, and each carrying a 30-pound bomb fitted with a time-fuse, these caused little damage and few casualties.

Early 20th century. True military aviation began with perfection of the navigable airship and airplane. The brothers Wilbur and Orville Wright, who made the first powered, sustained, and controlled flights in an airplane on Dec. 17, 1903, believed such an aircraft would be useful mainly for military reconnaissance. They considered that it might even make wars impossible by exposing every movement of opposing armies. When they received the first contract for a military airplane in February 1908, it called for an aircraft capable of carrying two persons at a speed of at least 40 miles per hour for a distance of 125 miles. This first machine, an improved version of their 1903 design, crashed during demonstration flights at Ft. Myer, Virginia, on September 17. A replacement was delivered nine months later as "Airplane No. 1, Heavier-than-air Division, United States aerial fleet."

Subsequently, the U.S. lost the initiative in military aviation. Between 1908 and 1913 it allocated only $430,000 for further buildup of its "aerial fleet." In the same period, France and Germany each spent an estimated $22,000,000 on military and naval aviation, Russia about $12,000,000, and even Belgium spent $2,000,000. The most formidable aircraft of the era were the huge metal-framed airships designed and built in Germany by Graf Ferdinand von Zeppelin. Typical was the L.3, which served with the German Navy at the outbreak of World War I. This craft had a length of 490 feet (150 metres), gas capacity of 795,000 cubic feet, and three 210-horsepower Maybach engines, which gave it a speed of 47 miles per hour (about 76 kilometres per hour) and a ceiling of 6,000 feet (1,800 metres). It could carry five 110-pound (50-kilogram) high-explosive bombs and 20 6½-pound incendiary bombs at a time when most military airplanes were without any form of weapons, being intended only for reconnaissance.

Experiments with airplane armament were made spasmodically after 1910, when August Euler took out a German patent on a machine gun installation. In 1911 a French Nieuport two-seater was fitted with a machine gun, while in England Maj. Robert Brooke-Popham of the Air Battalion, Royal Engineers, armed his Blériot monoplane in the same way. He was promptly ordered to remove the gun.

A new low-recoil machine gun, developed by Col. Isaac Newton Lewis in the U.S., was tested on a Wright Model B biplane. In trials at College Park, Maryland, on June 2, 1912, the gun was fired in the air, and a request was made to the U.S. Army Ordnance Department for ten more guns. The request was refused; Colonel Lewis left for Europe and, in January 1913, formed a company named Armes Automatiques Lewis, at Liège, Belgium, and began manufacture of what was to become a standard aerial weapon, the Lewis gun.

Bombing techniques were evolved simultaneously, with the U.S. again pointing the way. Dummy bombs were dropped on a target in the form of a ship by Glenn Curtiss on June 30, 1910. This test was followed by the dropping of a real bomb and the devising of the first bombsight. In England the Royal Flying Corps (RFC) fitted some of its aircraft with bomb carriers, which consisted of a kind of pipe rack beside the observer's cockpit in which small bombs were retained by a pin. The pin was pulled out over the target by tugging on a string. It was primitive but it worked, and, as the Naval Wing of the RFC became more ambitious, attempts were made, successfully, to drop torpedoes from Short and Sopwith seaplanes, following a little-known earlier air launch of such a weapon in Italy.

This practice suggested a great future for torpedo aircraft, and ships to carry them were already being developed. In 1910–11 a Curtiss biplane had been flown from and onto wooden platforms erected over the decks of two cruisers anchored in Hampton Roads, Virginia, and at San Francisco, California, respectively. In May 1912 a pilot of the Naval Wing, RFC, flew a Short S.27 biplane from HMS "Hibernia" while the ship was steaming at 10½ knots. One further step toward the aircraft carrier of the future came in the following year. The old cruiser "Hermes" was fitted with a short deck from which seaplanes took off on wheeled trolleys that were fitted under their floats and dropped away as the machines became airborne.

Thus, by 1914, fighter, bomber, reconnaissance, and carrier-based aircraft all were evolving. Some had been used in action. The first use of an airplane in war was on Oct. 23, 1911, during the Italo-Turkish War, when an Italian pilot made a one-hour reconnaissance flight over enemy positions between Tripoli and al-'Azīzīyah, in North Africa, on a Blériot XI monoplane. This flight was

First bombing raid

followed nine days later by the first bombing raid, when a pilot dropped four heavy grenades on 'Ayn Zārah and the oasis of Tājūrā. Psychological warfare was tried next, on Jan. 10, 1912, when propaganda leaflets were showered from the air, suggesting that the local Arabs abandon the Turks and surrender. The first reconnaissance photographs of enemy positions were taken on February 24–25, and on April 19 motion pictures were made of an enemy encampment from the military airship P.3.

Other nations were quick to follow the Italian example. When the first Balkan War began in 1912, the Bulgarians dropped finned canisters of explosive by hand on Turkish-held Adrianople from the rear seat of Blériot monoplanes.

WORLD WAR I

At the outbreak of war in August 1914, Germany had the largest air force and probably the best, as many of its 260 aircraft were powered by the highly efficient Mercedes liquid-cooled engines that had made possible a series of altitude and duration records. The RFC had 63 airplanes; its former Naval Wing, now the Royal Naval Air Service (RNAS), had 39 landplanes, 52 seaplanes, and 7 airships. The French Air Force, although smaller than that of its enemy with 156 airplanes, had behind it the world's leading aircraft industry. For the first two years of the war, this industry supplied equipment to the British and other allied forces, as well as to the French Air Force. Every engine in the 73 airplanes taken to France and Belgium by the RFC and RNAS in August 1914 was of French manufacture, and 22 of the airplanes were wholly French. Not until April 1918 did the British air services cease to fly airplanes built in France.

Zeppelins. At the start of the war the German Navy had one Zeppelin airship, and the Army six, plus three converted commercial Zeppelins and three smaller airships. The army made the mistake of sending its Zeppelins over heavily defended areas in daylight, and three were shot down before the end of the first month of the war. Thereafter the German Navy achieved outstanding successes in the North Sea with such aircraft by keeping constant watch on allied shipping movements from the Zeppelins, each of which was reckoned to be worth five or six cruisers. Naval airships also spearheaded German air attacks on Britain. In 51 raids between Jan. 19, 1915, and Aug. 5, 1918, they dropped 5,806 bombs totalling 196.5 tons; 557 people were killed and 1,358 injured. The raids caused sizeable damage, had a considerable effect on British civilian morale, hampered war production, and kept twelve RFC squadrons with 110 airplanes and 2,200 personnel and a large force of anti-aircraft guns and searchlights manned by 12,000 personnel engaged on home defense.

Finest of the Zeppelins was LZ-70, commanded by Fregattenkapitän Peter Strasser, who led the naval airship service throughout its entire period of wartime successes and ultimate defeat. The LZ-70 was 740 feet (225 metres) long, able to fly above 16,000 feet (4,880 metres), and had a range of 7,500 miles (12,070 kilometres), which enabled Strasser to plan an air raid on New York by three such craft. Before this could be attempted, the LZ-70 was shot down, and the big rigid (metal-framed) airship was never again employed as a combat aircraft. Smaller, nonrigid airships were used throughout World War I by the British for antisubmarine patrol, convoy escort, and coastal reconnaissance. This type was revived by the U.S. Navy during World War II and was notably successful in protecting convoys against submarine attack.

Unpowered, captive balloons also were used in large numbers in both world wars. Most were of the elongated Drachen type, with stabilizing tail fins. All of the opposing armies used them for observation in World War I, and the crews were among the first to be provided with parachutes for use if the balloons were punctured or set on fire by enemy action. By World War II, such craft had become so vulnerable that they could be used only in unmanned form as anti-aircraft barrage balloons. Over Britain, by making enemy bombers fly high, they reduced the average number of shells required to destroy an airplane from 20,000 to under 3,000 by February 1941. Later, they brought down 232 German V-1 pilotless "flying bombs" over England, when the missiles flew into their cables; but there is no military service today that continues to use lighter-than-air balloons or airships of any kind.

Reconnaissance. When World War I began, few military leaders believed that heavier-than-air flying machines were useful for anything but reconnaissance (*i.e.*, "aerial cavalry" to keep track of enemy troop movements). So, while the German Zeppelins began raiding cities in Belgium and France from the first days of the war, reconnaissance airplanes began the task that was to remain of primary importance throughout the war. They soon proved their worth. If the German armies that had advanced to the River Marne had not been held, the war might have ended in a few weeks. That they were held was due in no small measure to the complete and accurate information on enemy dispositions furnished by aerial reconnaissance.

Aerial photography

Techniques slowly improved. Photography was added to visual observation by the RFC on Sept. 15, 1914, when photographs were taken of German positions during the battle of the Aisne. Developed and printed on the ground, they were forerunners of a huge photographic map of the entire western front, in thousands of sections, that was kept constantly up-to-date as an essential aid to tactical and strategic combat planning in the latter stages of the war. Radio was also coming into use at this time as a means of passing messages between reconnaissance aircraft and the ground.

Development of fighters. One of the key tasks undertaken by air forces was artillery spotting. This action involved circling endlessly over the battlefield, spotting where the shells from friendly artillery were falling, and correcting the aim of the gunners to ensure economy of effort. The value of this activity, plus that of photoreconnaissance, persuaded the military leaders of both sides that the time had come to deny freedom of movement to enemy air forces. From the start of the war, some reconnaissance crews had carried revolvers or rifles. A few lucky shots had forced opponents down. One RFC pilot even persuaded his opponent to surrender by performing such hazardous manoeuvres immediately above the German aircraft that the pilot was happy to land. Clearly, however, the real answer was to fit a machine gun to aircraft that would have the specific task of fighting other aircraft.

The Vickers company of Britain had pioneered such an airplane in 1913, when they displayed their E.F.B.1 (Experimental Fighting Biplane No. 1), named the "Destroyer," in a London air show. Although typical of its time in general layout, with a pusher engine aft of its crew, it was unique in carrying in its nose a belt-fed Vickers-Maxim machine gun with a field of fire of 60 degrees up and down and left and right. From it was evolved the F.B.5 "Gunbus," which began reaching RFC squadrons in France in February 1915. Its top speed was only 70 miles per hour, and the new 100 horsepower Gnome Monosoupape engine was so unreliable at first that one pilot logged 22 forced landings in 30 flights. Nevertheless, the "Gunbus" proved the potential of the fighting airplane.

The "Gunbus"

It was realized from the start that greater success depended on higher speed. The fastest airplanes of the time were single seaters with a tractor engine (*i.e.*, with the propeller at the nose). The difficulty was to devise a method of firing a machine gun forward from such an airplane without shooting away its own propeller. Some British Bristol Bullets and French Nieuports overcame the problem by having a Lewis gun mounted centrally above the top wing, firing above the propeller. The main drawback of this arrangement was the difficulty of changing the Lewis's 47-round drums of ammunition. An alternative idea was to mount the gun at an angle on the side of the fuselage, so that it fired outward past the propeller. This arrangement called for exceptional skill on the part of the pilot, as the airplane had to be flown crabwise when making an attack.

A far superior idea was the interrupter or gun-synchro-

nizing gear, which enabled bullets to pass between the blades of a spinning propeller. This device had been invented before the war by several people, but the designs were pigeonholed and forgotten. Franz Schneider, chief designer of a German company, came nearest to success. He patented his synchronizing gear on July 15, 1913, and fitted it in due course to a prototype two-seat monoplane. While on its way to the front for combat testing in 1915, however, this aircraft was lost and no replacement was built.

Only in France did the prewar concepts lead to modest success in combat. A machine gun and an interrupter gear had been fitted to a Morane-Saulnier monoplane. This arrangement worked quite well, except that some bullets tended to "hang fire" and hit the propeller. To avoid damage, a steel plate was attached to each propeller blade to deflect bullets that would otherwise have hit it. In the spring of 1915, the prewar French sporting pilot Roland Garros fitted similar deflector plates and a machine gun to a Morane-Saulnier type N monoplane and shot down three German aircraft on April 1, 13, and 18. On the 19th, Garros himself force landed with engine trouble and was taken prisoner. With the secret of the "mystery Morane" revealed, the German authorities ordered their leading manufacturer, Anthony Fokker, to fit deflector plates and guns to his "scouts." Instead, his designers evolved a practical interrupter gear and this device, embodied in the Fokker Eindecker (monoplane) series of single-seat scouts (see Figure 1), gave the German Air Force complete superiority over the Western Front.

Fokker Eindecker

By courtesy of John W.R. Taylor

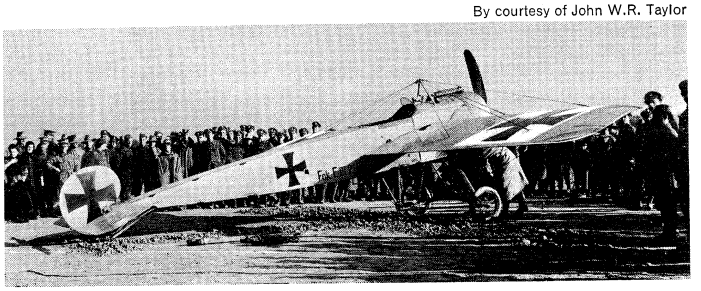


Figure 1: Fokker Eindecker, German fighter plane of World War I.

The operation that became known as the "Fokker Scourge" lasted from October 1915 until May 1916. The British and French air forces were almost shot from the sky by the Fokker Eindeckers, whose synchronized machine guns more than offset low performance (top speed of the E.I was only 82 miles per hour). The British B.E.2c reconnaissance biplanes suffered in particular because their designers had concentrated on giving them inherent stability for optimum observation capability, so that they lacked the manoeuvrability to elude attacking Fokkers. The observer had a machine gun but was so surrounded by struts and wires in the forward cockpit that he could seldom fire it effectively.

The airplanes that ended the Fokkers' period of mastery were new versions of the French Nieuport, with a gun above the top wing, and the British D.H.2 and F.E.2b "pusher" fighters with a nose-mounted gun. These latter machines represented the limit of usefulness of the "pusher" configuration. By mid-1916, the British and French air forces had efficient interrupter gears of their own and thereafter both sides virtually standardized on tractor biplanes and triplanes with one or two forward firing synchronized machine guns, an arrangement that was to remain the normal single-seat fighter armament until the mid-1930s. The allied interrupter gears were of original design.

Sopwith Camel

Many variations of armament were tried experimentally. After standard Sopwith Camel fighters had achieved a reputation as deadly trench-strafers, one aircraft of this type was fitted with two fixed guns, pointing downward through the floor of the fuselage, so that it could rake enemy troops with fire while flying fast and level above their trenches. As early as 1911, Louis Blériot had tested on the ground a gun that fired through the hollow propeller-shaft. The idea was revived by Hispano-Suiza in the form of the *moteur-canon* of 1917. A 37-millimetre Puteaux single-shot shell-firing gun, mounted between the cylinder blocks of a 220-horsepower Hispano engine and firing through the propeller shaft, was used successfully by the French fighter aces Georges Guynemer and René Fonck, but the low muzzle velocity, excessive vibration, and high standard of marksmanship demanded by the *moteur-canon* prevented wider acceptance in World War I. Other large-calibre guns were tried in both fixed and flexible (manually aimed) installations. Large firework-type rockets, carried on the wing struts, and upward-firing machine guns were used on a limited scale, mainly against airships and observation balloons. Canisters of steel darts, known as *fléchettes*, were showered without much success on troops on the ground and airships in flight.

Only two variations of the standard two-gun tractor single-seat fighter formula met with enduring success. The first was the evolution of the fighter-bomber, with small bombs added to fixed guns for ground-attack duties. A typical example was the installation of racks for four 25-pound bombs under the centre fuselage of some Camels. Second was the two-seat fighter, pioneered by another Sopwith design, the 1½-Strutter. In addition to a flexible gun in the rear cockpit, this plane had a fixed forward-firing gun fitted with synchronizing gear. Unaware for a time of the existence of the latter, German pilots made frontal attacks on 1½-Strutters—a hitherto safe direction from which to approach a two-seater—only to be met by fire from the front gun. The two-seat formula was continued even more successfully by the later Bristol Fighter, which became one of the most formidable combat types of 1917–18.

Development of bombing aircraft. The main task of all fighters was to attack enemy reconnaissance and bombing aircraft. The early development of the bomber owed much to Comdr. Charles Rumney Samson, RNAS. Impatient at being limited to a passive reconnaissance role after he had been sent to Belgium with ten aircraft at the outbreak of war, he decided to initiate a strategic bombing offensive with Zeppelins as the primary target. The first air raid into German territory, with four aircraft, on Sept. 22, 1914, was ineffective. The second, made by two single-seat, 80-horsepower Sopwith Tabloids, each carrying a few 20-pound bombs, on October 8, was a spectacular success. One pilot attacked Cologne railway station. Another dropped his bombs on the airship sheds at Düsseldorf, one of which was destroyed, along with the latest Zeppelin, Z.9, inside it.

Handley Page

On November 21, three RNAS pilots attacked the Zeppelin works at Friedrichshafen. Convinced of the potential of strategic bombing, the Admiralty's Air Department commissioned Handley Page, Ltd. to develop "a bloody paralyser of an aeroplane." This plane flew on Dec. 18, 1915, as the prototype H.P. O/100, which was to become the world's first heavy night bomber. Meanwhile, other air forces began building and putting into service strategic day bombers. Among the first were French Voisins. Known as "chicken coops" because of their abundance of struts and wires, these were built in thousands and served throughout the war. The type L, which equipped four artillery observation squadrons in August 1914, had an engine of only 80 horsepower and was used in early 1915 to carry about 130 pounds of small bombs that simply lay in the bottom of the cockpit until the time came for the observer to drop them overboard. Later models had various engines, rated at up to 350 horsepower, and were equipped alternatively as attack aircraft, carrying up to 660 pounds of bombs, or *canon* types with a 37-millimetre gun mounted in the nose. None flew faster than 84 miles per hour; so the Voisins operated mainly under cover of darkness in the last year of the war.

Italy, too, was quick to appreciate the value of heavy and persistent air attacks on enemy targets. Its big three-engined, twin-tailboom Capronis were among the finest bombers of World War I (see Figure 2). Their rear gun-

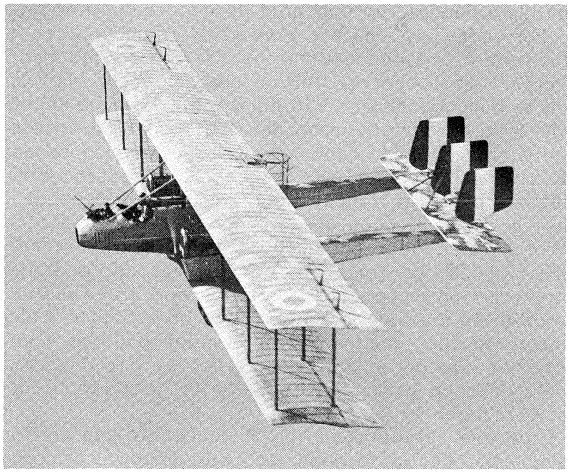


Figure 2: Italian Caproni bomber of World War I.
By courtesy of John W.R. Taylor

ners were some of the war's great heroes, for they often stood on open platforms behind the wings as the aircraft flew slowly over the Alps and back in winter during long raids on targets in Austria. Even larger in size were the Russian Ilya Mourometz bombers of the tsar's Eskadra Vozdukhnykh Korabley (Squadron of Flying Ships). Designed by Igor Sikorsky, now best remembered as the world's outstanding helicopter pioneer, these biplanes spanned about 100 feet and were descended from his "Russky Vityaz" of May 1913, the world's first successful four-engined airplane. About 80 were built, and they made 400 raids on German targets with the loss of only one plane. The best-known German strategic bombers of

The Gotha

World War I were the Gothas, twin-engined "pusher" biplanes, which made several daylight raids on London in formation in the summer of 1917. All large German bombers tended to be called Gothas by the people they attacked. In fact, the raiding forces often included other types. The German Air Force also operated a family of giant four-engined metal bombers known as *Riesenflugzeug*, or R-planes. Most of those used in action were produced by the Staaken and Zeppelin companies. Typical was the Staaken R.VI number R.25, with four 260-horsepower Mercedes engines, which attacked St. Pancras railway station, London, on Feb. 17–18, 1918. The plane spanned 138 feet 5½ inches, was 72 feet 6 inches in length, and had a take-off weight of 25,269 pounds, which included a crew of seven and a bomb load of up to 4,000 pounds.

The Gothas and R-planes, which were inspired by Sikorsky's Ilya Mourometz, took over Germany's strategic bombing offensive when the Zeppelins became too vulnerable to fighter attack. The threat the planes posed, more than the limited destruction they caused, prompted integration of the RFC and RNAS into the Royal Air Force (RAF) in April 1918 and formation of the Independent Force of the RAF in France in that year. Successor to the 41st Wing, which had carried out a nonstop offensive against enemy military centres by day and night in the winter of 1917–18 and the following spring, it was the first independent air force in the world formed specifically for a strategic bombing role. Its four squadrons of Handley Page O/100 and O/400 night bombers, single squadron of F.E.2b night bombers, and four squadrons of D.H.4, D.H.9, and D.H.9A day bombers dropped 550 tons of bombs in the last five months of the war, 390 tons by night. About 40 percent of the total fell on German aerodromes, with such effect that not a single allied airplane was destroyed by enemy bombing during that period.

The achievements of the Independent Force influenced the entire future development of strategic bomber forces. No less important was the example of the D.H.4, which was fast enough to elude most enemy fighters in daylight, thus pioneering a formula that was to produce many years later the British Mosquito of World War II and postwar jet-powered aircraft that relied on speed, rather than guns, for their defense.

Development of naval aviation. Equally significant progress was made in naval flying in World War I, notably by the RNAS. Two distinct categories of combat aircraft emerged. One was the long-range overwater reconnaissance and antisubmarine airplane, operated from coastal bases; the other was the ship-borne airplane, which followed the pattern of land-based air forces by evolving from reconnaissance craft to, eventually, a range of offensive and defensive combat types.

Most efficient of the coastal-based airplanes were large twin-engined flying boats. These had their beginnings in the Curtiss "America," which was being prepared for an

Curtiss "America"

attempted transatlantic flight just before World War I. One of the two designated crew members was an Englishman, John Porte, who rejoined the RNAS at the outbreak of war and persuaded the Admiralty to order a number of flying boats similar to the "America." They were followed successively by larger Curtiss machines of the same general design and then by the "F" flying boats, evolved from them under Porte's leadership at Felixstowe, England. Despite their bulk, these aircraft were sufficiently fast and manoeuvrable to engage enemy airplanes and Zeppelins in combat. The flying boats also pioneered the technique of searching for submarines in a methodical pattern by flying strictly controlled courses shaped like a huge spider web over a patrol area in the North Sea. It took a surfaced submarine ten hours to cross the area, and a single flying boat could search a quarter of the "web" in five hours. The first positive success of a flying boat was logged on May 20, 1917, when a Curtiss "Large America" bombed and sank the submarine U.C.36.

Carrier-based air power also advanced rapidly under the stress of war. In early 1916 the first landplanes (Sopwith Pups) were being flown off the 200-foot (61-metre) decks of primitive carriers that had been converted from merchant ships. The planes still had to alight on the sea and be hauled back on board by crane at the end of a flight until, on Aug. 2, 1917, a pilot landed a Pup on the takeoff deck of HMS "Furious" while the ship was under way. The concept of the true aircraft carrier had been born, and the "Furious" was soon fitted with a rear landing deck in addition to the foredeck.

Much more research and development were required before the modern carrier was perfected. For example, the Pups and 1½-Strutters based on the "Furious" had skid landing gear fitted with hooks that engaged ropes extending fore and aft along the landing deck to bring them to a halt. This primitive device, coupled with the unpredictable air currents flowing around the superstructure of the midship, caused many landing accidents. Nevertheless, it was now possible to plan more formidable naval aircraft, and in October 1918 a squadron of Sopwith Cuckoos, each able to carry an 18-inch torpedo, was

Torpedo aircraft

embarked in HMS "Argus." The war ended before it could go into action; but the RNAS had already used torpedoes dropped from Short seaplanes to sink enemy ships in the Mediterranean, and the Cuckoo, with its modest top speed of 103 miles per hour and endurance of four hours, heralded the eventual demise of the battleship in the face of air-power dominance at sea.

Alternative concepts, such as towing aircraft behind fast warships on lighters from which they could take off when within range of air or surface targets, were soon abandoned. The only adjunct to the carrier that persisted was the idea of carrying one or two defensive fighters or reconnaissance aircraft on board cruisers and battleships. During World War I, such aircraft were stowed on and took off from wooden platforms built over gun turrets, which were turned into the wind for takeoff. After the war, the turret-platforms were replaced by catapults, enabling heavier, faster airplanes to be launched.

Air transport and training. Military air transport did not emerge on any scale in 1914–18, the most notable operations being those in which food and supplies were dropped over a brief period to troops cut off or besieged by enemy forces. The first such operation was mounted

in April 1916, when the British dropped bags of food and other supplies to troops under siege at Kūt al-Imāra (al-Kūt), Mesopotamia. No proper equipment was available for the task, and the bags were simply slung on each side of the aircraft's fuselages or on the landing gear.

The training airplanes of 1914–18 were mainly older types, no longer good enough for combat duty, plus small numbers of operational fighters, bombers, and reconnaissance aircraft. The whole concept of training pilots for military duty, however, was changed by Maj. Robert Smith-Barry, RFC, at the School of Special Flying, Gosport, England. Instead of training his pupils on slow, sedate Farmans, he standardized on the Avro 504J, with the 100-horsepower Gnome Monosoupape engine. Pupil, not instructor, sat in the rear (pilot's) seat, though dual controls were fitted. And the whole essence of instruction was "not to prevent flyers from getting into difficulties or dangers, but to show them how to get out of such situations satisfactorily and, having done so, to make them go and repeat the process alone." For the first time, military pilots flew into action as complete masters of their airplanes, and the sort of flying pioneered at Gosport was adopted eventually at training schools throughout the world.

BETWEEN THE WORLD WARS

In the 21 years between the end of World War I and the start of World War II, the military airplane underwent a complete transformation. The best fighter of 1918 was probably the German Fokker D.VII. Most airplanes of the period were all wood, with fabric covering. The D.VII differed in having a sturdy steel-tube fuselage structure. It was a biplane and, when powered by a 160-horsepower Mercedes engine, had a top speed of 117 miles per hour. Armament was two machine guns. The British Supermarine Spitfire of 20 years later was an all-metal monoplane, with a 1,030-horsepower Rolls-Royce Merlin engine, maximum speed of 355 miles per hour and armament of eight machine guns. Its landing gear was retractable; its pilot sat in an enclosed cockpit, could breathe oxygen at high altitudes, and was in constant contact with his base and other pilots by radio. Even then engineers and scientists were evolving a revolutionary new kind of aeroengine and electronic equipment that would enable pilots to shoot down aircraft they could not see.

Fokker D.VII

The transition took place, understandably, from 1934 to 1939. During World War I France had built 67,982 aircraft, Britain 55,093, Germany 47,637, and Italy about 20,000. Of these, a total of more than 57,000 remained in service with the French, British, and German air forces as the war ended. The United States had relied mainly on its allies for aircraft used in action in France but had produced 15,000 airplanes during the 21 months of its participation in the war and was at last creating a sound aircraft industry.

With so many airplanes available and the war over, no government was willing to spend large sums of money on new military aircraft, and the industry faced a recession in the early 1920s. As an economy measure, which could be tolerated in areas where no air or anti-aircraft opposition was to be expected, there was a concentration on general-purpose aircraft when new types were ordered to replace the 1918 veterans. Typical were the big British twin-engined Vickers Vernons, which could be used for bombing as well as for their intended task of transporting troops or casualties. Equally versatile were the Westland Wapitis, which finally replaced D.H.9As (but used many 9A components) and which were designed for both reconnaissance and light bombing.

Other nations with colonial commitments, such as France and Italy, evolved similar general-purpose types; but the most significant technical advance in the 1920s was the switch to metal structure, still fabric covered, to provide the strength needed to cope with various climates throughout the world and steadily improving engine performance.

Design improvements. When the more drastic changes came, they resulted from research born of competition in sport flying rather than from military requirements. For the international seaplane contests for the coveted Schneider Trophy and for the national Pulitzer races in the 1920s, the Curtiss company produced a remarkable series of high-speed biplanes for the U.S. Army and Navy. Far cleaner in design and more streamlined than any of their predecessors or contemporaries in these events, the CR, R2C, R3C aircraft were built around closely cowled Curtiss liquid-cooled engines developing up to 610 horsepower in the R3C-2. Piloted by Lt. James Doolittle, an R3C-2 won the 1925 Schneider Trophy contest with a speed of 232.5 miles per hour, compared with the winning speed of 145.62 miles per hour achieved in 1922 before the Curtiss seaplanes took part in the event.

Curtiss liquid-cooled engines

During a visit to the U.S., the British manufacturer C.R. (later Sir Richard) Fairey was so impressed by the standard of streamlining offered by the Curtiss D-12 engine that he acquired license rights to build it. Fairey installed this engine in his Fox two-seater, which advanced the speed of RAF day bombers by 50 miles per hour. The Fox was, in fact, faster than any single-seat fighter then in service and had a great influence on military design in Europe. Many of the outstanding fighters and bombers that followed it into service had closely cowled liquid-cooled engines in the tradition of the D-12. The U.S., home of the Curtiss engine, favoured the air-cooled radial engine, offsetting its less streamlined shape by careful cowling design and by the weight saving and reduced vulnerability in combat offered by absence of a cooling system.

In the year that Doolittle won the Schneider Trophy contest, an even more revolutionary design had appeared briefly on the scene. Designed by R.J. Mitchell of Supermarine, the S.4 seaplane was a wooden cantilever monoplane that set an unprecedented standard of streamlining. It crashed because of wing flutter before it could display its full potential in the contest, but from it evolved a series of aircraft that won the contest three times. The Supermarine S.6B, which achieved the last victory, in 1931, gave Britain permanent possession of the trophy; raised the world speed record above 400 miles per hour; and inspired Mitchell to design the Spitfire fighter (see Figure 3).

By courtesy of John W.R. Taylor

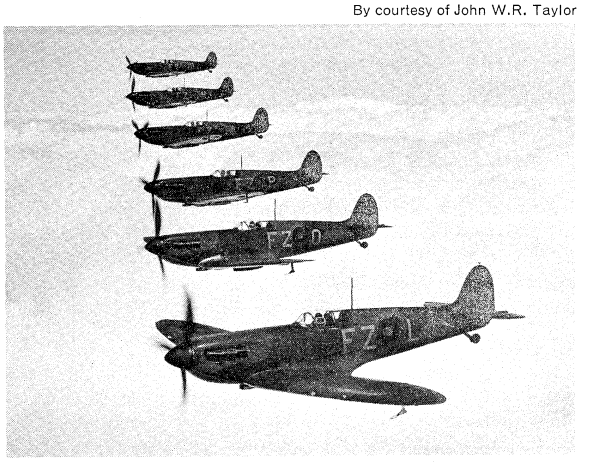


Figure 3: Supermarine Spitfires, British fighter planes of World War II.

The S.5, winner of the trophy in 1927, had come at the period when aircraft were switching from wooden to metal construction. Its wings were wire braced for added strength, but no such external bracing was needed when the Spitfire prototype was built and flown nine years later, on March 5, 1936. Earlier military aircraft had utilized a metal basic structure still covered largely with fabric. The Spitfire was typical of the new high-speed designs of the mid 1930s in having metal covering as well as metal structure, with the skin helping to bear the stress loads and so permitting a reduction in the weight of interior structure. Like the slightly earlier Hawker

Spitfire

Hurricane, which was to be its partner in the Battle of Britain in 1940, the Spitfire also had the then-revolutionary armament of eight 0.303-inch Browning machine guns, mounted in the wings outboard of the propeller so that no interrupter gear was needed.

Bombers. Bombers evolved in parallel with fighters, changing to high-strength metal construction in the late 1920s and to monoplane design, which brought higher speeds, in the early 1930s. In 1921–23, the fiery leader of the U.S. Army Air Service, Gen. William ("Billy") Mitchell, had used formations of Martin MB-2 twin-engined bombers to sink several old German and U.S. battleships at sea. Although the ships were stationary and defenseless, the exercises left no doubt of the dominant part that air power would play in future warfare at sea. Mitchell claimed that they proved the obsolescence of large naval craft and pressed the point so strongly that he was eventually court-martialled and suspended from duty, only to be honoured more than 20 years later, after his death, when the truth of his claims had been borne out by events.

The lessons taught by Mitchell's bombing experiments added to the achievements of World War I bombers, helped to maintain an interest in strategic attack in the interwar years, even during the period when military appropriations were at their lowest. Thus, in 1931, the Boeing Aircraft Company used experience gained with its Monomail transport to produce the B-9 bomber. Predating such fighters as the Spitfire, the B-9 was the progenitor of all modern combat aircraft, with its all-metal cantilever monoplane design, semiretractable undercarriage and variable-pitch propellers—*i.e.*, propellers whose blade angle could be varied. Two 600 horsepower engines gave it a speed of 188 miles per hour, representing a 50 percent improvement over the big Keystone biplanes then in service, without any reduction in bomb load.

Within months of its first flight, the B-9 was overshadowed completely by the Martin B-10 of 1932, which brought the biggest single advance in bomber design since the Handley Page O/400. To the innovations of the B-9 it added enclosed cockpits and an internal weapons bay for its 2,260-pound bomb load. Maximum speed went up to 213 miles per hour, making the B-10 faster than the fighters of its day. Following its success, the U.S. Army ordered prototypes of two gigantic bombers, the four-engined 72-ton Douglas XB-19 and the Boeing XB-15. Neither type went into production, but Boeing built in 1935 a smaller four-engined machine known as the Model 299 that became the prototype of the B-17 Flying Fortress. The concept of this famous bomber was to make possible penetration to any target in daylight by fitting sufficient defensive armament to battle past fighter opposition. It led to development of one of the great combat aircraft of World War II, although combat experience was to prove that formations of day bombers could never protect themselves adequately by their own concentrated cross fire without a fighter escort.

Flying Fortress

Unlike the U.S. Army Air Corps and the RAF, the reformed German Air Force and the Soviet Air Force continued to regard the close support of troops as their primary duty. Even heavy bombers had to be capable of dive bombing, and the Junkers Ju 87 dive bomber became one of the most highly rated German types of the late 1930s. Dive bombing was not new. It had first been tried experimentally with a Sopwith Camel in World War I and was developed into a routine combat tactic by the U.S. Navy and Marine Corps in the 1920s. In the first fleet demonstration, on Oct. 22, 1926, the Curtiss F6C-2 Hawks of Squadron VF-2 made almost vertical simulated dive attacks from 12,000 feet on the heavy ships of the Pacific Fleet as they left San Pedro. Surprise was so complete that ship commanders felt unanimous that such attack would succeed over any defense. Thereafter, dive bombing became a basic U.S. naval aviation tactic. This development was not overlooked by the Imperial Japanese Naval Air Service, which had also been taught the technique of torpedo bombing by a British training mission in 1921–23.

Carrier aircraft. Development of the aircraft carrier progressed rapidly in the interwar years. Most of the essential features had been perfected by the early 1930s, when ship-based aircraft had arrester hooks for picking up cables strung transversely across the deck to bring them to a halt, and folding wings that enabled them to be taken by elevator to below-deck servicing hangars. Because of the additional equipment that naval aircraft had to carry and the strengthened structure necessary for deck landing, they tended to be inferior in performance to land-based types. As radio and navigation equipment became more extensive and important, the Royal Navy also began to favour a crew of two, even for fighters, which imposed further penalties on performance. The Royal Navy also retained a biplane configuration for its Swordfish and Albacore torpedo-bombers into World War II, when all other combat types were monoplanes, believing that the resulting slow-flying capability and manoeuvrability offered the best possibility of penetrating between a screen of escort ships and the capital ships that would be the prime target of an attack.

Other developments. All categories of aircraft benefitted from the technical progress made in the interwar years by engine, equipment, and instrument manufacturers. Superchargers, to boost engine power at high altitudes, were in common use in the 1930s. With crew members now seated in enclosed cockpits and provided with oxygen systems, fighters could maintain full effectiveness at an altitude of 15,000–20,000 feet (4,570–6,100 metres). Combat pilots were taught blind flying for safety and improved operational readiness at night or in bad weather; and first-line aircraft were fitted with blind-flying instrument panels as standard equipment. Variable-pitch propellers, for optimum performance at both take-off and in cruising flight, had been introduced on German Messerschmitt Bf 109 fighters during the Spanish Civil War; they were being produced for RAF Hurricanes and Spitfires in 1939–40. The Bf 109 had also pioneered the use of heavy-calibre cannon armament to ensure the greatest possible destruction against metal-skinned aircraft in the short periods during which the target could be kept in the gun sight at rapidly increasing combat speeds. British fighters were being adapted to take 20-millimetre guns by the time World War II began.

Armour plating behind the pilot's seat and bullet-proof windshields were becoming standard on fighters by 1939. Gun turrets for bomber defense had been pioneered by Machines Motrices in France, and a license-built version of their turret appeared on the British Boulton Paul Overstrand bomber in 1934. This international collaboration and the independent work of the Frazer-Nash Company in the U.K. led to the installation of hydraulically operated multigun turrets on such later RAF fighters as the Defiant and all the aircraft that were used in the wartime British bomber offensive. The U.S. Army Air Force claimed that its highly secret Norden bombsight provided such accuracy that "a bomb could be placed in a pickle barrel from 20,000 ft." Perhaps most important of all, the pilots of RAF Fighter Command were trained to be guided toward their targets by ground radar operators, eliminating the need for wasteful and wearying standing patrols.

Norden bombsight

WORLD WAR II

Bombers. The early air fighting in 1939–40 pointed quickly to the need for further technical improvements. RAF Wellington bombers when intercepted in daylight proved particularly susceptible to fire when their fuel tanks were ruptured. This weakness led to universal adoption of self-sealing fuel tanks on combat aircraft and caused the RAF Bomber Command to switch largely to night operations for the duration of the war. Operational analysis revealed that its heavy and costly operations had achieved only limited effect by the autumn of 1941. Only one aircraft in three was dropping its bombs within five miles of its designated target; over the heavily defended Ruhr this accuracy average fell to one aircraft in ten. This knowledge lent urgency to the development of radio and radar navigation aids, culminating in such equip-

ment as the H2S radar, carried entirely on board the bombers, which could produce maplike pictures of the terrain beneath the aircraft through clouds and in all weather so that identification of targets was always possible. Such equipment, allied to the availability of a new generation of four-engined bombers (the Stirling, Halifax, and Lancaster) and ever-larger bombs, eventually swamped the German defenses by night. Bomber losses were kept to within acceptable limits by their heavy defensive armament, which totalled ten machine-guns in four turrets on the Lancasters. This effectiveness continued even after the German fighters began carrying radar to aid interception.

At the outbreak of war, bombs normally ranged in weight up to 500 and 1,000 pounds (227 and 454 kilograms). The Lancaster was designed to carry bombs of up to 4,000 pounds; it was adapted progressively to accommodate weapons weighing 8,000, 12,000, and finally 22,000 pounds, the largest bombs dropped in World War II.

Another successful RAF bomber was the all-wood Mosquito, designed to avoid the use of metals that were scarce in wartime, and fast enough with a speed of 415 miles per hour in its Mk.XVI version to need no defensive armament. Although small in size, with a crew of two, Mosquitoes had sufficient range to attack Berlin and could carry a 4,000 pound bomb. Variants served as radar-equipped night fighters, as close-support fighter-bombers, and as reconnaissance aircraft.

Day bombers made parallel progress in World War II, notably with the U.S. Army Air Force. As the war continued, the B-17 Flying Fortress and its partner, the four-engined B-24 Liberator (see Figure 4), acquired progressively heavier armament, until the B-17G had no fewer than 13 50-calibre (0.50-inch) guns in chin, nose, dorsal, centre-fuselage, ventral, waist, and tail positions. These bombers also were fitted with turbo-supercharged engines that enabled them to fly at or above a height of 30,000 feet (9,100 metres). Even so, losses were high until such escort fighters as the P-51 Mustang were produced with sufficient range to accompany the bombers throughout their longest raids.

By courtesy of the U.S. Air Force; photograph, from the Air Force Museum, Wright-Patterson Air Force Base, Ohio



Figure 4: U.S. Consolidated B-24 Liberator bombers attacking oil refineries at Ploesti, Romania, May 31, 1944.

Ju 87 dive bomber

Germany's Ju 87 dive bomber (see Figure 5) achieved considerable success in the early stages of the war in Europe but proved vulnerable when matched against the more numerous modern fighters of the RAF in the summer of 1940. Short-range attack in daylight was undertaken mainly by conventional twin-engined bombers with heavy defensive armament from that time, although the fighter-bomber gradually assumed greater importance.

Fighter-bombers and fighters. Among the first types to demonstrate the effectiveness of a small, high-speed strike aircraft flying at ground level was the Soviet Ilyushin IL-2 Stormovik. Heavily armoured for protection against ground fire and defended by a gunner in the rear of the two-seat cabin, the IL-2 could fly at up to 280 miles per hour and was able to attack ground targets with cannon, bombs, and rockets. It was the first close-support type to employ rockets in vast quantities and had a great influence on the adoption of such weapons by other allied air forces. In 1944, the RAF evolved the "cab rank" tactic. Under radio control from the ground, rocket-firing Typhoon fighter-bombers were directed into a line of aircraft one after another (line astern) to attack German armoured columns. This method played a major part in the Normandy campaign.

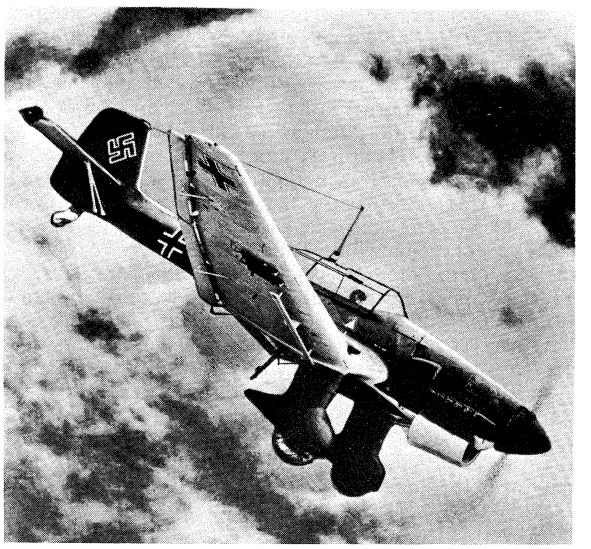


Figure 5: German Junkers Ju 87 dive bomber.
UPI Compix

Another major development in fighter tactics dates from the autumn of 1940, when RAF Blenheims went into action against German night bombers, using on-board radar to locate their targets. Satisfactory results from such attacks came only with the introduction of faster Beaufighter and Mosquito night fighters. The radar itself evolved rapidly from an early version using a number of separate widely spaced aerials to a type that could be wholly enclosed in a small nose "thimble" radome. Interception armament was normally a mix of machine guns and 20-millimetre cannon, with four guns of each type in the Mosquito fighter.

Rockets and missiles. A major advance came with the addition of small unguided air-to-air rockets on German fighters. These rockets proved highly effective but were never used on a large enough scale to affect the course of the war in the air. Similarly, the air-to-air and surface-to-air guided missiles under development in Germany were not perfected in time to play any significant part in the war. Air-to-surface missiles did demonstrate their effectiveness on a number of occasions, notably as anti-shipping weapons. One of their most spectacular early successes occurred when the Italian battleship "Roma" was sunk by direct hits with Fritz X radio-controlled glide bombs launched from Dornier Do 217K-2 bombers on Sept. 9, 1943. Later, Azon radio-controlled bombs dropped by U.S. Liberator squadrons in Southeast Asia proved particularly valuable for pinpoint attacks on Japanese-held bridges.

In general, the German guided missiles were too late to play any major part in World War II, although they had great influence on later military strategy and tactics. Unguided or semiguided V-1 flying bombs and V-2 rockets were used in considerable numbers against London and other targets in 1944–45, each carrying a heavy explosive warhead. Allied to the atomic bomb, which, dropped on two cities in Japan in August 1945, brought World War II to an end, the V-2 foreshadowed the intercontinental ballistic missiles of the postwar era.

Naval aircraft. At sea, air power fulfilled its promise of making the big capital ships obsolete, notably during operations in the Pacific. At Pearl Harbor, on Dec. 7, 1941, Japanese carrier-based dive bombers and torpedo-carrying aircraft struck a heavy blow at the U.S. Pacific Fleet. In later battles, such as the Coral Sea and Midway, U.S. carrier-based aircraft employed similar tactics to inflict crippling blows on Japanese naval forces at sea. Such battles were the first that could be won decisively by one of the fleets engaged without the two surface forces ever coming within sight of each other.

A few of the more successful carrier-based aircraft of World War II, such as the British Swordfish biplane, seemed thoroughly obsolete by comparison with their land-based counterparts. In contrast, naval fighters progressed rapidly to the stage where they were as good as —in some cases better than—contemporary air force types. The performance of the Japanese Mitsubishi A6M, known to the Allies as the "Zero," far overshadowed that of Allied fighters opposing it in the early stages of the war in the Pacific.

Of the 10,937 "Zeroes" built, many were used for kamikaze (suicide) attacks on Allied ships in 1944–45, side-by-side with other combat types and smaller numbers of specially designed suicide aircraft. A peculiarly Japanese tactic, kamikaze operations entailed diving bomb-laden aircraft into targets at sea and were highly effective in terms of both ships sunk and lowering of Allied morale. The 1,228 Japanese naval aircraft expended in kamikaze attacks sank 34 ships and damaged 288 others. Off Okinawa, they inflicted more casualties in the U.S. fleets than were lost as a result of Japanese army action against the invading troops in the long battle ashore.

As the war progressed, maritime aircraft used a variation of the interceptor's search radar to find enemy surface vessels and submarines at sea. Standard attack weapons were bombs, rockets, depth charges, and air-dropped mines; but some aircraft were fitted with large guns, such as the 57-millimetre six-pounder Molins gun in the Mosquito XVIII and the Army 75-millimetre gun carried by the B-25G and H versions of the North American Mitchell. Flying boats were used on a large scale for maritime patrol and antisubmarine duties; but the steady improvement in range of land-based aircraft, coupled with the construction of concrete runways for large landplanes all over the world, initiated the gradual postwar disappearance of combat flying boats from the air forces of most countries except the Soviet Union and Japan.

Flying boats

Reconnaissance. Reconnaissance proved as vital as it had been in World War I. It was impossible to plan a major bombing offensive without thorough and up-to-date coverage of the target area beforehand. It was equally important to discover after each attack precisely how much damage had been done. Sometimes routine sorties brought surprising results, as when a Spitfire photographic sortie revealed that the Germans were using defensive radar in western France in November 1940, at a time when the RAF believed it had a monopoly of such equipment. Later, a chance reconnaissance photograph taken over Peenemünde enabled a sharp-eyed woman photo-interpreter to suggest that the enemy had a flying bomb almost ready for service—a discovery that precipitated a heavy raid on this German rocket research centre.

Many different types of aircraft were used for reconnaissance, but the most successful were probably the high-speed fighters, stripped of armament and all other disposable items to save weight and to increase performance, which raced low to the target areas, took their photos in a few quick passes, and then escaped homeward at maximum speed. High-flying aircraft also took photographs on motion-picture film that could be made up into mosaic maps of vast areas, pointing the way to postwar progress in aerial surveys for map making.

Air transport. Equally significant for postwar civilian operations were the major advances in air transportation made during the war. Mass drops of parachute troops had been pioneered by the Soviet Union in the 1930s, but the German Air Force first used the technique operationally, notably during the invasion of Crete. More than 15,000 of the 22,000 soldiers who captured the island in May 1941 went by air, in 700 transport aircraft and 80 gliders. The price they paid was so high that the Germans never again mounted a major airborne operation, but airborne and parachute troops played a leading role in many subsequent allied assaults. The troop-carrying glider was one of the developments of World War II that had no continuing place in postwar air forces; but the transport airplane was only at the beginning of its useful life. The main transports used in the war were the Douglas DC-3 (C-47) on the allied side and the Junkers Ju 52 on the German side. Their capability was best shown in Burma, where a British Army of 300,000 men was supplied entirely from the air by 23 squadrons of aircraft in the last year of the war against Japan. Their success was such that designers all over the world began work on specialized military freighters with nose- and tail-loading features, roller conveyors in the floor, and other ideas that would speed turnaround on the ground and facilitate the air-dropping of supplies and equipment in flight. None of these airplanes was ready in time to play a significant part in the war but, again, they pointed the way to postwar progress in both military and commercial air freighting.

Jet and rocket propulsion. No wartime development was more significant than the appearance of the first jet-propelled airplanes. Frank Whittle had run the first turbojet, in England, on April 12, 1937, but it was the German Heinkel He 178 that, on August 27, 1939, became the first jet-propelled airplane to fly. Both nations pressed on as rapidly as possible to perfect jet fighters during the war, and the RAF's Gloster Meteor and German Messerschmitt Me 262 were both operational by the summer of 1944. They and the other jet aircraft evolved during the war years played only a small part in the fighting; but their high speed and operating ceiling left little doubt that the days of the piston-engined combat airplane were numbered, and in 1945 a Meteor raised the world air speed record in one jump from 469 to 606 miles per hour.

Heinkel He 178

It is also worth noting that Germany had used operationally a small tailless rocket-powered interceptor, the Me 163, which had a speed of 596 miles per hour. High fuel consumption gave it a powered endurance of only 2½ minutes after its climb to operating height, and several pilots were killed when remaining propellants mixed and exploded during landing. Although ME 163s achieved some success in attacks on allied bomber formations, they remain the only rocket-powered airplanes ever to have achieved operational status.

THE SUPERSONIC AGE

Two years after the end of World War II, on Oct. 14, 1947, Maj. Charles Yeager of the U.S. Air Force became the first person to fly faster than the speed of sound. He made his supersonic flight in a Bell X-1 rocket-powered research airplane, launched in the air from the bomb-bay of a B-29 Superfortress of the kind that had dropped the atomic bombs on Japan. Six months later, on April 26, 1948, the prototype North American XP-86 Sabre jet fighter also flew at supersonic speed in a shallow dive during its development tests, and military aviation entered the supersonic age.

Supersonic fighters. German wartime research had shown that the use of swept-back wings eased the shock-wave problems associated with flying at very high speeds. The Sabre had swept-back wings; so did the Soviet MiG-15 fighter that flew a few weeks after the prototype XP-86, on Dec. 30, 1947. These two aircraft set the pattern for most fighters that followed, although some have adopted a delta-wing layout, while others, notably the F-104 Starfighter, have benefitted from subsequent aerodynamic knowledge to revert to unswept wings. All use a gyro-stabilized gunsight of the type developed during World War II, and their instrumentation and equipment have become steadily more sophisticated. The latest all-weather fighters, in particular, are semi-automatic. After takeoff, they can be flown automatically on preset courses by linking their automatic pilot with a computer-controlled flight system. Radar equipment can search for,

locate, interrogate, and "lock on" to an enemy airplane, fly the fighter toward it by controlling the flight system, fire the armament automatically at the optimum moment, and then break away for the return to base. Self-contained inertial navigation and Doppler-radar systems, allied to other navigation aids, offer high standards of navigational accuracy (see also NAVIGATION). Head-up displays enable the pilot to see all vital information in pictorial form projected on his windscreen without the need to glance down at his instruments during critical periods of flight or landing.

Guided weapons, mostly of the self-homing type, have become standard interceptor armament but are supplemented by heavy-calibre guns and unguided rockets for close-range combat. Support of ground forces remains a major fighter duty, with the weapons of World War II supplemented by air-to-surface missiles and tactical nuclear weapons. Napalm incendiary tanks, utilized in the last years of World War II, have remained a primary fighter weapon in postwar campaigns. A few fighters capable of flying at three times the speed of sound (Mach 3) have appeared, notably the experimental Lockheed YF-12A in the U.S. and the MiG-23 in the Soviet Union. These have to be built of special heat-resistant metals, but almost all other aircraft in service in 1970 continued to have aluminum alloy basic structures that limit maximum speed to about Mach 2.5.

Supersonic bombers. Interceptors have been supplemented rather than replaced by ground-launched missiles; the same is true of bombers. For more than 20 postwar years the jet-propelled heavy bombers of the U.S., U.K., and the U.S.S.R. maintained the deterrent policy of peace through the mutual fear of atomic attack. This role has been taken over largely by intercontinental thermonuclear missiles based in the U.S. and U.S.S.R. and on board submarines (see ROCKETS AND MISSILE SYSTEMS). Nevertheless, both the U.S. and U.S.S.R. retain forces of long-range bombers armed with nuclear missiles and other weapons. These aircraft have more extensive versions of the automatic navigation systems carried by fighters and can also make use of electronic countermeasures (ECM) devices to jam enemy search and fighter/missile-control radar, decoys that give the same signal as a bomber on enemy radar screens, missiles that home on the emissions from defense radar, and automatic landing systems for safe touchdown in fog or bad weather.

F-111 bomber

Other recent military aircraft. Close-support aircraft include such airplanes as the General Dynamics F-111, which is in the class of a light bomber and has a "swing-wing"; *i.e.*, a wing whose geometry can be varied in flight to combine high operational speed with the ability to land and take off comparatively slowly. This development reflects the need for combat aircraft to be as independent as possible of fixed bases in a nuclear age. Most first-line combat types are designed to operate from comparatively short, unprepared strips, and this requirement has led to development of a number of STOL (short takeoff and landing) and VTOL (vertical takeoff and landing) aircraft. Typical of the former is the Swedish Saab-37 Viggen, which uses a tandem-wing layout to combine a Mach 2 performance with the ability to operate from especially prepared sections of highway. The first true VTOL type in service in the 1970s was the Hawker Siddeley Harrier (see Figure 6), which uses a single vectored-thrust turbo-fan, with rotating exhaust nozzles, for both vertical take-off and forward propulsion.

Maritime patrol aircraft, powered increasingly by turbofan engines offering high speed with lower fuel consumption, carry highly complex equipment to detect and track modern deep-diving, nuclear-powered submarines. This equipment includes radar, magnetic anomaly detection (MAD) gear in a tail "sting" to search out metallic objects underwater, sonobuoys that are dropped into the water to listen for and pinpoint submarines, a device that can detect traces of smoke or fumes emitted by a ship or submerged submarine, and a searchlight, as well as a variety of attack weapons and ECM equipment. Reconnaissance aircraft also carry ECM devices and rely heavily on electronic and infrared sensors to supplement their film



Figure 6: Hawker Siddeley Harrier VTOL fighter plane.
By courtesy of Hawker Siddeley Aviation, Ltd.

cameras. Their tasks are to locate and photograph targets, using radar and conventional photographic techniques, and to probe enemy electronic defense systems to discover and evaluate the types of radio and radar equipment that are in use. They do this by offshore patrols, just outside territorial limits and, more rarely, by overflights. The best-known type used for overflights since the war was the Lockheed U-2, which could supplement its long endurance at very high altitude by shutting down its engine and gliding for a time. Successor to the U-2 is the Lockheed SR-71, which can cruise at above Mach 3 at 80,000 feet (24,240 metres) for very long periods. At the other extreme are small pilotless "drones," which can be launched either in the air or from the ground, for reconnaissance flights over enemy territory or in a battle area, carrying photographic, radar, or television sensors. In peacetime, however, most work of this kind is performed by reconnaisance satellites.

Lockheed U-2

Carrier-based aircraft differ little from their land-based counterparts; in some cases, such as the American McDonnell F-4 Phantom II, the same basic types serve both afloat and ashore. The early warning aircraft, a specialized carrier type, has a large radar to detect aircraft or ships approaching the fleet over which it circles. Some airplanes in this category, such as the U.S. Grumman E-2 Hawkeye, can also control interceptor fighters defending the fleet; and this kind of AWACS (airborne warning and control system) airplane has appeared in land-based air forces to detect low-flying enemy raiders and direct interceptors toward them. The first aircraft of this type was a Soviet conversion of the Tu-114 turboprop transport, known to North Atlantic Treaty Organization forces as "Moss."

Modern air forces make use of many other types of aircraft. Transports are often similar to commercial designs, ranging from small STOL liaison machines to such huge transports as the USAF Galaxy, largest airplane in the world, with a payload of 265,000 pounds (120,000 kilograms). Training aircraft are mostly jets, although a few piston-engined aircraft are sometimes used for the most elementary stages of flying instruction. Flying tankers, usually converted from transports or bombers, are used to refuel combat types in flight and so increase their range or endurance.

Helicopters. The helicopter first made its impact during the Korean War (1950–53), when it proved its ability to put down men, guns, and supplies in otherwise inaccessible places. By evacuating casualties speedily, it reduced the death rate to the lowest figure in military history. To these basic duties, the military helicopter has since added the retrieval of damaged aircraft in its flying crane role, antisubmarine search and attack from platforms on virtually any kind of ship, and even close support of ground forces in its "gunship" form, with armament of guns, rockets, grenade throwers, and missiles. No aircraft illustrates better the all-important "go-anywhere" versatility of modern military aviation.

BIBLIOGRAPHY. WALTER RALEIGH and H.A. JONES, *The War in the Air*, 7 vol. (1922–37), is the classic history of air fight-

ing in World War I, centred on the British air services. Similar works treating World War II are DENIS RICHARDS and HILARY ST. G. SAUNDERS, *Royal Air Force, 1939–45*, 3 vol. (1953–54); and W.F. CRAVEN and J.L. CATE (eds.), *The Army Air Forces in World War II*, 7 vol. (1948–58). A detailed history of the Royal Air Force is recorded by JOHN W.R. TAYLOR and P.J.R. MOYES, *Pictorial History of the R.A.F.*, 3 vol. (1968–70). Weapons and equipment used by the British air services in World War II are discussed in M.M. POSTAN, D. HAY, and J.D. SCOTT, *Design and Development of Weapons* (1964). Books describing and illustrating military aeroplanes from 1909 are JOHN W.R. TAYLOR (ed.), *Combat Aircraft of the World* (1969), and *Jane's All the World's Aircraft* (annual). Aircraft used by individual services are treated by F. GORDON SWANBOROUGH and PETER M. BOWERS, *United States Military Aircraft since 1909* (1963), *United States Navy Aircraft since 1911* (1968); OWEN THETFORD, *Aircraft of the Royal Air Force since 1918*, 5th rev. ed. (1971), *British Naval Aircraft since 1912*, 2nd ed. (1962); PETER GRAY and OWEN THETFORD, *German Aircraft of the First World War*, 2nd ed. (1970); and R.J. FRANCILLON, *Japanese Aircraft of the Pacific War* (1970).

(J.W.R.T.)

Air-Cushion Machines

Air-cushion vehicles, sometimes called ground-effect machines, are characterized by their movement, in which a significant portion of their weight is supported by forces arising from air pressures developed around the craft, as a result of which it hovers in close proximity to the earth's surface. It is this proximity to the surface that chiefly distinguishes such craft from aircraft, which derive their lift from aerodynamic forces created by movement through the air.

Two main classes of air-cushion vehicles exist: those that generate their own pressure differential irrespective of forward speed; and those, more closely related to true aircraft, that require forward speed before the pressure differential can be generated. The former are classed as aerostatic craft (ACV's); the latter are called aerodynamic ground-effect machines (GEM's).

This article outlines the development of the air-cushion vehicle since the formation of its basic concept, indicating the wide range of uses to which it has been applied and the methods of construction and operation.

HISTORY

Perhaps the first man to research the air-cushion vehicle concept was Sir John Thornycroft, a British engineer who, in the 1870s, began to build test models to check his theory that drag on a ship's hull could be reduced if the vessel were given a concave bottom in which air could be contained between hull and water. His patent of 1877 emphasized that "provided the air cushion could be carried along under the vehicle" the only power that the cushion would require would be that necessary to replace lost air. Neither Thornycroft nor other inventors in following decades succeeded in solving the cushion-containment problem. In the meantime aviation developed, and pilots early discovered that when they were flying very close to land or water surface their aircraft developed greater lift than in free air. It was soon determined that the greater lift was available because wing and ground together created a "funnel" effect, increasing the air pressure. The amount of additional pressure proved dependent on the design of the wing and its height above ground. The effect was strongest when the height was between one-half and one-third of the average front to rear breadth of the wing (chord).

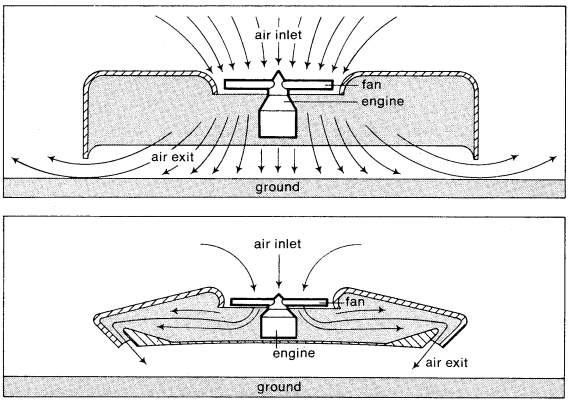
Practical use was made of the ground effect in 1929 by the German Dornier Do X flying boat, which achieved a considerable gain in performance during an Atlantic crossing when it flew close to the sea surface. World War II maritime reconnaissance aircraft also made use of the phenomenon to extend their endurance.

In the 1960s U.S. aerodynamicists developed an experimental craft, making use of a wing in connection with ground effect. Several other proposals of this type were put forward, and a further variation combined the airfoil characteristics of a ground-effect machine with an air-cushion lift system that allowed the craft to develop its own hovering power while stationary, then build up forward speed, gradually transferring the lift component to its airfoil. Though none of these craft got beyond the experimental stage, they were important portents of the future because they suggested means of using the hovering advantage of the air-cushion vehicle and overcoming its theoretical speed limitation of around 200 miles per hour, above which it was difficult to hold the air cushion in place. These vehicles are known as ram-wing craft.

In the early 1950s, engineers in the United Kingdom, United States, and Switzerland were seeking solutions to Sir John Thornycroft's 80-year-old problem. Christopher Cockerell of the U.K. is now acknowledged to have been the father of the Hovercraft, as the air-cushion vehicle is popularly known. During World War II he had been closely connected with the development of radar and other radio aids and had retired into peacetime life as a boatbuilder. Soon he began to concern himself with Thornycroft's problem of reducing the hydrodynamic drag on the hull of a boat with some kind of air lubrication.

Cockerell (since knighted) bypassed Thornycroft's plenum chamber (in effect, an empty box with an open bottom) principle (see illustration), in which air is pumped

By courtesy of T.E. Sweeney



(Top) Plenum chamber. (Bottom) Peripheral jet.

directly into a cavity beneath the vessel, because of the difficulty in containing the cushion. He theorized that if air were instead pumped under the vessel through a narrow slot running entirely around the circumference, the air would flow toward the centre of the vessel, forming an external curtain that would effectively contain the cushion. This system is known as a peripheral jet. Once air has built up below the craft to a pressure equalling the craft weight, incoming air has nowhere to go but outward and experiences a sharp change of velocity on striking the surface. The momentum of the peripheral jet air keeps the cushion pressure and the ground clearance higher than it would be if air were pumped directly into a plenum chamber.

Cockerell's original Hovercraft design

To test his theory, Cockerell set up an apparatus consisting of a blower that fed air into an inverted coffee tin through a hole in the base. The tin was suspended over the weighing pan of a pair of kitchen scales, and air blown into the tin forced the pan down against the mass of a number of weights. In this way the forces involved were roughly measured. By securing a second tin within the first and directing air down through the space between, Cockerell was able to demonstrate that more than three times the number of weights could be raised by this means, compared with the plenum chamber effect of the single can.

Cockerell's first patent was filed on December 12, 1955, and in the following year he formed a company known as Hovercraft Ltd. His early memoranda and reports show a prescient grasp of the problems involved in translating the theory into practice. Such problems still concern designers of Hovercraft more than a decade later, and some of Cockerell's ideas have yet to be fully explored. He forecast, for example, that some kind of sec-

ondary suspension would be required in addition to the air cushion itself. Another of his ideas still to be developed deals with the recirculation of air in the peripheral jet so that part of it is used over and over, improving efficiency and reducing the power required.

Realizing that his discovery would not only make boats go faster but would allow the development of amphibious craft, Cockerell approached the Ministry of Supply, the British Government's defense-equipment procurement authority. The air-cushion vehicle was classified "secret" in November 1956, and a development contract was placed with a British aircraft and seaplane manufacturer. In 1959 the world's first practical ACV was launched. It was called the SR.N1.

Originally the SR.N1 had a total weight of four tons and could carry three men at a maximum speed of 25 knots (one knot equals one nautical mile per hour) over very calm water. Instead of having a completely solid structure to contain the cushion and peripheral jet, it incorporated a six-inch-deep "skirt" of rubberized fabric. This development provided a means whereby the air cushion could easily be contained despite unevenness of the ground or water. It was soon found that the skirt made it possible to revert once again to the plenum chamber as a cushion producer. Use of the skirt brought the problem of making skirts durable enough to withstand the friction wear produced at high speeds through water. It was necessary to develop the design and manufacturing skills that would allow skirts to be made in the optimum shape for aerodynamic efficiency.

Skirts of rubber and plastic mixtures, four feet deep, had been developed by early 1963, and the performance of the SR.N1 had been increased by using them (and incorporating gas-turbine power) to a payload of seven tons and a maximum speed of 50 knots.

The first crossing of the English Channel by the SR.N1 was in 1959, symbolically on the 50th anniversary of Louis Blériot's first flight across the same water. Manufacturers and operators in many parts of the world became interested. Manufacture began in the United States, Japan, Sweden, and France; and in Britain, additional British companies were building craft in the early 1960s.

The sidewall craft

In 1963 the first major variation of the basic air-cushion vehicle theme was produced in the form of sidewall craft. This was a nonamphibious vessel that had a solid hull down each side, with a plenum chamber beneath the hull sealed by flexible skirts at the bow and stern. In the displacement mode, the central hull section floated in the water with the sidewalls well submerged, but when air was pumped into the plenum chamber, the hull was raised above the water and the sidewalls themselves were submerged for only some 12 inches (30 centimetres), considerably reducing the hydrodynamic drag.

The sidewall ACV has several advantages over the amphibious craft, although its use is confined to water: first, water propellers can be used, allowing a much greater freedom of control, especially at low speeds; second, the sidewalls themselves give the craft better stability and reduce the problems that are inherent in all-round flexible skirts. In the early 1970s sidewalls were once again coming into favour, especially among American manufacturers who saw a market for a high-speed marine freight carrier that would not need an amphibious capability.

The years 1962–64 were a peak period for worldwide interest in Hovercraft, but by the early 1970s only the British had produced what could truly be called a range of craft, and this against considerable odds. There were signs, however, that U.S., Soviet, and French manufacturers were seriously contemplating re-entry into the field, and that Australia and Japan also were becoming ACV-minded.

The stagnation of the intervening seven years can be explained by the failure of air-cushion vehicles to live up to what many people thought was their original promise. Cockerell and others had forseen many of the difficulties, but some second-generation designers, builders, and, particularly, operators thought that the simple Hovercraft would be the answer to a variety of problems that at that stage of development were considerably beyond the scope of the craft available.

In the first place, amphibious craft generally needed to be air propelled. Directional control was imprecise, precluding their use on highways. As already mentioned, the design of and materials used in flexible skirts had to be developed from the first, and not until 1965 was an efficient and economic flexible-skirt arrangement evolved, and even then the materials were still being developed.

Problems of engine corrosion

Another major problem arose when aircraft gas-turbine engines were used in a marine environment. Although such engines, suitably modified, had been installed in ships with some success, their transition into Hovercraft brought out their extreme vulnerability to saltwater corrosion. By its very nature the air-cushion vehicle generates a great deal of spray when it is hovering over water, and the spray is drawn into the intakes of gas turbines in amounts not envisaged by the engine designer. Even after considerable filtering, the moisture and salt content is high enough to corrode large modern gas-turbine engines to such an extent that they need a daily wash with pure water and even then have a considerably reduced lifespan between overhauls.

The costs of engine overhauls and skirt maintenance and repairs have probably been the main factors retarding the advancement of air-cushion vehicles. Skirt development has proceeded extremely rapidly in the first decade since SR.N1 and by 1980 will probably no longer constitute a problem. Jet-engine corrosion may be solved by new materials or possibly by intake design to limit spray ingestion. In the meantime, some manufacturers are bypassing the gas-turbine difficulty by using high-speed marine diesel engines in multiple units. These are cheaper, more economical to run, and relatively free from corrosion problems but for a given power output are considerably heavier than their gas-turbine counterparts.

The history of the air-cushion vehicle principle also includes the use of air-cushion support in other applications, both for transportation and for support as such. These include air-cushion transporters, trains, and even beds.

DESIGN, CONSTRUCTION, AND OPERATION

Basic elements of the ACV

The basic elements of an air-cushion vehicle are a hull, beneath which a skirt system is attached and on which accommodation for passengers, crew, and freight is built; a propulsion system; and a lift system that feeds air into the plenum chamber below the craft in order to provide a cushion. The propulsion and lift systems can be driven by the same power plant or by separate units. If a common power plant is used, the system is known as an integrated lift-propulsion system. Some early craft had only one airflow generating system, which was used for both lift and propulsion, but optimum efficiency for both requirements was difficult to achieve simultaneously and separate systems are generally used.

The power-to-weight ratio is as critical at the design stage of an ACV as it is in an aircraft. In the ACV it not only determines the payload and performance of the craft but also the ground clearance between the surface and the skirt. The greater the ground clearance, the more efficiently the propulsion forces available can be used. Theoretical design operating weights are essential for comparison and evaluation purposes, but in practice it has been found that air-cushion vehicles can be overloaded by as much as 100 percent of the design payload and still operate.

To obtain the best power-to-weight-to-strength relationships, structural fabrication of air-cushion vehicles has been based more on aviation than on marine practices. Hull structures are of marine aluminum skin, welded or rivetted onto aluminum webs or frames. The enclosed spaces are usually sealed so that the airtight compartments thus formed provide natural buoyancy. More recent craft have aluminum honeycomb panelling separated by frames to provide the basic buoyancy raft, and considerable areas of glass-fibre structure have also been incorporated.

Early craft had a hole located near the centre of the buoyancy raft through which air was fed to the plenum chamber beneath, but the development of skirt and other techniques led to the ducting of fan air to the edge of the raft, where it was fed downward into the plenum chamber in the manner of a peripheral jet.

Skirt arrangements. Skirts themselves have developed from a simple curtain designed to enclose the cushion into complicated geometrical shapes that contain the cushion, duct the air, and, in some cases, provide a degree of secondary suspension. The simple curtain was quickly replaced by what is now known as a bag skirt. In the shape of a semicircle, this is fastened around the perimeter of the craft; the lower edge is taken inward and upward and is fastened inboard, below the hull. The inflated skirt forms a semicircular cross section. If air is fed through ducts in the top hull so that it inflates the skirt and then is allowed to escape through holes on the inside edge of the bag into the plenum area, the skirt acts as natural ducting, and by varying the size of the holes it is possible to vary the pressure ratio of bag inflation to plenum pressure.

The problem with bag skirts is that the lowest part of the bag quickly wears away and the bag itself tears, allowing air to escape and releasing the cushion pressure. In 1965 it was decided to lengthen the bag skirt by suspending a curtain-type skirt from it. But instead of a straightforward curtain arrangement, the skirt was split into small segments, each of which acted independently from the others. This segmented, or finger-type, addition to the basic bag skirt became the version most commonly used because worn segments could be replaced quickly and economically and because the independent action of each finger allowed the whole skirt to conform much more closely to the operating surface beneath, reducing drag and air-cushion losses.

Materials used in the skirts have varied from the original rubberized fabric, through pure rubber and nylon, to a lamination of nylon and a proprietary plastic known as neoprene. Bondings between the different layers have to be especially strong; otherwise the fabric delaminates under the severe conditions of wear and loses its tear resistance.

Power plants. Power plants used for air-cushion vehicles are generally gas-turbine engines; the output shaft is driven by a turbine that is not mechanically connected to the main compressor-turbine assembly. In this way the engine can be independent of the fan or propeller that it drives, and the free turbine will not begin to rotate until gas from the engine is allowed to pass over its vanes. This allows the craft to remain stationary and on the ground until the driver decides to move, even though the engines are delivering power. The fans used to provide air pressure for lift are usually of the centrifugal type, in which air is fed in through the centre and driven out at considerably higher pressure around the circumference.

Propulsion

Propellers are generally similar to those used for aircraft, although because the air-cushion vehicles travel in the 0–60 knot speed range and can move in reverse, a standard aircraft propeller designed to operate best at higher speeds is inefficient. Hovercraft propellers can be fixed or mounted on swivelling pylons, which allow the craft to be manoeuvred quite accurately, independently of the rudders on which fixed propellers rely. Rudder effectiveness depends to some extent on the forward speed of the craft, and at very low speeds rudders are not efficient as a means of turning.

Other propulsion methods that have been tried in the past include ducted fans, which are quieter than normal propellers but tend to be large and cumbersome. Sidewall craft can be propelled by water screws or by water jets.

ACV operation. Operations on which air-cushion vehicles have been used have been largely confined to commercial passenger-carrying ferry services across stretches of water, varying between three and 25 miles wide, and to certain military operations. Although scheduled services have been run for experimental periods in the United States, Canada, Sweden, and Italy, it is only in Britain and France that such services have survived longer than a season. In the early 1970s, a 170-ton car-carrying craft was so well established on routes across the English Channel that a considerable amount of traffic was being taken from sea ferries, and air services were virtually closed down.

Nonpassenger civil applications have also been found. Craft have been successfully used for seismic survey parties, either over shallow-water areas or in the desert, and in search-and-rescue operations from international airfields at Vancouver, British Columbia, at Auckland, New Zealand, and at San Francisco.

Military uses

Military uses have been more diverse. The main British use has been as a troop carrier during amphibious assaults and as a logistics follow-up craft during the post-assault period. The United States Army has used the Hovercraft successfully in actual operations in Vietnam, both as a patrol craft and as a means of covering the vast area of marsh and paddy field that surrounds the Mekong Delta. Future military uses include mine-countermeasure work, antisubmarine work, aircraft carrying, and missile launching.

The mainstream of Hovercraft development as such has remained in Britain. Although United States firms built experimental craft in the mid-1960s, American interest declined as the pressure of the Vietnam War tended to encourage technologists to improve their established disciplines rather than develop new ones. Other countries also dropped out when the technical difficulties of skirts and the lift-propulsion system became too great for the financial or technical resources available.

The three British companies that pioneered ACV manufacturing merged their ACV interests in the 1960s into one company known as the British Hovercraft Corporation, which in the early 1970s still remained the world's premier manufacturer.

Cockerell's patent and other patents were taken up by a subsidiary of the National Research Development Corporation, a peculiarly British body set up to encourage the funding and backing of inventions. The subsidiary was known as Hovercraft Development Ltd., and, because of the patents that it held, could control the manufacture of skirted air-cushion vehicles in many countries of the world.

The line along which ACV's developed was:

SR.N1 (1959) 3½ to 7 tons, single engine, ducted fan propulsion, speeds between 25 and 50 knots.

SR.N2 (1962) 19 tons, four engines driving two pylon-mounted air propellers; speed about 73 knots.

SR.N3 (1963) 37½ tons, four engines driving two pylon-mounted air propellers; speed about 75 knots.

SR.N5 (1964) 3½ tons, one engine driving a fixed, variable-pitch propeller; speed about 50 knots. Capacity, 18 passengers.

SR.N6 (1965) 4½ tons, one engine driving a fixed, variable-pitch propeller; speed about 60 knots. Capacity, 38 passengers.

SR.N4 (1968) 177 tons, four engines driving four pylon-mounted air propellers; speed 65 knots. Capacity, 30 cars and 254 passengers.

BH.7 (1969) 48 tons, one engine driving a pylon-mounted air propeller: speed about 65 knots. Capacity, 72 passengers and six cars (although the first three craft delivered were purely military versions).

Those somewhat bare statistics suggest the burst of creative energy in the Hovercraft field in the early 1960s and the subsequent slowing down and rationalization of craft into practical machines. In the period between 1960 and 1964, some six other companies in the world also were building their own prototypes or large-scale models, but because of technology problems many of these were never followed up by production models.

The exceptions were in the Soviet Union and France. Little is known about Soviet ACV development, except that after some five years of trials and experimentation large passenger-carrying craft were being operated on the Volga River in 1969 and that military craft were being tested in the Black Sea in 1970.

A French company, formed in 1965, built two amphibious craft that, carrying up to 90 passengers, operated a commercial service based at Nice in 1969. The French

designs are basically the same as any other amphibious craft with the major exception of the skirts, which instead of forming one large plenum chamber are grouped together in a series of "mini-skirts" side-by-side along the length of the craft. Compartmenting the cushion in this way is said to improve stability and directional control. In the early 1970s it was announced that a larger craft, based on similar principles and carrying 32 cars and 260 passengers, would be put into production.

Non-amphibious Hovercraft

The development of nonamphibious Hovercraft along the sidewall principle began in 1962. For various reasons the manufacturer halted production, but the idea of a fast, essentially marine-oriented craft continued to appeal to designers, and a company known as Hovermarine was formed in 1965 to build a smaller sidewall craft, some 12 of which were sold. This is the HM.2, which carries about 65 passengers and is designed for short and medium ferry routes. Mechanical and skirt-design problems caused difficulties that led to liquidation of the parent British company, but in 1970 an American company took over the HM.2 and it appeared that its future would be assured. The basic advantage of sidewall craft is that since they are purely marine, it is possible to equip them with marine propellers and operate them in the same way as high-speed boats. Nevertheless, because a great part of their weight when under way is supported by an air cushion, they can be classified as air-cushion machines.

It is along these lines (sidewall craft propelled by marine methods) that two United States manufacturers have designed much larger craft under government contracts. The first versions are 100-ton test craft, which will be used to evaluate the potential of multithousand-ton surface effect ships. The two models are somewhat similar in appearance; both are powered by six gas turbines. One craft uses propellers, the other water-jet propulsion. Another United States military application, an amphibious assault landing craft for the navy, offers potential uses in commercial operations.

OTHER APPLICATIONS

Air-cushion trains. Once air-cushion suspension was proved practicable in Hovercraft, the system was quickly applied to other forms of transport, and it soon became clear that a tracked vehicle, similar to a train or monorail, would benefit considerably from the lack of friction inherent in an air-cushion system. A French company was the first in the world to produce a practical device, and a later version of its machine was due to be in service as a high-speed link between Orléans and Paris by the mid-1970s. The system uses air-cushion pads above and at the side of a single concrete track to support the "aerotrain," while propulsion is via a large ducted fan mounted at the rear.

In Britain tracked air-cushion vehicle development is also under way, with a construction of a "Hovertrain," propelled by a relatively silent linear induction motor that has no moving parts and picks up current as it moves along the track.

Research is also proceeding in other countries. Air-cushion trains have speed potentials of up to 300 miles per hour; track costs are relatively low because of the simple concrete structure involved, which can be elevated on pylons, laid on the surface, or sunk in tunnels. Engineers in Britain, the United States, France, and Germany see this kind of high-speed surface transport as a means of connecting large urban centres with each other and with international airports.

Industrial applications. The other major area in which air-cushion technology has proved itself useful is in moving loads over surfaces that would be impossible for wheeled or tracked vehicles. One of the first of these applications was an air-cushion amphibious military vehicle. Another was an air-cushion truck built in France. Air-cushion load lifters for specialized applications were already in regular use in the early 1970s. One of the first was a British heavy-load carrier, designed specifically for the transport of large electrical transformers over bridges that were not stressed for the weights involved when normal wheeled transport was used. The transformers, weighing up to 300 tons, must be transported by road from the factory to the often remote power-station site. The costs of strengthening bridges sufficiently to take the load are far higher than those of fitting a special skirt and air-blower system to a conventional truck.

ACV load carriers

Similar systems on a smaller scale have been developed for transport of many types of unusual loads over awkward ground. They find particular application in the Arctic regions, where roads are often lacking and where oil-drill rigs are being constructed or where surveying is being carried out. Air-cushion vehicles have shown a large economic advantage over helicopters.

On an even smaller scale air-cushion pallets are used in many industries in the United States and Britain to carry heavy loads across factory floors or along production lines. Again the main benefit is the reducing of strain on floors and the ease of propulsion with reduced friction.

One unique form of air suspension may be employed in hospitals. The Hover-bed is a device on which a patient is supported with the minimum of body contact and surface pressure. The bed is being tested by the British Medical Research Council and is expected to be of particular use in cases in which the patient is burned over a large area of the body. Air support in such cases not only relieves pressure and pain but provides a film of sterile air that actually helps to heal the wound.

Sporting Hovercraft. Still another application promises to be that of recreation. In the mid-1960s enthusiasts in many parts of the world began to build their own Hovercraft, powered by automotive engines and using homemade propellers, fans, and skirts.

By the 1970s, Hover clubs existed in more than seven countries. The movement is probably most active in Britain, followed by the United States and Australia. Of the several hundred amateur-built craft, a few were capable of operating successfully over both land and water with more than one person aboard. Although power plants and other equipment are readily adaptable, the problem of providing sufficient lift was considerable. A craft capable of lifting 200 pounds (90 kilograms) needs a cushion area of 8 × 4 feet (1.3 × 2.4 metres). When weights of 1,000 pounds are involved, the area must be much greater. Another inhibiting factor in the sport's growth is the restrictions in most countries on operating air-cushion craft in public areas.

The future. Air-cushion vehicles have not yet fulfilled their original promise. Conventional skirted craft have not yet been shown to be completely economical in commercial use, although in certain military applications they are almost ideal. The ram-wing craft described earlier shows promise for over-water routes.

One area of research in which manufacturers are concentrating in the 1970s is that of secondary suspension systems to iron out skirt undulations and possibly reduce spray ingestion.

BIBLIOGRAPHY. WILLIAM T. GUNSTON, *Hydrofoils and Hovercraft* (1970), is a comprehensive, easy-to-understand survey that is well-illustrated and includes a glossary of terms and a table of principal Hovercraft and ACV's. More technical, requiring some college-level mathematics, is G.H. ELSLEY and A.J. DEVEREUX, *Hovercraft Design and Construction* (1968), an excellent textbook, based on lectures given at the Isle of Wight Technical College, which deals with the theory and practice of ACV-building. ROBERT L. TRILLO, *Marine Hovercraft Technology* (1971), covers operational problems to a greater extent. See also *Jane's Surface Skimmers* (annual), a review of ACV's and hydrofoils giving technical details of craft available, operators, and consultants, with brief bibliography; and *Air-Cushion Vehicles* (bimonthly), a technical review of the state of the Hovercraft art including design, construction, and operation of ACV's of all sizes.

(Jo.Be.)

Air Law

Air law as currently defined is the body of law directly or indirectly concerned with civil aviation. Aviation in this context extends to both heavier-than-air and lighter-than-air aircraft. Air-cushion vehicles are not regarded as aircraft by the International Civil Aviation Organization (ICAO), but the practice of individual states in this

regard is not yet settled. Because of the essentially international character of aviation, a large part of air law is either international law or international uniform law (rules of national law that have by agreement been made internationally uniform). The earliest legislation in air law was a 1784 decree of the Paris police forbidding balloon flights without a special permit. Insofar as international air law is concerned, it need hardly be mentioned that an international agreement or an amendment thereto is binding only on states that are parties to it.

THE LAW OF AIRSPACE

Sovereignty. A basic principle of international air law is that every state has complete and exclusive sovereignty over the airspace above its territory, including its territorial sea. At the turn of the 20th century the view that airspace, like the high seas, should be free was sometimes advanced. But the principle of airspace sovereignty was unequivocally affirmed in the Paris Convention on the Regulation of Aerial Navigation (1919) and subsequently by various other multilateral treaties. The principle is restated in the Chicago Convention on International Civil Aviation (1944). Airspace is now generally accepted as an appurtenance of the subjacent territory and shares the latter's legal status. Thus, under the Geneva Convention on the High Seas (1958) as well as under international customary law, the freedom of the high seas applies to aerial navigation as well as to maritime navigation. Vertically, airspace ends where outer space begins.

Right to regulate the entry of foreign aircraft

It follows from the principle of airspace sovereignty that every state is entitled to regulate the entry of foreign aircraft into its territory and that persons within its territory are subject to its laws. States normally permit foreign private (*i.e.*, nongovernmental and noncommercial) aircraft to visit or fly through their territory without too much difficulty. Such aircraft registered in states that are parties to the 1944 Chicago Convention are, under the convention, allowed into the territories of all other contracting states without prior diplomatic permission if not engaged in the carriage of passengers, mail, or cargo for reward.

Commercial air transport is divided into scheduled air services and nonscheduled flights. Charter flights fall mostly, but not invariably, into the latter category. Under the Chicago Convention, contracting states agree to permit aircraft registered in the other contracting states and engaged in commercial nonscheduled flights to fly into their territory without prior diplomatic permission and, moreover, to pick up and discharge passengers, cargo, and mail, but in practice this provision has become a dead letter.

For scheduled air services, the privilege of operating commercial services through or into a foreign country was, at the time of the 1944 Chicago conference, split into five so-called freedoms of the air. The first is the privilege of flying across a country nonstop; the second, of flying across with a stop for technical purposes only. These two freedoms are also known as transit rights. A large number of ICAO members are parties to the 1944 International Air Services Transit Agreement, placing these rights on a multilateral basis. The other freedoms of the air are known as traffic rights, referring to passengers, mail, or cargo carried on a commercial service. The third of the five freedoms is the privilege of bringing in and discharging traffic from the home state of the aircraft or airline; the fourth is that of picking up traffic for the home state of the aircraft or airline; the fifth is that of picking up traffic for or discharging traffic from third states in the territory of the state granting the privilege. This fifth freedom is the main bargaining point in the exchange of traffic rights among states. Attempts have been made since 1944 to create other freedoms, but each new freedom usually represents in practice a new restriction.

Efforts to conclude a widely acceptable multilateral agreement on traffic rights were unsuccessful, and such rights have continued to be handled through bilateral international agreements. These agreements fix the routes to be served, the principles governing the capacity of the agreed services (frequency of the service multiplied by the carrying capacity of the aircraft used), and the procedures for the approval of fares and tariffs by the respective governments. Most agreements require that airlines operating the same routes consult among themselves before submitting their fares to the two governments concerned for approval, and many agreements specify the International Air Transport Association (IATA), an association of airlines, as the organ for such consultations. The right to carry domestic traffic between points within a state is normally reserved to that state's own airlines. A bilateral agreement signed at Bermuda in 1946 between the United Kingdom and the United States set a pattern that has generally been followed, although the formal Bermuda-type agreement is likely to be accompanied by confidential memoranda attaching various restrictions.

Private rights. The principle of airspace sovereignty in international law is probably well reflected in the maxim, *Cujus est solum ejus est usque ad coelum et ad inferos* ("he who owns the land owns what is above and below it"). In private law the acceptance of this maxim for a long time posed little difficulty, and the Code Napoléon of 1804 adopted it almost verbatim; in more recent times, however, it is more than questionable whether such a principle can be accepted without qualification. Both the German Civil Code (1896) and the Swiss Civil Code (1907), while recognizing the principle of *Cujus est solum*, adopted a functional approach, limiting the right of the owner to such a height and such a depth as are necessary for his enjoyment of the land. In common-law countries the courts have arrived at a broadly similar position. In France, too, both the doctrine and the courts have refused to take *Cujus est solum* literally. In one celebrated case, *Clément Bayard* v. *Coquerel* (1913), the Court of Compiègne, lending judicial authority for the first time to the theory of abuse of rights, awarded damages to a plaintiff whose balloon had been destroyed by "spite structures" erected by the defendant on his own land and ordered the offending spikes to be taken down.

In the course of the 1920s it became clear in most countries, either through judicial decisions or express legislation, that aircraft would be allowed to fly over the private properties of others in normal flight in accordance with aeronautical regulations. This immunity applies only to the mere passage of the aircraft and does not extend to damage caused by it or to other encroachments on the use or enjoyment of the land, such as excessively low flights.

THE LAW RELATING TO AIRPORTS

In most countries airports may be privately, municipally, or nationally owned and operated, and the siting of an airport may be subject to town and country planning or zoning regulations. Whether or not the establishment of an airport requires special permission, aircraft leaving or entering a country will normally be required to do so at an airport having customs and immigration facilities. Airports that are open to public use are generally subject to some form of licensing or control in order to ensure compliance with minimum safety standards. Members of ICAO, in order to comply with their obligations under the Chicago Convention, have to make certain that such airports are open to aircraft of all other ICAO members under the same conditions as they are open to national aircraft. Restrictions may also be imposed on the noise level of aircraft taking off or landing, as well as the general level of noise, vibration, smoke, and so forth that may result from the operation of airports. In order to secure safety of flight, restrictions may be imposed on the use of lands adjoining an airport, such as the height of buildings or the planting of trees. Practice varies as to whether such restrictions are regarded as true measures of planning or zoning or as takings of private property for public use, which require the payment of compensation.

Some legal systems exempt the airport owner, operator, and users from liability for low flights over neighbouring properties, noise, vibration, or other forms of disturbance, provided that all the regulations and conditions

laid down for the operation and use of the airport are complied with. In the absence of such immunity, granted by law or obtained privately from adjacent landowners, the owners, operators, and users of airports are basically liable, in much the same way as other occupiers of land, for any substantial impairment of the use or enjoyment of neighbouring lands.

THE LEGAL STATUS OF AIRCRAFT

The channelling of rights and duties through the principle of nationality

Nationality. Among the most important points resolved in the 1919 Paris Convention were that aircraft should have a nationality, that they should have the nationality of the state in which they were registered, and that no aircraft could be validly registered in more than one state. The 1944 Chicago Convention retained these principles. While both conventions preclude dual or multiple registration, the ICAO Council in 1967 recognized the possibility of joint registration of aircraft by a number of states, and even "international registration"—without, however, specifying what the latter meant. The principle that every aircraft, at least every one that flies outside its country of origin, must have a nationality is of cardinal importance in air law, inasmuch as it enables a number of rights and duties to be either directly grafted onto the aircraft or channelled through the aircraft to a variety of persons. At the international level, moreover, it ensures that there will be no aircraft for which there is not a state answerable.

Under the 1944 Chicago Convention an aircraft, in order to benefit from the privileges conferred by the convention, must comply with its terms. Many of these terms are further elaborated in annexes to the convention. According to Article 20 of the convention, as among the contracting states, "every aircraft engaged in international air navigation shall bear its appropriate nationality and registration marks." Under Article 31, "every aircraft engaged in international navigation shall be provided with a certificate of airworthiness issued or rendered valid by the State in which it is registered"; in 1960 a number of European countries signed, at Paris, a multilateral agreement relating to Certificates of Airworthiness for Imported Aircraft, which is open to accession by other states, designed to facilitate mutual recognition of certificates of airworthiness for import and export purposes. Under Article 30(a) of the Chicago Convention,

> aircraft of [*i.e.*, having the nationality of] each contracting State may, in or over the territory of other contracting States, carry radio transmitting apparatus only if a licence to install and operate such apparatus has been issued by the appropriate authorities of the State in which the aircraft is registered.

As regards the operating personnel of the aircraft, the Chicago Convention provides that

> the pilot of every aircraft and the other members of the operating crew of every aircraft engaged in international navigation shall be provided with certificates of competency and licences issued or rendered valid by the State in which the aircraft is registered.

When an aircraft registered in one contracting state is in or over the territory of other contracting states,

> radio transmitting apparatus may be used only by members of the flight crew who are provided with a special licence for the purpose, issued by the appropiate authorities of the State in which the aircraft is registered.

In addition, the convention prescribes that

> there shall be maintained in respect of every aircraft engaged in international navigation a journey log book in which shall be entered particulars of the aircraft, its crew and of each journey. . . .

All of the above documents must be carried by "every aircraft of a contracting State, engaged in international navigation," as well as the appropriate manifests if passengers and cargo are carried.

The channelling of all these rules concerning the aircraft and its crew through the state of registry can give rise to problems when an aircraft is leased or chartered for any length of time to operators of a different nationality ("interchange of aircraft"). They can sometimes be resolved by a temporary transfer either of de facto control or of registration to the state of the operator.

The provision and operation of ground and other air navigation facilities, as well as the establishment and enforcement of air navigation rules and air traffic control, are the responsibility of the territorial state. So is investigation of accidents, though among ICAO members, under the Chicago Convention the state of registry

> shall be given the opportunity to appoint observers to be present at the inquiry and the State holding the inquiry shall communicate the report and findings in the matter to that State.

Among ICAO members, over the high seas the Rules of the Air established by ICAO apply. Enforcement rests primarily with the state of registry, which is also responsible for investigating accidents occurring over the high seas. A body known as Eurocontrol, established in 1960 by the Brussels Convention Relating to Co-operation for the Safety of Air Navigation, represents an attempt at international cooperation in air-traffic control by a number of western European states.

Private rights in aircraft

Registration of aircraft for nationality and public-law purposes is to be distinguished from registration for purposes of private law. Some legal systems treat aircraft simply as ordinary movable property. Others require all sales of, and other transactions relating to, aircraft, such as mortgages, to be effected in writing and recorded in a public registry before they may be invoked against third parties. Yet others regard only rights duly recorded as valid. If aircraft are to be used as security for credit or loans, a system of recording of rights with international recognition of the rights so recorded has obvious advantages. To this end, a Convention on the International Recognition of Rights in Aircraft was concluded in Geneva in 1948. Few states accepted it at first, but, with the rising cost of modern aircraft, interest in the convention increased. Its wide acceptance will have the side effect of bringing about much greater uniformity in rules of private law governing rights in aircraft.

Acts and occurrences on board aircraft. *Criminal jurisdiction.* Although some systems of national law still adhere to the view that ships and aircraft are part of the territory of the state the nationality of which they possess, this is merely a crude metaphor. In international law, a distinction has to be made between three types of state jurisdiction: territorial jurisdiction over national territory and all persons and things therein; quasi-territorial jurisdiction over national ships and aircraft and all persons and things thereon; and personal jurisdiction over all other nationals and all persons under a state's protection, as well as their property. In case of conflict, territorial jurisdiction overrides quasi-territorial jurisdiction and personal jurisdiction, while quasi-territorial jurisdiction overrides personal jurisdiction.

For a long time, the failure of states to extend their criminal laws to their aircraft while they were outside national territory posed a serious problem. As long as an aircraft is flying in the national airspace of some state, the law of that state is applicable. When a crime has been committed during an international flight, however, there may be difficulty in pinpointing when and where it occurred and hence in determining the state the law of which has been violated. Unless the criminal law and jurisdiction of the state of registry have been extended to the aircraft during the period it is outside the state of registry, there may be none applicable; over the high seas, for example, there would clearly be a gap in the law. This, together with the realization that with the constant increase in air traffic the incidence of offenses on board aircraft was bound to rise, led to the conclusion in 1963 at Toyko of the Convention on Offences and Certain Other Acts Committed on Board Aircraft, obliging the contracting states to extend their criminal law and jurisdiction to aircraft of their registry when they are outside national territory. The convention furthermore gives the aircraft commander power to ensure law and order on board his aircraft and to disembark any offender in any contracting state in which the aircraft lands.

The Tokyo Convention

Civil jurisdiction. In most countries the general civil law applies, except as modified. In the interest of avoiding statelessness, most states confer their nationality on those born on aircraft of their registry; but there is in air law no general principle of the law of the flag (*i.e.*, the law of the state of registry) being applicable to every occurrence on board. There are, however, various international agreements that affect the exercise of civil jurisdiction by states. A few may be mentioned. The most important is doubtless Article 28 of the 1929 Warsaw Convention on International Carriage by Air, as subsequently modified by Article 8 of the 1961 Guadalajara Convention and amplified by Article 12 of the 1971 Guatemala City Protocol. Under Article 28, an action arising from an "international" carriage by air may be brought only before the courts of certain contracting states and no others. The 1933 Rome Convention on Precautionary Arrest of Aircraft, which has not been widely accepted, exempts aircraft actually used on government services or in commercial transport from precautionary attachment. In other cases, the giving of an adequate bond "shall prevent the precautionary attachment or give a right to immediate release." Among ICAO members, Article 27 of the Chicago Convention provides that, subject to certain conditions, aircraft of the contracting states on an international flight are exempt from seizure or detention on patent claims in the territory of other contracting states, without having to deposit a security. Under the 1952 Rome Convention on Surface Damage, in principle, actions may be brought only before the courts of the contracting state in which the damage occurred.

Crimes against aircraft. *Piracy.* The 1958 Geneva Convention on the High Seas intends to be declaratory of general international law when it defines the offense of piracy principally as

> any illegal acts of violence, detention or any act of depredation, committed for private ends by the crew of the passengers of a private [*i.e.*, nongovernmental and not noncommercial] ship or a private aircraft, and directed: (a) on the high seas, against another ship or aircraft, or against persons or property on board such ship or aircraft; (b) against a ship, aircraft, persons or property in a place outside the jurisdiction of any State.

The convention defines the effect of piracy under international law as follows:

> On the high seas, or in any other place outside the jurisdiction of any State, every State may seize a pirate ship or aircraft, or a ship taken by piracy and under the control of pirates, and arrest the persons and seize the property on board.

A state, having done so, may decide upon the penalties to be imposed and the fate of the ship, the aircraft, or the property. This definition of piracy deliberately excludes acts committed for political motives, as well as acts confined within a ship or aircraft, such as mutiny or the hijacking of an aircraft by its passengers or crew. Although some states, for example, the United States, have in their own laws categorized hijacking as aircraft piracy, this in itself is unable to bring about the consequences of piracy under international law.

Hijacking. Unlawful seizure is the legal name that states at the international level have given to aircraft hijacking. Thus, the 1963 Toyko Convention obliges contracting states to take all appropriate measures to restore control of an aircraft hijacked in flight to its lawful commander, and obliges the state in which the aircraft lands to allow the passengers and crew to continue their journey, and to return the aircraft and its cargo to those lawfully entitled to possession. In response to a wave of hijackings that began in 1968, the 1970 Hague Convention for the Suppression of Unlawful Seizure of Aircraft was concluded in an effort to prevent hijackers from finding immunity in any of the contracting states.

Far more grievous than hijackings as regards the number of persons injured or killed and of aircraft damaged or destroyed have been the many acts of sabotage or violence committed against aircraft and civil aviation installations. This led to the conclusion at Montreal in 1971 of the Convention for the Suppression of Unlawful Acts Against the Safety of Civil Aviation, on roughly the same lines as the 1970 Hague convention dealing with the problem of unlawful seizure.

LIABILITY TOWARD PASSENGERS, SHIPPERS, AND THIRD PARTIES ON THE SURFACE

Passengers, baggage, and cargo. Problems of liability arising from the international carriage of passengers, baggage, and cargo by aircraft are dealt with by the widely accepted Warsaw Convention on International Carriage by Air, which was concluded in 1929, amended in 1955 by the Hague Protocol, supplemented in 1961 by the Guadalajara Convention, and further amended in 1971 by the Guatemala City Protocol. The 1971 protocol was much influenced by the so-called Montreal agreement of 1966, which was an agreement among airlines accepting the terms of the United States government for withdrawing its denunciation of the Warsaw Convention. Not every party of the Warsaw Convention has accepted the 1955 or the 1971 amendment protocol; nor is a party that accepts either protocol required to denounce the treaty which the protocol amends. The result is that there are in fact three separate treaties (Warsaw, Warsaw–Hague, and Warsaw–Hague–Guatemala City) with a number of states being parties to more than one.

The contracting states to the Warsaw convention seek to reduce and eliminate problems of conflict of laws by agreeing to adopt uniform rules in their internal laws governing the international carriage of persons, baggage, or cargo by aircraft. The qualification "international" has the special meaning attributed to it by the convention and, in order to avoid ambiguity, carriage governed by the Warsaw Convention is often referred to as Warsaw carriage. Some states have extended the rules of the convention to carriage not governed by the convention. The rules established by the convention are mandatory in the sense that the parties to the contract of carriage cannot vary them, especially not to the detriment of the consumer. They are also exclusive in the sense that no additional claims can be brought against the carrier outside the terms of the convention.

In the event a passenger is injured or dies, or baggage or cargo is damaged or lost during "international" carriage, the convention makes the carrier prima facie liable. In order to avoid liability, the carrier has to prove that he and his servants and agents have not been at fault. His liability may also be excluded in whole or in part if the injured person is wholly or partly to blame for the damage. In return for the presumption of his liability, the carrier is given the benefit of limited liability, a benefit which he forfeits if it is proved that the damage resulted from his willful misconduct or that of his servants and agents. The convention lays down detailed rules on the issuance and contents of the appropriate documents of carriage (passenger ticket, baggage check, and airway bill) and enforces them with the penalty of absolute and unlimited liability.

Limits of liability

The limits of the carrier's liability were set out in French gold francs. The limit of liability for death or injury was set at 125,000 gold francs per passenger ($8,300 at the 1970 exchange rate). This was subjected to much criticism, principally in the United States. The Hague Protocol doubled the limit, but this was still considered by the United States as inadequate. Frustrated by its unsuccessful efforts to obtain agreement to raise the limit further, the United States government in 1965 gave notice of denunciation of the Warsaw Convention. The Montreal agreement was the result, in which carriers supposedly agreed voluntarily, when carrying passengers to, from, or by way of the United States, (1) to accept the principle of absolute liability, subject only to the exception of contributory negligence, and (2) to raise the limit of their liability to $75,000 inclusive, or $58,000 exclusive, of legal fees or costs. The Guatemala City Protocol adopted the principle of absolute liability for the carriage of both passengers and baggage and fixed the limits per passenger at 1,500,000 gold francs ($100,000) exclusive of legal costs for death or injury, and 150,000 gold francs ($1,000) for damage to, or loss of,

baggage. These limits are unbreakable: they apply even in the event of willful misconduct. For the carriage of passengers, the Guatemala City Protocol further lays down rules on possible supplementary compensation schemes and periodic increases of the new limit. The rules concerning documents of carriage, which can have important effects on the carrier's liability, were substantially modified at The Hague and again at Guatemala City.

Third parties on the surface. While aircraft may fly over private lands in normal flight, most national laws protect the landowner against excessively low flights and other undue interference with his possession and use of land, such as the performance of aerobatics and buzzing. In addition, normally, every system of national law allows third parties on the surface (of the earth) to recover for damage done by an aircraft in flight, or by persons and things in or falling from it, though in some countries there are limits to the amount recoverable. Some national laws base liability on fault. The majority, however, hold the owner or the operator of the aircraft absolutely liable for any damage caused. Insurance against third-party liability is compulsory in many states.

Surface damage

In 1933 a convention was concluded in Rome on surface damage caused by foreign aircraft. This was supplemented by the 1938 Brussels Insurance Protocol. Both have since been superseded by the 1952 Rome Convention on the same subject. The convention applies only to surface damage caused in a contracting state by aircraft registered in another contracting state. It accepts the principle of absolute liability, subject to the exception of contributory negligence, providing however that

> there shall be no right to compensation . . . if the damage results from the mere fact of passage of the aircraft through the airspace in conformity with existing air traffic regulations.

The convention makes the operator of the aircraft liable. The main raison d'être of the convention is said to be its scheme of compulsory insurance; in return, the liability of the operator is limited according to the weight of the aircraft, unless the damage is caused by a deliberate act of the operator, his servants, or agents, done with intent to cause damage. An unauthorized user of the aircraft also incurs unlimited liability. Actions to recover compensation may be brought only in the state in which the damage occurred. An important element in the convention is that judgments rendered under it are enforceable in all other contracting states. The convention came into force in 1958, but the number of states accepting it is not very high. The wording of Article 1(1) of the convention quoted above gives rise to doubt whether damage resulting from sonic booms is covered. In view of the treatment accorded to passengers by the Guatemala City Protocol, a revised convention establishing absolute, unlimited, and adequately secured liability for all damage caused by aircraft to third parties on the surface would seem indicated.

THE INTERNATIONAL CIVIL AVIATION ORGANIZATION

Under the 1919 Paris Convention, an International Commission for Aerial Navigation (ICAN, or CINA) was created with headquarters in Paris. In 1937 an Inter-American Technical Aviation Conference decided on the creation of a Permanent American Aeronautical Commission (CAPA). Both were superseded by the establishment in 1947 of the International Civil Aviation Organization (ICAO) under the 1944 Chicago Convention. The Soviet Union joined in 1970, making ICAO membership almost universal.

Functions of the ICAO

ICAO is a specialized agency affiliated to the United Nations, with headquarters in Montreal. It has a plenary Assembly, which holds the budgetary power; a permanent Council elected triennially, which, among other things, adopts the Annexes to the Convention; an Air Navigation Commission; an Air Transport Committee; and a plenary Legal Committee, which discusses matters of air law of common interest to members and prepares draft conventions for their consideration and acceptance.

One of the most important functions of ICAO is the preparation and periodical revision of international standards and recommended practices relating to civil aviation. It has done much to standardize aeronautical regulations throughout the world. Among other functions of ICAO may be mentioned those in connection with joint support programs among members for financing air navigation facilities and its technical assistance program. The ICAO Council, under the Chicago Convention, may also function as either a conciliation body or a judicial organ in disputes between members.

BIBLIOGRAPHY. The most recent and comprehensive general work covering both international and English air law is *Shawcross and Beaumont on Air Law*, 3rd ed. by P.B. KEENAN, A. LESTER, and P. MARTIN, 2 vol. (1966), with its supplements. An introductory work for the student is G.A. SEABROOKE, *Air Law* (1964). A.D. MCNAIR, *The Law of the Air*, 3rd ed. (1964); and P.C. NATHAN and A.R. BARROWCLOUGH, "Civil Aviation," in *Halsbury's Laws of England*, 3rd ed., vol. 5, pp. 1–248 (1953), both deal with the subject from the standpoint of English law. For the United States, see A.J. MATHES and T. MATTERN, *Manual of Aviation Law* (1952), a booklet that commends itself by its scope and brevity. Much more elaborate are F. DE BILLYOU, *Air Law*, 2nd ed. (1964); and R.W. FIXEL, *The Law of Aviation* (1967). ROBERT R. WRIGHT, *The Law of Airspace* (1968), is on private rights and is not limited to air law, while D.H.N. JOHNSON, *Rights in Air Space* (1965), deals primarily with international law and is especially useful for its treatment of the historical evolution of the subject and the rules of air warfare. On the Chicago Convention, on ICAO, and on multilateral and bilateral agreements governing the operation of international nonscheduled flights and scheduled air services, see B. CHENG. *The Law of International Air Transport* (1962). H. DRION, *Limitation of Liabilities in International Air Law* (1954), a standard work, deals with both the Warsaw Convention and the Rome Convention. B. CHENG, "The Law of 'International' and 'Non-international' Carriage by Air," *The Law Society's Gazette* (1963–64), covers the Warsaw Convention, the Hague Protocol, and the Guadalajara Convention. A.F. LOWENFELD and A.I. MENDELSOHN, "The United States and the Warsaw Convention," *Harvard Law Review*, 80:497–602 (1967), is a more or less official apologia for absolute liability and the Montreal agreement. These were also the subject of two symposia: *Journal of the Royal Aeronautical Society* (1967); and "Symposium on the Warsaw Convention as Modified by the Montreal Agreement," *Journal of Air Law and Commerce*, 33:519–726 (1967). T. BUERGENTHAL, *Law-Making in the International Civil Aviation Organization* (1969); and J. SCHENKMAN, *International Civil Aviation Organization* (1955), are both works of great scholarly merit. U.S. LIBRARY OF CONGRESS, LAW LIBRARY, *Air Laws and Treaties of the World*, 3 vol. (1965), for the use of the Senate Committee on Commerce, is a mine of information. D.A. COOPER (ed. and trans.), *The Air Code of the U.S.S.R.* (1966; orig. pub. in Russian, 1961), has useful notes and explanatory comments.

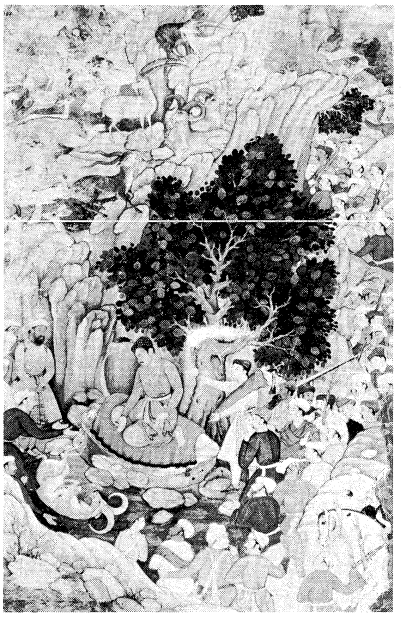
Air law is a dynamic subject, and much of its literature is found in periodicals, of which may be mentioned: *Aeronautical Journal* (London); *Current Legal Problems* (London); *ICAO Bulletin* (Montreal); *International Law Association Reports* (London); *Journal of Air Law and Commerce* (Dallas); and the *Yearbook of Air and Space Law* (Montreal).

(Bi.C.)

Akbar

Akbar, the greatest of the Mughal emperors of India, extended Mughal power throughout northern India and into the Deccan and established an elaborate administrative service on military lines, in which the highest posts were open to Hindus as well as Muslims. Although he never renounced Islām, he took an active interest in other religions, persuading Hindus, Parsis, and Christians, as well as Muslims, to engage in religious discussion before him. Illiterate himself, he encouraged scholars, poets, painters, and musicians, making his court a centre of culture.

Early life. Abū-ul-Fatḥ Jalāl-ud-Dīn Muḥammad Akbar was born at Umarkot, Sind, on October 15, 1542; he was descended from Turks, Mongols, and Iranians—the three peoples who predominated in the political elites of northern India in medieval times. Among his ancestors were Timur (Tamerlane) and Genghis Khan. His father, Humāyūn, driven from his capital of Delhi by the Afghan usurper Shēr Shāh Sūr, was vainly trying to estab-



Akbar, miniature portrait from the *Akbar-nāmeh* by Abū-ul-Fazl, *c.* 1600. In the India Office Library, London.
By courtesy of the India Office Library, London

lish his authority in Sind. Soon Humāyūn had to leave India for Afghanistan and Iran, where the Shāh lent him some troops. He regained his throne in 1555, ten years after Shēr Shāh Sūr's death. Akbar, at the age of 13, was made governor of the Punjab.

Humāyūn had barely established his authority when he died, in 1556. Within a few months his governors lost several important places, including Delhi itself, to Hemū, a Hindu minister who claimed the throne for himself. But a Mughal force defeated Hemū on the historical battlefield of Pānīpat, which commanded the route to Delhi, thus ensuring Akbar's succession.

At Akbar's accession his rule extended over little more than the Punjab and the area around Delhi, but under the guidance of his chief minister, Bayram Khān, his authority was gradually consolidated and extended. The process continued after Akbar forced Bayram Khān to retire in 1560 and began to govern on his own—at first still under household influences but soon as an absolute monarch.

Imperial expansion. Akbar first attacked Mālwa, a state of strategic and economic importance commanding the route through the Vindhya hills to the Deccan and containing rich agricultural land; it fell to him in 1561.

Toward the zealously independent Hindu Rājputs, a military race inhabiting rugged, hilly Rājasthān, Akbar adopted a policy of conciliation and conquest. Successive Muslim rulers had found the Rājputs dangerous, however weakened by disunity. But in 1562, when Raja Bihārī Mal of Amber (Jaipur), threatened by a succession dispute, offered Akbar his daughter in marriage, Akbar accepted the offer. The Raja acknowledged Akbar's suzerainty, and his sons prospered in the Mughal ruler's service. Akbar followed the same feudal policy toward the other Rājput chiefs. They were allowed to hold their ancestral territories, provided that they acknowledged Akbar as emperor, paid him tribute, supplied him with troops when required, and concluded a marriage alliance with him. The emperor's service was also opened to them and their sons, offering tempting rewards both in honour and in money.

Rājput policies

But Akbar showed no mercy to those who refused to acknowledge his supremacy. When, after protracted fighting in Mewār, Akbar captured the historic fortress of Chitor in 1568, he massacred its inhabitants. Even though Mewār did not submit, the fall of Chitor prompted other Rājput *rajas* to accept Akbar as emperor in 1570 and to conclude marriage alliances with him, although the state of Mārwār held out until 1583.

One of the notable features of Akbar's government was the extent of Hindu, and particularly Rājput, participation. Rājput princes attained the highest ranks, as generals and as provincial governors, in the Mughal service. Discrimination against non-Muslims was reduced by abolishing the taxation of pilgrims and the tax payable by non-Muslims in lieu of military service. Rājput princes had ruled unhampered in previous centuries, and some Rājputs and many other Hindus had cooperated with previous Muslim governments. Yet Akbar was far more successful than any previous Muslim ruler in winning the cooperation of Hindus at all levels in his administration. The further expansion of his territories gave them fresh opportunities.

Further expansions

In 1573 Akbar conquered Gujarāt, an area with many ports that dominated India's trade with western Asia, and then turned east toward Bengal. A rich country with a distinctive culture, Bengal was difficult to rule from Delhi because of its network of rivers, always apt to flood during the monsoon. Its Afghan ruler, declining to follow his father's example and acknowledge Mughal suzerainty, was forced to submit in 1575. When he rebelled and was defeated and killed in 1576, Akbar annexed Bengal.

Toward the end of his reign, Akbar embarked on a fresh round of conquests. Kashmir was subjugated in 1586, Sind in 1591, and Qandahār in 1595. Mughal troops now moved south of the Vindhya Mountains into the Deccan in peninsular India. By 1601 Khāndesh, Berār, and part of Ahmadnagar were added to Akbar's empire. His last years were troubled by the rebellious behaviour of his son, Prince Salīm, who was eager for power. Akbar died at Āgra in 1605.

Administrative reform. Previous Indian governments had been weakened by the disintegrating tendencies characteristic of pre-modern states—the tendency of armies to split up into the private forces of individual commanders and the tendency of provincial governors to become hereditary local rulers. Akbar combatted this trend by instituting comprehensive reforms that involved two fundamental changes. First, every officer was, at least in principle, appointed and promoted by the emperor, instead of his immediate superior. Second, the traditional distinction between the nobility of the sword and that of the pen was abolished: civil administrators were assigned military ranks, thus becoming as dependent on the emperor as army officers.

These ranks were systematically graded from commanders of ten to commanders of 5,000, higher ranks being allotted to Mughal princes. Officers were paid either in cash from the emperor's treasury or, more frequently, by the assignment of lands from which they had to collect the revenue, retaining the amount of their salary and remitting the balance to the treasury. Such lands seem to have been transferred frequently from one officer to another: this increased the officers' dependence on the emperor, but it may also have encouraged them to squeeze as much as they could from the peasants with whom their connection might be transitory. Politically, the greatest merit of the system was that it enabled the Emperor to offer attractive careers to the able, ambitious, and influential. In this way, Akbar was able to enlist the loyal services of many Rājput princes.

Akbar's reforms required a centralized financial system, and thus by the side of each provincial governor (*sūbadār*, later called nawab) was placed a civil administrator (*dīwān*) who supervised revenue collection, prepared accounts, and reported directly to the emperor. As a further safeguard against abuses, Akbar reorganized the existing network of newswriters, whose duty was to send regular reports of important events to the emperor. Akbar also seems to have instituted more efficient revenue assessment and collection in an effort to safeguard the peasants from excessive demands and the state from loss of money. But such efficiency could only have been enforced in the areas directly administered by the central government. This excluded the lands under tributary rulers such as the Rājputs and also the lands assigned for the maintenance of Mughal officers.

Financial reforms

Yet, notwithstanding Akbar's reforms, travellers' accounts indicate that the Indian peasants remained impoverished. The official elite, on the other hand, enjoyed great wealth; liberal patronage was given to painters, poets, musicians, and scholars, and luxury industries flourished. Akbar also supported state workshops for the production of high-quality textiles and ornaments.

Personality and assessment. Akbar maintained a luxurious and brilliant court, at which elaborate ceremonial emphasized his distance from other men, though he was careful to cultivate public opinion outside court circles. Every morning at dawn he stood at an open window to be seen and reverenced by the people. Foreign observers commented on the graceful manner in which he accepted little gifts from the people and showed himself ready to hear the complaint of any man who dared to approach him.

Physically, he was strong and could withstand hardship on campaigns. Although he seems to have been no more than five feet seven inches tall, he impressed observers as a dominating personality. Clearly, although he was illiterate, he had a powerful and original mind. His unprejudiced inquiries into Christian doctrines misled the Jesuit missionaries he invited to his court into thinking that he was on the point of conversion. He persuaded the Muslim theologians at his court to accept him as arbiter on points of Islāmic law in dispute among them. Although this seems to have been little more than an expression of his systematic approach to problems, the orthodox were offended. He gave further offense by the religious discussions he encouraged between Muslims, Hindus, Parsis, and Christians. These discussions were continued by a small group of courtiers who shared with Akbar a taste for mysticism. Although their doctrines and ceremonies, known as the Divine Faith (Dīn-e Ilāhī), assigned a central place to Akbar himself, it would be an oversimplification to ascribe political motives to these developments.

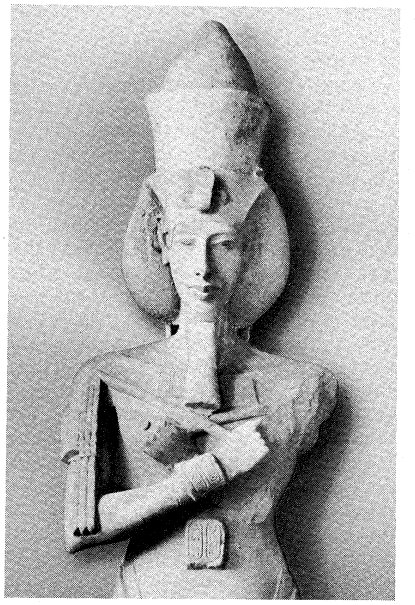
Begun in 1570 and abandoned in 1586, Akbar's capital of Fatehpur Sīkri, near Delhi, was evidence of the resources he could command. Its combination of Hindu and Muslim architectural styles symbolized the contact of cultures that he encouraged. Similarly, he commissioned the translation of Sanskrit classics into Persian, giving illustrated copies to his courtiers. He also received with enthusiasm the European pictures brought by the Jesuits, and his painters incorporated European techniques of realism and perspective in the distinctive Mughal style (characterized by a vivid treatment of the physical world) that began to develop during his reign. Akbar's reign was an example of the stimulating effects of cultural encounter. It has also often been portrayed as a model for future governments—strong, benevolent, tolerant, and enlightened. Effective government in a country geographically as vast and socially as complex as India demands a wide measure of social support: Akbar understood this need and satisfied it.

BIBLIOGRAPHY. VINCENT A. SMITH, *Akbar the Great Mogul, 1542–1605*, 2nd ed. (1919), is still the standard biography. ASHIRBADI LAL SRIVASTAVA, *Akbar the Great*, vol. 1, *Political History, 1542–1605*, vol. 2, *Evolution of Administration, 1556–1605* (1962–67), is a reliable account.

(K.A.B.)

Akhenaton

Akhenaton (sometimes written Ikhnaton, also known as Amenhotep IV or Amenophis IV), one of the world's great religious reformers, was king of Egypt in the 14th century BC. He reigned for 17 years (1379–1362 BC), during the age of Egypt's imperial power. Radical cultural changes were current in his day, and he attempted to emphasize some of them. As a result, later generations treated him as a heretic. Modern scholarship has divided sharply as to whether he was a noble idealist and the first monotheist in history or a bigoted zealot, whose attempted reforms brought Egypt to disaster. This article is principally concerned with Akhenaton the reformer, the aspect that is most available to historical description and interpretation.



Akhenaton, detail of the sandstone pillar statue from the Aton temple at Karnak, *c.* 1370 BC. In the Egyptian Museum, Cairo.
Hirmer Fotoarchiv, Munchen

Egyptian religion and culture before Akhenaton's reign. The religion of ancient Egypt was static and traditional, urging that the gods had given a good order and that it was necessary for man to hold firmly to the order. When changes did occur, religion tried to incorporate them into the system as though they came from the creation. By the time Akhenaton took the throne as the fourth pharaoh named Amenhotep, the 18th dynasty (1567–1320 BC) had run for nearly 200 years, and there had been a century of imperial conquest and control of foreign lands. Egypt dominated Palestine, Phoenicia, and Nubia. The nation was powerful, rich, and courted by lesser princes. To maintain these gains, a military and political group controlled the culture. Since the Egyptian state had always been theocratic, ruled by a god or gods, according to traditional beliefs, this group interlocked with the priesthood. The richest and most powerful of the gods, such as Amon of Thebes or Rę of Heliopolis, it was held, dictated the purpose of the state. The king had to apply to the gods for oracles directing his major activities. In return for wealth, elegance, and the role of the leading actor in a drama of imperial success, the pharaoh had relinquished his religious (and military) authority to others.

A century before Akhenaton, the energetic pharaoh Thutmose III had conquered nearby Asia and Africa. His successors continued his vigorous method of life, but, when the conquered territories were firmly held, that vigour turned from warfare to sports. Akhenaton's father, Amenhotep III, was a mighty hunter in his youth, but the son was weak physically and could not follow the pattern of outdoor feats. His activities were intellectual.

Beginnings of change before Akhenaton

The sudden spread of empire had excited the Egyptian culture. Architecture became less firmly planted and soared upward in assertiveness. Art lost some of its heavy cubism and became softer and rounder. Egyptian soldiers and officials lived in foreign countries, and foreigners lived in Egypt. The sharp differences between the people of the Nile Valley and the people abroad were blurred. Egyptian gods had temples in other countries, and foreign gods were introduced into Egypt. Gods and goddesses were concerned about Asia and Africa, as well as Egypt. Hymns before Akhenaton's reign show that the spread of empire meant the spread of religion. Gods became universal. Egypt had already combined gods, with Amon and Re (the sun-god) becoming Amon-Re or even Amon-Re-Harakhte. This permitted the Egyptians to think of the gods as unified forces, which was a prelude to monotheism.

There were other breaks in tradition. The royal line, wherever possible, had been kept pure by marriages with princesses of the king's own family. Amenhotep III defied this custom by marrying a commoner named Tiy. She apparently enjoyed unusual power in the palace without abandoning her loyalty to her husband. Akhenaton was a child of this marriage.

Empire was held firmly by garrisons abroad, which assured the favourable flow of trade to Egypt. Near the Nile Valley there were rich gold mines, so that the country could dominate both trade and politics. Messengers travelled between Egyptian and foreign cities carrying letters written in Babylonian cuneiform, the international language of the day, on clay tablets. This correspondence shows the imperial power and elegance of Egypt, which seemed to be assured of its unending dominance over all the nearby countries.

Akhenaton's early reign. Scholars disagree whether Amenhotep III associated his son Amenhotep IV on the throne for several years of coregency or whether the younger king succeeded to rule after the death of his father. It is here assumed that the older king died before the younger pharaoh gained power. The latter still used the family name Amenhotep, and on his ascension he still worshipped the old gods, especially Amon of Thebes and a sun-god, Re-Harakhte. His first buildings near Thebes were started in the older, massive architecture, using huge stone blocks and showing the worship of Re-Harakhte. The art was traditional, even though the figures of men and gods were carved in a softer outline than they had been a century earlier.

Within his first few years as pharaoh there were changes. He abandoned the temple to Re-Harakhte and began to build a place to worship a new form of sun-god —the disk of the sun, called the Aton. It had been a little known deity for two generations before him. The Aton was never shown in human or animal form, except insofar as the extended rays of the sun disk might end in hands to confer blessings upon men. This was the life-giving and life-sustaining power of the sun. He has no image in the hidden sanctuary of a temple but was to be worshipped out in sun-warmed openness. The buildings for the Aton were of a new kind. The massive solidity of the older temples was given up, and walls were run up of much smaller stones and were jammed with excited little scenes in a feverish new art. When artistic inventiveness was encouraged, forms were exaggerated to the point of caricature. Since the young king had a drooping jaw, a scrawny neck, sloping shoulders, a pot belly, and thick thighs, these features were carved in a grotesque way. The shape of the King became the flattering pattern for his followers, so that they also were shown with thin necks and round bellies. The King's wife, Nefertiti, may be beautiful in some of the sculptures made of her, but the new art often showed her as though she were a misshapen hag. Egyptian art changed from a static statement of eternity into a liveliness that is both fascinating and repelling. It came down from eternity to the here and now, with pictures of the King presiding over a specific ceremony, kissing his wife, or gnawing on a bone at the dining table. As long as art had shown the indefinite future, it had no exact time or place; now it told stories about what happened to the royal family and their followers.

Worship of Aton

Life and rule at Akhetaton (Amarna). The new temples were built at Karnak, near Thebes, a region dominated by the god Amon and by the families that had run the state for several generations. The new king had to break away sharply from this traditional setting. In the sixth year of his reign he changed his name from Amenhotep ("Amon is Satisfied") to Akhenaton ("He Who Serves the Aton"), thus formally declaring his new religion; and he moved his capital from Thebes more than 200 miles (300 kilometres) north of a desert bay on the east side of the Nile River, a place now called Tell el-Amarna or Amarna. Here he began to build a new city, which he called Akhetaton ("the Place of the Aton's Effective Power"). He took an oath that he would never go beyond the bounds of this place, which seems to mean not that he would never leave it but that he would not push the city limits beyond designated boundary stones. He was now free from the hostile forces at Thebes. Akhenaton, Nefertiti, and their six daughters gave themselves over to the new "truth." Their family life was open to the public. They worshipped the Aton in a temple open to the sunlight. The newly made nobles and officials gave their devotion to Akhenaton and Nefertiti.

Amenhotep becomes Akhenaton

Although the Egyptian texts always asserted that the king was a god and therefore the source of every benefit for the land, the complexities of a large empire, the activities of the bureaucrats, and the enticements of royal privilege had made Akhenaton's predecessors captives of the state. His political reforms were hence reactionary, inasmuch as he tried to recapture the old authority of the king. If the focus is only on the spectacular new elements that he introduced, the fact that his domination of rule was a restoration of a very old "truth" may be obscured.

The new city at Amarna must have had charm. Officials lived in spacious villas, with trees, pools, and gardens. Indoors the walls were painted in the free flowing new art, with marsh scenes near the floor and floral bouquets near the ceiling. Amarna art ranged from the very gracious, such as the famous bust of Nefertiti, to the grotesque. Everything was lively. Probably the elegant fragility of the portrait of Nefertiti displeased the traditionalists. Instead of presenting a solidity that might last forever, it gave a delicate and fleeting impression. The new "truth" came down to Earth. The prime minister was shown running in front of the King's chariot, an exertion that would have been unthinkable in the staid old times. Scenes of the busy market and the soldiers' guardroom, with lively comments of the people, are depicted. The present-day viewer of this ancient art feels as though he were there.

The new religion. The religion of the Aton is not completely understood. Akhenaton and Nefertiti worshipped only this sun-god. For them he was "the sole god." Akhenaton had dropped his older name Amenhotep, and the name "Amon" was also hacked out of the inscriptions throughout Egypt. Here and there the names of other gods and goddesses were removed, and sometimes the words "all gods" were eliminated. The funerary religion dropped Osiris, and Akhenaton became the source of blessings for the people after death. The figure of Nefertiti replaced the figures of protecting goddesses at the corners of a stone sarcophagus. Yet Akhenaton and Nefertiti directed their worship only to the Aton. It was the closest approach to monotheism that the world had ever seen.

The hymn to Aton

The king addressed a beautiful hymn to his god, expressing gratitude for the benefits of life. The Aton, says the hymn, gave these blessings not only to the Egyptians but also to "Syria and Nubia" and to "all distant foreign countries," to "all men, cattle, and wild beasts," to the lion coming from his den, the fish in the river, and the chick within the egg. Men live when the sun has risen, but at night the dark land is as if dead. It has often been pointed out that this hymn has a remarkable similarity to Psalm 104 in the Bible. Both the hymn and the psalm reflect a common family of ideas, according to which God or the god is praised for his bounties.

The Aton religion was a happy nature worship, without an ethical code. Men were asked only to be grateful to the sun for life and warmth. It was unlike the awful austerity of the great gods of former Egypt, who might punish man for disobedience. It was quite unlike the heavy demands that the Hebrew god would lay upon his people. In the Aton religion there was no "Thou shalt . . ." and no "Thou shalt not. . . ."

An aesthetic and intellectual religion, it probably had no deep roots. There is no evidence that the people worshipped the Aton happily; in their tombs they prayed to the Aton but also to Akhenaton and Nefertiti. The people wanted to see a definite force. Akhenaton was the same god-king who had always ruled Egypt, and Nefertiti could substitute for all the former mother goddesses. The

King and Queen seem to have accepted this worship as their just due.

The decline and end of Akhenaton's reform movement. The politics of the time must have been troubled. Although the ruling classes had been shorn of their powers, there was still an army. It may have been restless, because the documents show that Akhenaton paid little attention to it. Without a strong army and navy, foreign trade began to fall off, and internal taxes began to disappear into the pockets of local officials, finally causing the discontented priesthood and civil officials to combine with the army to discredit the new movement. Akhenaton was able to withstand these forces, but his weaker successors could not.

The Amarna Letters, discovered in the ruins of Tell el-Amarna from an archive of international correspondence directed by Asian princes to the courts of Amenhotep III and Akhenaton, reflect the new situation. The army commanders and high commissioners in Palestine and Syria were neglected. The local princes, who had seen their advantage in trading with Egypt, became despondent when Egypt did not answer their appeals for support. Hostile forces arose, ambitious princes in Palestine and Syria, invaders from the eastern desert, and the venturesome Hittites to the north. The Amarna Letters, as well as the archives found at the Hittite capital, show the disintegration of the Egyptian Empire in Asia. Loyal princes were forced to flee their cities. Aggressors, aided by the Hittites, captured territory from the Egyptian army. It may be that Egypt lost all of its holdings except the southwest corner of Palestine. Akhenaton's preoccupation with ideas and ideals cost Egypt its proud empire.

Akhenaton may have given in a little in the face of these disasters. In the 12th year of his reign the queen mother Tiy, a practical little woman, made a visit to Amarna. There is some evidence that he modified his extremism after that. The matter is confused, involving Akhenaton's estrangement from Nefertiti and the promotion of his young son-in-law Smenkhkare as a favourite. Since Smenkhkare apparently returned to Thebes, compromise seems to have been in the air.

When Akhenaton died, he was succeeded briefly by Smenkhkare and then by a second son-in-law, Tutankhaton. The latter was forced to change his name back to Tutankhamen, dropping the Aton and embracing Amon, to abandon Amarna and move back to Thebes, and to pay penance by giving the old gods new riches and privileges. When the tomb of Tutankhamen was discovered in the hills of western Thebes in 1922, it gave a final illustration of the sumptuous glories of Amarna art. A few years after the death of this young king, the army took over the throne in the person of General Horemheb. He instituted counterreforms in order to restore the old system fully.

Assessment

Akhenaton was a strange figure, spiritually and physically. Representations of his peculiar, unmanly body have been studied by pathologists with no unanimous conclusions. Some modern scholars have also questioned his ability to father children, but the presence of six daughters would certainly indicate that he was potent. Despite conflicting statements in the literature, it now seems certain that his mummy has never been found. Anciently and modernly he has been a controversial person, but the very fury of the controversy shows that he was a major figure of ancient history. The strong and changing forces of his day shaped his determined nature, and yet he stood estranged from his day in the strength of his ideas and ideals.

BIBLIOGRAPHY. The classic statement about Akhenaton was given by JAMES H. BREASTED in *A History of Egypt* (1905); and in the *Cambridge Ancient History*, vol. 2, ch. 6 (1924). Less admiring views are in SIR ALAN GARDINER, *Egypt of the Pharaohs*, ch. 9 (1961); and CYRIL ALDRED, *Akhenaten: Pharaoh of Egypt* (1968). Excavations at Tell el-Amarna are treated in T.E. PEET *et al.*, *The City of Akhenaten*, 3 vol. (1923–51). J.D.S PENDLEBURY, *Tell el-Amarna* (1935), was a handy summary up to that date. The art at Amarna is well represented in N. DE G. DAVIES, *The Rock Tombs of El Amarna*, 6 vol. (1903–08); and in J.D. COONEY, *Amarna Reliefs from Hermopolis in American Collections* (1965). The cuneiform correspondence called the Amarna Letters is treated by W.F. ALBRIGHT and T.O. LAMBIN in the *Cambridge Ancient History*, 3rd ed., vol. 1, ch. 4 (1966). A brief statement about the new religion is in JAROSLAV CERNY, *Ancient Egyptian Religion*, pp. 61–66 (1952). The hymn to the Aton was translated by J.A. WILSON in *Ancient Near Eastern Texts Relating to the Old Testament*, ed. by JAMES B. PRITCHARD, 3rd ed., pp. 369–371 (1969). An important technical report is R.G. HARRISON, "An Anatomical Examination of the Pharaonic Remains Purported to Be Akhenaten," *Journal of Egyptian Archaeology*, 52:95–119 (1966).

(Jo.A.W.)

Akiba ben Joseph

Akiba ben Joseph, 1st- to 2nd-century-AD Jewish sage in Palestine (born *c.* 40), was one of the principal founders of rabbinic (or Talmudic) Judaism and its literature, which became and remained the central, normative source for Jewish religion and life until modern times. In references to his life, scattered about the vast writings that make up this literature, fact and fiction are intermingled. Soon after his death, Akiba became a hero of popular legend. According to legend, he was an illiterate shepherd who turned to study in his maturity, becoming a student after 40. His devoted wife, Rachel, supported him both morally and materially during this arduous period of late learning (12 years according to one account). His principal teachers were the great masters of the Law, Eliezer ben Hyrcanus and Joshua ben Hananiah, themselves disciples of Johanan ben Zakkai, the illustrious founder of the academy at Jabneh. When, after his long period of study, Akiba returned home and was greeted by his wife, he reputedly said to his pupils: "What I am and what you are comes from her." He established his academy in Bene Beraq (near present-day Tel Aviv-Yafo) and the leading sages of the following generation, especially Meïr and Simeon ben Yoḥai, were his disciples. Among his disciples the sources also include an Egyptian proselyte, Minyamin (Benjamin).

Akiba perfected the method of biblical interpretation called "Midrash," whereby legal, sacral, and ethical tenets, sanctioned by tradition, were viewed as being implied in Scripture. Thus, Scripture, in addition to its overt meaning, is understood as replete with inferred teaching; it is, in fact, all-encompassing. The "Written Law" of Scripture and the "Oral Law" of tradition are ultimately one. Many midrashic works of the 2nd century originated in Akiba's school. In addition, he collected the oral traditions about the conduct of personal, social, and religious life (Halakha, "law"), and arranged them systematically, thereby laying the foundation of what was to become the Mishna, the first postbiblical code of Jewish law. (Akiba has been called "the father of the Mishna.") His apprehension of Scripture was opposed by the eminent contemporary midrashic exegete Rabbi Ishmael ben Elisha, who taught that "the Torah speaks in the language of men" and should not be forced to yield special meanings but be interpreted solely by means of set, logical rules of interpretation.

Teachings

Akiba's importance lies both in his achievements as a rabbinic scholar and in the impact of his personality on his time. He was strict in matters of law ("No pity in judgment!"—*i.e.*, compassion is irrelevant in establishing what the law is or means), but he opposed the death penalty. Modest in his personal life, he disapproved of parading one's knowledge. He counselled: "Take your place a few seats below your rank until you are asked to take a higher place." He respected the role of the woman in life and attributed the redemption of the Israelites from Egyptian bondage to the meritoriousness of women of that generation. "Man and his wife—the divine presence is between them," he taught. Known for his concern for the poor, he was appointed charity overseer. The term "to do the affairs of heaven" (religious deeds) he interpreted to mean "to scatter one's money among the poor."

As judge he addressed litigating parties: "Know before whom you are standing. You are standing before him whose word created the world, not before Akiba ben Joseph."

His lectures were on legal subjects, scriptural exegesis, and religious thought. For him the central teaching of Judaism resided in the commandment, "love your neighbour as yourself." God's love for man is expressed in that he created man in his image. Man has freedom of will ("Everything is foreseen, yet freedom of choice is given"); his deeds determine his fate, yet his true reward will be granted only in the world to come. In the present life there is much suffering, but "suffering is precious" and man should praise God for it. The people of Israel, who in a special sense are "God's children," have the task to "proclaim the glory of God to all the nations of the world." Akiba interpreted the Song of Solomon as a dialogue of love between Israel and God. For the sake of this love Israel withdraws from the affairs of the world. In these teachings—partly in answer to early Christian tenets—Akiba laid the basis for an ideology of Israel in dispersion among the nations of the world.

About the year 95, Akiba and other sages journeyed to Rome. Arriving at the seaport Puteoli they beheld the power and grandeur of the empire. While his companions wept, remembering the victory of Rome over Judaea some two decades ago, Akiba remained calm. If God is so kind to the wicked Romans, he explained, he will, in the end, be even kinder to Israel. He was equally calm when he visited the ruins of the Jerusalem Temple, destroyed by the Romans in the year 70. The prophecies of doom have come true, he commented; now we may anticipate the fulfillment of the prophecies of reconstruction.

The Bar Kokhba revolt: Akiba's martyrdom

Scholarly opinion is divided on the extent of Akiba's participation in an ill-fated rebellion against Rome (132–135) led by Simeon Bar Kokhba (also bar Koziba or ben Kosba). Some consider Akiba to have been the spiritual force behind the uprising. Others take note of the Talmudic report that Akiba considered Bar Kokhba to be the promised messianic king but see no evidence of further action on his part. Akiba was, it is true, apprehended by the Romans, imprisoned in Caesarea, and finally martyred (*c.* 135), but his offense is recorded as having been his continued public teaching rather than revolutionary activity. He accepted the agony of martyrdom serenely (he was flayed alive, according to tradition), grateful for the opportunity to fulfill the commandment to "love God . . . with all your life," which he always interpreted to mean "even when he takes your life." His last words were, "the Lord is one," the final words of the Jewish confession of faith ("Hear, O Israel! The Lord is our God, the Lord is one").

BIBLIOGRAPHY. LOUIS FINKELSTEIN, *Akiba: Scholar, Saint and Martyr* (1936), is the only extensive biographical narrative, with emphasis on Akiba's intellectual development.

(N.N.G.)

Alabama

One of the Southern states of the United States, Alabama comprises 51,609 square miles (133,667 square kilometres) forming a roughly rectangular shape, elongated in a north to south direction. Tennessee is the bounding state to the north, Georgia to the east, and Mississippi to the west. A westward coastal extension of northern Florida blocks access to the waters of the Gulf of Mexico except in the Mobile Bay region of Alabama's southwestern corner.

The ambience of the Deep South adds vivid character to the local landscape; it has permeated the troubled history of the region; and it touches, inescapably, much of the contemporary way of life in the state. It has also molded the images—and indeed the stereotypes—that the name of the state conjures up in the minds of outsiders.

In northeast Alabama, the broken terrain and fine scenery of the southern fringe of the Appalachians are often reminiscent of areas farther north, yet, in a southwesterly progression, past the belching smokestacks of the modern industrial giant of Birmingham, the Alabamian landscapes acquire a marked Southern appearance. The band of prairie lowland known as the Black Belt or Cotton Belt has rich soils that cradled a rural, cotton-producing way of life central to the state's development. Southward again, the historic houses, iron grills, and slave-made brickwork of old Mobile, the striking ranks of azaleas blossoming in the Gulf breezes, and the swaying, moss-draped oaks of the coastal fringe add an almost tropical note.

The landscape of Alabama has witnessed many of the major crises in the settlement of the continent and in the development of the modern nation. It was a battleground for European powers vying for the lands of the New World, for the savage fights between the white settlers and the Indians, for the internecine struggles of North and South in the Civil War, and for the forces of economic and social change that have extensively altered many aspects of the Deep South in recent decades. Although the Alabama of the early 1970s continued to trail near the bottom of the states in many significant social rankings, there has been some improvement in race relations, particularly in the area of school desegregation and in the election of blacks to political offices. The economy has also shown several growth points. Yet Alabamians and outsiders alike tend to agree that the state's Deep South heritage remained complex and often troubled. (For related information see UNITED STATES; UNITED STATES, HISTORY OF THE; CIVIL WAR, U.S.; and NORTH AMERICA.)

THE HISTORY OF ALABAMA

The Indians. The earliest and certainly the longest established inhabitants of what is now Alabama were Indians. Visible traces of their occupancy, which spanned almost 10,000 years, may be seen in the great mounds which snake across the landscape near the river valleys. Many place names in the state also betray an Indian origin. The principal Indian groups at the time of the initial European exploration of the region were the Chickasaw, in the northwest; the Cherokees, in the northeast uplands; the Upper Creeks, or Muskogees, in the centre and southeast; and the Choctaws of the southwest. Tribal boundaries, given the shifting nature of the way of life, were often uncertain.

European rivalry, settlement, and growth. The first known European explorers were of Spanish descent and arrived at Mobile Bay in 1519. In subsequent decades others penetrated inland, discovering several of Alabama's rivers. The main thrust of exploration, however, came from the northwest, when the great explorer Hernando de Soto, with an army of more than 500 men, entered the interior from the Tennessee River in 1540. His expedition—memorable for its discovery of the Mississippi River, the knowledge it gained of a wide band of southern Indian cultures, and its role in securing the whole of the southeast region of what became the United States to European settlement—is of importance to Alabama for the knowledge it gained during an extensive criss-crossing of the state. More tragically, a battle with the ill-equipped warriors of the Indian chief Tuscaloosa, on October 18, 1540, resulted in a slaughter which took the lives of at least several thousand of the indigenous peoples: it has been described as the bloodiest single encounter between whites and Indians in what was to become the United States. In 1559 a Spanish fort was established in Mobile Bay.

The de Soto expedition

The ensuing 250 years or so were to be characterized by struggles between French, British, and Spanish authorities for control of the region, often in shifting alliances with the Indians of the area. The French founded a community at Ft. Louis in 1702, the first permanent European settlement in Alabama. The British, too, had made a number of trips to the region from the Carolinas; but the French settlements, part of a string of forts arcing down from Canada and designed to contain the British, were more numerous. Port Dauphin, also founded in 1702, on Dauphin Island, witnessed the arrival of the first peoples of African descent, when a slave ship unloaded its human cargo in 1719.

The Treaty of Paris, in 1763, gave Mobile to Britain and established that country's supremacy in the area. In 1783 American control was recognized, although the

Spanish maintained a foothold in the south of the state until the United States, claiming the area as part of the Louisiana Purchase, finally established authority throughout the state in 1813. Washington County, in the southwest, was established as early as 1800, but the Alabama Territory was created only in 1817, and the state established in 1819. General Andrew Jackson, meanwhile, had inflicted a decisive defeat on the Creeks at the Battle of Horseshoe Bend in 1814, while the Chickasaw Indians had ceded their northern lands in 1805, followed a few months later by the Choctaw and Cherokees. The influx of population following the securing of the region and the introduction of the cotton gin meant a rapid removal of the Indians to the west. Some of their descendants now live in Oklahoma, while only a few hundred Creeks remain today in the southern part of the state.

The antebellum period. By 1820 Alabama's population was slightly over 125,000, including about 500 free Negroes. By 1830 there were 300,000 residents, 38 percent of them slaves, and cotton grown on the plantations of the so-called Black Belt was the principal money crop. The Southern way of life had begun to flourish. Up to the Civil War, domestic politics centred around the removal of the Indians, difficulties concerning the banking system, and, with ominously increasing frequency, the questions of state and federal rights and slavery. By 1860 the population was close to 1,000,000: nearly half were black. All but 5 percent of the state's population was rural in character. The University of Alabama (chartered 1820, opened 1831) and other educational institutions were flourishing, including (under an 1852 act) the beginnings of a public school system. The antebellum period also saw the introduction of the first railroad, the first canal, the first textile mill, and the first factory for manufacturing gin-milling machinery. By 1850 the mining of coal and iron and the manufacture of steel were obviously going to become big business in the state.

The Civil War and its aftermath. From 1860 to about 1900, the history of Alabama falls into three distinct periods: the Civil War, Reconstruction, and the reaction to Reconstruction. In 1861 Alabama seceded from the Union and joined the Confederate States of America, which was founded in Montgomery, its first capital. The state legislature conscripted soldiers and appropriated several millions of dollars for military operations and for the support of the families of soldiers. Some 35,000 of the 122,000 Alabamians who served in the war died. Following the collapse of the Confederacy and the refusal of the state legislature to ratify the 14th Amendment to the United States Constitution, Alabama was placed under military rule in 1867, but it was readmitted to the Union in 1868.

Reconstruction and after

From 1868 to 1874, the state was in political turmoil. To many whites, this Reconstruction period was a "tragic era"; to many blacks, it was a period of opportunity and hope. The *Huntsville Advocate* asserted that "This is a white man's government and a white man's state," and the Ku Klux Klan gained considerable influence. There was struggle for political power among whites who had supported the Confederacy, whites known as Scalawags who were willing to cooperate with the "Carpetbaggers" (Northern white immigrants), whites who had been loyal or sympathetic to the Union, and the blacks. The blacks demanded free schools; most whites insisted that separate schools for whites and blacks were most appropriate. Although blacks participated in the constitutional conventions and in the state legislatures, their political power was not as strong as that of blacks in South Carolina, Mississippi, or Louisiana. In 1874, the white Democrats, most of whom had been supporters of the Confederacy, regained control of the state political machinery; blacks were rendered almost powerless until about 1964. Throughout this period, some blacks worked diligently to stimulate political activity among their fellows, to enlighten and influence fellow white citizens, and to influence the state and federal governments to guarantee their political and social rights.

In 1875 a state constitutional convention was held, and a new constitution ratified. In 1876 federal troops were removed from Alabama, and political efforts began to centre around the aim of restricting the participation of Negroes in government. By 1901 when another state constitution was ratified, there were virtually no Negroes in government, and a tide of social and political reaction was in full flood. The economy, meanwhile, had recovered somewhat from the devastation of the War and its aftermath, and heavy industry was beginning to flourish. Economic activity generally remained at a low level, however, and depressed agricultural conditions fanned a populist revolt in the small farms in the 1890s. Educationally, life began to develop again, with the reopening of old institutions and the founding of Tuskegee Institute as a major centre of Negro education in 1881.

The 20th century. By 1900 the state was largely rural in character, and the black population began a slow decline, which had reduced their numbers to barely one-third of the total population by midcentury. Patterns of racial segregation under the "Jim Crow" system were rigidly enforced within and without the law. The onset of the boll weevil in 1915 wrought havoc to the one-crop cotton economy of the Black Belt, forcing a diversification of the rural economy. An exodus of rural dwellers, mostly poor and black, began, both to the cities of the state and to the North, where cheap foreign labour supplies had dried up during World War I. Rural society began a slow decline, arrested in part by the activities of the Tennessee Valley Authority in the northern areas in the 1930s and by a cooperative agricultural extension service. Defense industries boosted the urban economy in the World War II years, and the contributions of the federal government became evident in support of agriculture and the provision of such services as road building, education, and welfare. Segregation, nevertheless, continued to give rigidity to the social framework and effectively excluded the black population from social and economic power.

Race relations in the 1950s and 1960s. Following the 1954 United States Supreme Court decision that compulsory racial segregation in public education was unconstitutional and the 1957 Civil Rights Act which weakened black voter registration restrictions, black Alabamians and their white allies, often from outside the state, were given encouragement and tools to work to improve race relations. Further aid stemmed from the Civil Rights acts of 1960 and 1964 and other legislation. Progress was nevertheless slow and bitter: the state acquired international significance as the site of such landmark civil rights actions as the bus boycott of 1955–56 in Montgomery; the "freedom riders" of 1961; the sending in of federal troops to enforce desegregation laws in 1963, a year in which four small Birmingham black children were killed by a bomb which destroyed their Sunday School, and the long "Freedom March" from Selma to Birmingham in 1965.

The gains of the civil rights movement

As a result of these activities, black citizens had, by the late 1960s and early 1970s, acquired access to better public services, broader educational opportunities, slightly improved economic opportunities, and freer political participation—the percentage of voting-aged blacks registered to vote having tripled. Blacks began to appear in small but increasing numbers in elective state and local government positions and by 1973 included four sheriffs and six mayors. Positions were also won in clerical and sales and—rarely—supervisory and managerial positions in commerce, but there were hardly any blacks employed in state government agencies. A number of educational and professional bodies as well as many schools had achieved a good measure of desegregation, though the churches of the state had not. Progress was slow but, nevertheless, significant, and national attention was beginning to focus on problems elsewhere.

THE NATURAL AND HUMAN LANDSCAPE

The natural environment. *Topography.* Although the average elevation of Alabama is about 500 feet above sea level, this represents a gradation from 2,407 feet (734 metres) above sea level atop Cheaha Mountain in

the northeast down across the Black Belt to the flat, low, southern Gulf Coast counties. Within this gradation, several relief regions may be distinguished.

The southern extremities of the Appalachians, entering Alabama from the northeast, cover almost half the state. In the far north, the Cumberland Plateau region, segmented by river action, thrusts south across the state line. Altitudes rise to 1,800 feet in the more rugged eastern portions. The Great Appalachian Valley forms another marked division to the east. Finally, a small triangular portion of the Piedmont Plateau juts across from Georgia at an altitude averaging 1,000 feet.

The character of the state changes markedly as the rugged, forest-clad hills and ridges of the Appalachian extremities give way to the lower country of the Coastal Plain. The plain has a number of subdivisions: in the north lie the rolling Fall Line Hills, while farther south the pine and hardwood belts add irregularity to the flat landscapes. Arcing into the heart of the lowlands, the Black Belt has been distinctive from its association with the cotton production that long dominated its rich soils. Finally, the 53 miles of coastline have occasional swamps and bayous, backed by timber growth on sandy soils and fronted by stretches of white-sand beaches.

Climate and soils. The Alabama climate is temperate, with an average annual temperature of about 63° F (17° C), mellowed by altitude to some 60° F (16° C) in the northern counties, and reaching 67° F (19° C) in the southern counties, although summer heat is often cooled somewhat by the winds blowing in from the Gulf. Occasionally on a hot summer day the temperature may rise to a pounding 100° F (38° C), while snow may occasionally fall in the northern counties, and frosts are periodic. The average summer temperature is 79° F (26° C); the winter average is 46° F (8° C). Rainfall is fairly well distributed throughout the year, with a yearly average of 53 inches (1,350 millimetres) and a concentration on the coast. Droughts are infrequent. These favourable conditions have given the state a long growing season, ranging from 200 days in the north to 300 days in the south.

There are four main soil zones. In the far north, the Tennessee Valley contains dark loams and the red clays which add vivid dashes of colour to the landscape when exposed. Farther south lie the varied soils of a mineral belt, and these are succeeded by the rich limestone and marl soils of the Black Belt. Along the coast are sandy loams and deep porous sands.

Drainage patterns. The Cumberland Plateau region drains to the northwest through the Tennessee River and the often deep valleys of its tributaries, while the rest of the state is drained southward through broader, lazier valleys. The Coosa and the Tallapoosa rivers join north of Montgomery to form the Alabama, which unites with the Tombigbee, draining the state's western portion. Their waters then part again and are discharged into Mobile Bay through the Mobile and Tensaw rivers.

Plant and animal life. The warm climate of Alabama has nurtured a rich plant cover, including over 125 tree varieties. Most of the thick forest cover coats the north and northeast. Pine trees predominate, while the stately oaks are found statewide, lending grace to the streets of the older towns and cities. Sweet gum and black walnut are also common, while the colourful red cedar is most abundant in the Tennessee Valley and the Black Belt, with the stately black cypress clustering around rivers and ponds. There are many varieties of shrubs and grasses, and bamboo, large canes, and mistletoe are widespread. Muscadines, scuppernongs, and blackberries also flourish. Long clumps of Spanish moss, hanging from the coastal oaks swaying easily in the Gulf breezes, have added much to the traditional image of the state.

Bird life, too, is rich. Bluebirds, cardinals, blue jays, mockingbirds, doves, woodpeckers, owls, hawks, yellowhammers, and an occasional eagle add flashes of colour to the scenery. Other wildlife includes rabbits, squirrels, opossums, foxes, wildcats, raccoons, muskrat, the occasional fleet-footed deer, and even a lumbering bear or two. Reptiles include the poisonous rattlers, moccasins, and corals as well as some nonpoisonous types, including black snakes. Alligators still exist in some of the dank swamps and bayous of the coastal regions.

The human imprint. *Rural settlement.* By the early 1970s, under half the population of Alabama was classified as rural, and the proportion was declining still further as rural residents sought new homes in the cities or, in many cases, took the long bus or train ride to the uncertain life of the decaying inner-city regions of the North. Thirty-two rural counties suffered a population loss over the decade of the 1960s. Much of the loss was from the old cotton region of the Black Belt, but general farm consolidation had raised farm size and income and driven many from the land. For poor white and, especially, for black rural dwellers, the rural way of life had nevertheless nurtured a rich oral tradition and culture, often under difficult conditions. The sleepy rural communities and the scattered homes of often direly impoverished sharecroppers and farm tenants nevertheless continue to give a stark aspect to life in the country, while the old mansions of the former plantations, where not fallen into decay, add their own stately note. Some of the isolated upland communities near Tennessee also suffered from poverty and depopulation, in spite of the general boost given by the regional planning of the Tennessee Valley Authority in the 1930s.

Rural traditions

Urban settlement. By 1960, for the first time, more than half Alabama's population was concentrated in ten counties, and the dominance of the urban areas has continued. The sprawling industrial giant of Birmingham and its metropolitan area continue to dominate the state, with its increasingly diversified steel-based regional economy employing about a quarter of the state's workers and producing about one-fourth of all its manufactures. The most developed region in the state is roughly contained in a rectangle bounded by a line linking Birmingham, Tuscaloosa, Anniston, and Gadsden. Outliers of similar development lie around Huntsville; Muscle Shoals in the northwest; Mobile in the southwest; and in a cluster around the capital, Montgomery.

THE PEOPLE OF ALABAMA

Ethnic and religious composition. The three-quarters of the Alabamian population of the early 1970s that were white are significant for their deep roots in the state: the number of foreign born is very small indeed, and most of these are descendants of 19th-century settlers from adjoining regions to the east and north. The 900,000 or so black Alabamians have equally deep roots, dating back to the days of chattel slavery and the evils of the African slave trade. Religious affiliations in the state are predominantly Protestant, with the various groupings in the black community having played an unusually prominent social role since the days when other outlets for such activity were denied. The vast majority of the churches in the state remain without integrated congregations.

Demographic trends. The birth rate in Alabama, especially in the areas of rural poverty, was somewhat above the national average for many years, but with the depopulation of rural areas and the expansion of the urban economy, the ratio was only slightly above the national average by the 1970s. One notable result of the exodus of younger people from the state, when considered in conjunction with an influx of retired persons, has been an increase in the numbers of old persons in Alabama; the number of persons over 65 went up by one-quarter during the 1960s. For nonwhites, infant mortality rates were twice as high, maternal mortality rates four times as high. Almost 60 percent of the total population was defined as urban by the early 1970s, with a marked concentration of population in the six standard metropolitan statistical regions centred on Birmingham, Mobile, Gadsden, Montgomery, Huntsville, and Tuscaloosa. The first two cities named had populations of over 300,000 and almost 200,000, respectively, in 1970 and dominated urban life in the state.

THE STATE'S ECONOMY

Among the 50 states, the relative status of Alabama may be indicated by its per capita income: it has ranked very

close to the bottom of the scale, usually around 48th place, for a number of years. This low status results, in large part, from the depressed state of agriculture, which continues to occupy about a quarter of the population: rural poverty thus drags down the state average, concealing more promising developments and a much stronger economic base in the urban areas.

Agricultural and industrial resources. The Alabamian rural economy of the early 1970s presented many facets which challenge the traditional view of a dependency on cotton. Although cotton still remained of local importance, it suffered a heavy blow with the onset of the boll weevil blight in 1915, and acreage continued to decline while mechanization and consolidation increased average farm size. A more important process was the diversification of agricultural production which brought a great increase in the acreage devoted to forestry, while cotton fields have been given over to pasture for dairy and beef cattle. In the north, notably in the Tennessee Valley area and the old Black Belt cotton region, the gross income from cattle was much more important than that from cotton, making up over half total agricultural product sales. The principal crops were cotton lint, peanuts, soybeans, and corn. Farm income, except in the case of the poorer tenants and sharecroppers, continued to rise, and the average value of a farm rose from $13,000 in 1949 to almost double this figure 20 years later.

Diversification in agriculture

Industrial development in Alabama was for long based heavily on the iron and steel industry of Birmingham, the development of which was facilitated by the presence in the state of accessible deposits of the essential iron ore, coal, and limestone. Other minerals of note include the famous white marble. Oil production in commercial quantities dates from 1944, and there are a number of important wells in the coastal regions. World War II defense industries gave an important impetus to the industrial economy, and, by the 1970s, the production of metal products, chemicals, and plastics had lessened the reliance on primary-metal production, although iron and steel production remained the largest industry. From 1960 the George C. Marshall Space Flight Center at Huntsville, producing Saturn booster rockets, has been a major contributor to the state's economy.

The role of government. In spite of a per capita income that ranked it 48th among the 50 states, Alabama spent a high percentage of its total revenue on education, health and hospitals, welfare, highways, and natural resources. Federal funds supported programs affecting agriculture, public education, a wide range of health and welfare projects, conservation, urban development and public works, and highway construction. The federal government supported the Air University in Montgomery, Craig Air Force Base near Selma, the George C. Marshall Space Flight Center, and the Redstone Arsenal at Huntsville, as well as several veterans hospitals and a part of the Tennessee Valley Authority operations.

Transportation. Together, the 26 major rivers of Alabama provide more than 1,600 miles of navigable waterways, while Mobile Bay has been deepened by a ship channel. Mobile has undergone a spectacular development as a modernized port in recent years and ranked among the top dozen seaports of the nation by the early 1970s. Although railroad transportation, as elsewhere in the United States, was suffering a relative decline, bus, truck, and airline traffic were increasing.

ADMINISTRATION AND SOCIAL CONDITIONS

The political framework. At the state level, Alabama is governed by a bicameral legislature and a governor and his cabinet. The legislature consists of some 35 senators and about 100 representatives, who meet biennially in regular sessions and are elected for four-year terms. The constitution, a complex and, some have claimed, outdated and often inadequate document, dates from 1901. Various poll tax and other tax provisions aimed at restricting black-voter registration have been declared illegal by the federal Supreme Court, and voter registration among blacks has made slow but significant progress in recent years. The chief administrative officers, ranging from the governor to an eight-member State Board of Education, are all elected for four-year terms. The state Supreme Court of nine elected members is the highest judicial body.

State and local government

At the county level, the chief elected officials are the various county commissioners, judges of probate, tax assessors and collectors, and boards of education. In the municipalities, there is no uniform system of government: the mayor–council form is most common, but some cities have the commission type, and some employ a city manager.

The Democratic Party of Alabama remained in firm control of the state government in the early 1970s, although there were some signs of a slightly increased Republican showing. Numerous black political organizations have helped increase black voting and participation in the political process.

Services. At the start of the 1970s about a third of state revenues were spent on education, about a quarter on highways, around 12 percent on public welfare, and about half this figure on health programs. Alabama's low income among the states had the effect of giving amounts per capita spent on vital services a very low ranking compared with other states. Educational expenditure per pupil, for example, was a mere $511 in the second school year of the 1970s, ranking the state 50th; the average staff salary was also low, ranking 44th.

Education. Alabama had 15 state-supported four-year colleges, many junior colleges and trade schools, and almost two dozen private colleges and universities. Enrollment in public colleges was over 80,000 and that in private colleges over 16,000, annually. Black enrollment was higher in the private colleges (just over 6,000) than in the public system (almost 5,000). Most of the black students were enrolled in institutions specifically established for them, although a small but increasing number are attending formerly all-white colleges.

Health, welfare, and correction. In rural areas and among nonwhites, educational and economic status and opportunities are lower and fewer, and health and medical resources and services are less available. In 1969, for every 100,000 persons there were 32 dentists, 71 physicians, and 168 registered nurses. There were less than five nonfederal hospital beds per 1,000 population for general, short term, and special care. Welfare payments in Alabama rank low by national standards and, it has been claimed, have helped to swell the numbers leaving the state. Penal institutions number about half a dozen prisons and several camps for youthful offenders. Controversy over conditions in these institutions has occasionally erupted, and some measure of reform has subsequently been introduced.

CULTURAL LIFE AND INSTITUTIONS

There is a clear separation in Alabama between the cultural life of the state, as evidenced by many fine libraries and museums, and a rich vein of folklore and tradition in the rural areas, notably in the black community. Storytelling, in particular, has attracted the attention of folklore specialists, while the deeply felt compositions of rural blues singers have been an important element in the development of modern American popular music, including many varieties of jazz.

Folk culture

The more than 2,500,000 volumes of the state library system are distributed through city, county, and regional libraries: the last-named utilize mobile library vehicles. Special library collections include those at the University of Alabama Medical Center in Birmingham; at the George C. Marshall Space Flight Center in Huntsville; and the famous Booker T. Washington Collection of black-history material at Tuskegee Institute. The Alabama Department of Archives and History was the first such department established in the United States. Leading museums and art galleries are found in Birmingham, Montgomery, and Mobile; the George Washington Carver Museum at Tuskegee Institute is also very well known.

Historic sites in Alabama are supervised by the State Parks System: they include the Mound State Monument

in Hale County and Ft. Morgan, standing at the entrance to Mobile Bay. The state Department of Conservation maintains a fine range of parks and public lakes.

Among the state's many weekly and daily newspapers, the *Birmingham News* has the largest circulation. Newspapers published in the black community include the semiweekly *Birmingham World*. Alabama was the first state to establish (1953) a statewide educational television network, operating under the guidance of the Alabama Educational Television Commission. In the late 1970s there were 207 radio and 17 television stations operating throughout the state.

PROBLEMS AND PROSPECTS

Alabama's major socio-cultural problems seem to be the difficulty of maintaining interracial cooperation; significant racial differentials in educational, political, economic, and occupational opportunities and services; the low levels of education and health of a high percentage of the population; low per capita income; high rates of unemployment; high percentage of poor families and of families on public welfare; and an inadequate educational system. On the positive side, diligent efforts are being made to improve the educational, health, and economic services, to attract industry, and to improve the transportation systems and facilities. There is less racial discrimination than there was in the early 1960s, especially in public services. In some areas of employment, increasing opportunities exist for blacks; and in government, the barriers are being lowered, though slowly.

BIBLIOGRAPHY. JESSE M. RICHARDSON (ed.), *Alabama Encyclopedia and Book of Facts*, vol. 1 (1965), a volume of comprehensive data on a variety of topics including an excellent chronology; FEDERAL WRITERS' PROJECT, *Alabama: A Guide to the Deep South* (1941; reissued 1973), a semipopular historical account of the politico-economic, educational, and cultural life within the state.

History: LILLIAN ESTELLE WORLEY, *Alabama's People* (1945), a historical description; THOMAS PERKINS ABERNATHY, *The Formative Period in Alabama, 1815–1828*, 2nd ed. (1965), the story of Alabama as a part of the Mississippi Territory; CHARLES SHEPARD DAVIS, *The Cotton Kingdom in Alabama* (1939), an economic report on the relations among slavery, economics, and the law; WALTER LYNWOOD FLEMING, *Civil War and Reconstruction in Alabama* (1949), an account of the factors leading to the secession of Alabama from the Union in 1861 and of conditions during the Civil War and Reconstruction; ALLEN JOHNSTON GOING, *Bourbon Democracy in Alabama, 1874–1890* (1951), an analysis and interpretation of the period; CHARLES GRAYSON SUMMERSELL, *Alabama History for Schools* (1957), a well-organized and well-written history of Alabama for upper elementary and secondary school children; ELLEN LLOYD TROVER (ed.), *Chronology and Documentary Handbook of the State of Alabama* (1972).

Society and culture: CARL CARMER, *Stars Fell on Alabama* (1934), a report of selected observations and experiences while travelling in the Alabama Black Belt, Red Hills, Conjure Country, Mobile, and the Bayou Country over a six-year period; CLARENCE CASON, *90° in the Shade* (1935), observations on the South—a psychograph of Southern people in their physical setting.

Race relations: JOHN EGERTON, "Alabama: Six Years After 'The Stand'," *State Universities and Black Americans*, pp. 24–34 (1969), an account of the policy and practice of admitting black students and employing black teachers at the University of Alabama; RAY JENKINS, "Majority Rule in the Black Belt: Greene County, Alabama," *New South*, 24:60–67 (1969), a journalistic account of the political action of blacks who were successful in winning most of the elective offices in the county government in 1969.

(C.G.G.)

Alagoas

Alagoas, the second smallest of Brazil's 21 states, is an agricultural region in the early stages of industrialization. Located in northeastern Brazil, on the northern bank of the São Francisco River, it is bounded on the north and west by the state of Pernambuco, on the east by the Atlantic, on the south by Sergipe, and on the west by Bahia. It has an area of 10,707 square miles (27,731 square kilometres), which amounts to 0.33 percent of the area of Brazil. Its coastline is about 138 miles long. In 1975 its population was approximately 1,786,000, of whom about 60 percent lived in rural areas and 40 percent in cities. The name of the state derives from its large number of lakes (*lagôas*). The capital, Maceió, has about 324,000 inhabitants. (For coverage of an associated physical feature, see SAO FRANCISCO RIVER.)

History. Pôrto Calvo in the north and Penedo, on the northern bank of the São Francisco River, were founded during the early years of the Portuguese occupation of Brazil, in the 16th century. In the early part of the 17th century the Dutch obtained a foothold but were soon expelled. At about the same time, *quilombos*—fortified settlements of fugitive slaves, with a distinct culture that had many African characteristics—were established in the forests. Alagoas was a district of Pernambuco until 1817, when it became a captaincy. Following the proclamation of Brazilian independence in 1822, it became a province of the empire, and in 1889 it became a state of the republic.

Physical geography. The climate is hot and humid along the coast and hot and dry in the interior. The terrain slopes downward from a semi-barren plateau to a narrow coastal plain. The principal rivers are the São Francisco and the Mundaú, Paraíba, and Coruripe. Lagoa Mundaú, or Lagoa do Norte (Lake of the North), and Lagoa Manguaba, or Lagoa do Sul (Lake of the South), are the largest of the state's many lakes.

Geologically, Alagoas is old in structure, consisting mostly of the southern part of the Serra da Borborema (Borborema Mountain Range). The Serra Lisa (Lisa Mountain) is the state's highest point. There are four zones of vegetation: the coastal plain; the Mata, or tropical rain forest; the Agreste, a shrubby savanna parkland; and the Caatinga, an arid region covered with underbrush and cacti. The animal life is rich and varied.

Four zones of vegetation

People. The Alagoans are a mixture of peoples of European, African, and Indian origin. White- and black-skinned peoples predominate on the coast, *caboclos*—copper-skinned people of mixed European and Indian descent—in the interior. The language of the state is Portuguese. Roman Catholicism is the principal religion.

Administration. The state has a governor and a legislative assembly. In addition to the Tribunal de Justiça (state court), which has 11 judges, there are also district judges. There are 94 municipalities.

Social conditions. In the late 1970s there were about 2,700 primary schools (both public and private), with a total enrollment of about 244,000 pupils; and about 80 secondary schools, with about 17,000 pupils. In addition to the University of Alagoas, which has more than 3,000 students, there is a state school of medical sciences and a private school of social service. Eight institutions provide technical instruction.

Abundant electric power is available from the Paulo Afonso hydroelectric project on the São Francisco River; the towns are all supplied with electricity. The water supply, however, remains inadequate. The São Francisco River is being canalized, both for navigation and to supply the hinterland with water for domestic use. Sanitary conditions are poor. Although there are hospitals and other medical establishments in the state, all are in a precarious condition because of shortages of staff and funds.

Economy. While Alagoas is still primarily an agricultural state, industrialization is making visible progress. The main agricultural products are sugarcane, beans, palm, coconuts, cassava, rice, maize (corn), cotton, and tobacco. Cattle raising is increasing in the Bacia Leitera, a dairying region in the west central part of the state. Sugar refining is the state's most important industry, others including textile manufacturing, dairy processing, iron smelting, and the processing of maize, rice, fertilizers, and hides. Many manufacturing industries are located in the industrial district eight miles from Maceió.

Deposits of amianthus, a fine, silky asbestos, are currently being mined. There are a number of oil wells in production, other wells are being sunk, and further prospecting continues. Oil production and exploration are centred on Arapiraca, Maceió, Marechal Deodoro, São Miguel, Coqueiro Sêco, and Piaçabuço.

Expanding oil production

Transport and communications. Transport and communication links between Alagoas and the rest of Brazil have not been sufficient to permit the state to keep full pace with material progress achieved in other areas of the Northeast. There are about 460 miles (740 kilometres) of federal highway and about 2,000 miles (3,200 kilometres) of state roads crossing the state. There are about 240 miles (385 kilometres) of railroad track. The capital, Maceió (population 323,600 in 1975), is the chief commercial city, and a modest foreign and coastwise trade is carried on through its port, Jaraguá. Inland water transport is available on the São Francisco River and on several lakes. Flights from Palmares Airport link Maceió to other Brazilian state capitals.

Cultural life. Among the state's cultural institutions are the Historical and Geographic Institute, the Academy of Letters, and a medical society. Alagoas maintains the tradition of popular festivals featuring dances and folk plays. There are several newspapers and radio stations in the state. (M.G.F.)

Alaska

When it became the 49th state of the United States in 1959, Alaska increased the nation's area by nearly 20 percent. The new area included vast stretches of unexplored land and untapped resources. But when Secretary of State William H. Seward negotiated its purchase from Russia in 1867, it was known as "Seward's Folly." Since its acquisition, its settlement and exploitation have been hindered by its distance from the rest of the nation, the climate and terrain, and the slowness of communications. Many problems still stand in the way of immigration and economic development, and Alaska continues to be the country's last frontier. About 46 percent of the 382,000 inhabitants (1976 estimate) lived in and around the city and borough of Anchorage. The current capital is Juneau, 573 miles (922 kilometres) to the southeast in the Panhandle region. A new capital site at Willow, however, was chosen by referendum in late 1976.

Area and boundaries

Itself a landmass of subcontinental proportions, Alaska lies at the extreme northwest of the North American continent and is the largest peninsula in the Western Hemisphere. Its 586,400 square miles (1,518,800 square kilometres) embrace some 15,000 square miles (39,000 square kilometres) of fjords and inlets, and its three faces to the sea have about 34,000 miles (55,000 kilometres) of indented tidal coastline and 6,600 total miles (10,600 kilometres) of coast fronting the open sea. The marine borders are the Arctic Ocean on the north and northwest, Bering Strait and Sea on the west, and the Pacific Ocean and Gulf of Alaska on the south. The land boundaries on the east cut across some 1,150 miles (1,850 kilometres) of high mountains to separate the state from the Canadian Yukon Territory and province of British Columbia. Rimming the state on the south is one of the Earth's most active earthquake belts, associated on the mainland with zones of volcanic activity. In the Alaska Range north of Anchorage, Mt. McKinley, at 20,320 feet (6,194 metres), is the highest peak in North America.

In the years following statehood, Alaska emerged as the first real proving ground of the country's stated resolve to find a balance between the effective development and management of its natural resources and the preservation of its irreplaceable wilderness and public recreation areas. Under the Statehood Act of 1958, about 27 percent of Alaska remained under federal jurisdiction as national forests or parks, wildlife preserves, Indian reservations, and the like. Another 70 percent remained vacant public-domain lands. By 1984, about 100,000,000 acres (40,450,000 hectares) of this, plus tidelands and other land equities, were to be selected by the state for community, recreational, and commercial use. In January 1972 the state filed for withdrawal of 76,600,000 acres of this optional land from the public domain, adding to 24,600,000 acres previously selected and leaving about 2,000,000 acres yet to be selected.

In the early 1970s the problem of development versus preservation was symbolized by conflicts between conservationists and petroleum companies over a pipeline from the oil-rich North Slope on the Arctic Ocean to the cities in the south and beyond. Further complications arose through a vigorously espoused native land-claims issue. In response to the energy crisis, however, a bill that cleared the way for construction of the pipeline was passed and signed in November 1973. (For information on related topics, see the articles ALASKAN MOUNTAINS; NORTH AMERICA; UNITED STATES; UNITED STATES, HISTORY OF THE; VALLEY OF TEN THOUSAND SMOKES.)

THE HISTORY OF ALASKA

Explorations. As early as 1700, native peoples of Siberia had reported the existence of a huge piece of land lying due east. An expedition appointed by the tsar and led by a Danish mariner, Vitus Bering, in 1728 determined that the new land was not linked to the Russian mainland, but because of fog it failed to locate North America. On Bering's second voyage, in 1741, the spectacular peak of Mt. St. Elias (18,008 feet, or 5,489 metres) was sighted, and men were sent ashore. Sea otter furs taken back to Russia opened a rich fur commerce between Europe, Asia, and the North American Pacific Coast during the ensuing century.

Russian pioneering

Early settlement. The first European settlement was established in 1784 by Russians at Three Saints Bay, near present-day Kodiak. It served as Alaska's capital until 1806, when the Russian-American Company, organized in 1799 under charter from the emperor Paul I, moved its headquarters to richer sea otter grounds in the Alexander Archipelago at Sitka. The company governed Alaska until its purchase by the United States in 1867. Alaska's first governor (then termed chief manager), Aleksandr Baranov, was an aggressive administrator whose severe treatment of the native Indians and Eskimos led in 1802 to a massacre at Sitka.

A period of bitter competition among Russian, British, and American fur traders was resolved in 1824 when Russia granted equal trade rights for all. The near extinction of the sea otter and the political consequences of the Crimean War (1853–56) were factors in Russia's willingness to sell Alaska to the United States. The Russian minister made a formal proposal in 1867, and, after much public opposition, the purchase was approved by Congress, and the United States flag was flown at Sitka on October 18, 1867.

Political growth. As a United States possession, Alaska was governed by military commanders for the War Department until 1877. During these years there was little internal development, but a salmon cannery built in 1878 was the beginning of what became the largest salmon industry in the world. In 1884 Congress established Alaska as a judicial land district, federal district courts were established, and a school system was initiated.

In 1906 the first representative to Congress, a nonvoting delegate, was elected, and in 1912 Congress established the Territory of Alaska, with an elected territorial legislature. Alaskans voted in favour of statehood in 1946 and adopted a constitution in 1955. Congressional approval of the Alaska statehood bill in 1958 was followed by formal entry into the Union on January 3, 1959.

Establishment of the territory

Mining booms. Other significant events in Alaska's history included early gold discoveries on the Stikine River in 1861, in Juneau in 1880, and on Fortymile Creek in 1886, and later the stampede to the Atlin and Klondike placer goldfields of adjoining British Columbia and Yukon Territory in 1897–1900. Gold discoveries followed at Nome in 1898 and at Fairbanks in 1903. The gold rush made Americans aware of the economic potential of this previously neglected land. The great hard-rock mines in the Panhandle were developed, and in 1898 a copper mine was discovered at McCarthy. Gold dredging in the Tanana River Valley was begun in 1903 and continued until 1967.

Economic growth. A dispute between the United States and Canada over the precise boundary between British Columbia and the Alaska Panhandle was decided by an Alaska Boundary Tribunal in 1903. The American view that the border should lie along the crest of the

Boundary Ranges was accepted. Between 1898 and 1900 a narrow-gauge railroad was built across the precipitous White Pass to link Skagway and Whitehorse in the Yukon, and shortly afterward the Cordova-to-McCarthy line was laid up the Copper River. Another milestone was the 538-mile Alaska Railroad connecting Seward with Anchorage and Fairbanks in 1923. In 1935 the government encouraged a farming program in the Matanuska Valley near Anchorage, and there are now well-established dairy herds and crop farming there and in the Tanana and Homer regions.

In 1942, during World War II, Japanese forces invaded Agattu, Attu, and Kiska islands and bombed Dutch Harbor on Unalaska. This aggression prompted the construction of large airfields as well as the Alaskan, or Alcan, Highway linking Dawson Creek, British Columbia, and Fairbanks with more than 1,500 miles of road. Both proved later to be of immense value in the commercial development of the state. A devastating earthquake in March 1964 affected the northwestern Panhandle and the Cook Inlet areas, destroying parts of Anchorage; a tsunami wave that followed wiped out Valdez; the coast sank 32 feet at Kodiak and Seward, while a 16-foot rise destroyed the harbour at Cordova.

Discovery of oil

Oil discoveries in the Kenai Peninsula and offshore drilling in Cook Inlet in the 1950s created an industry that by the 1970s ranked first in the state's mineral production. In the early 1960s a pulp industry began to utilize the forest resources of the Panhandle. Major paper-pulp mills were constructed at Ketchikan and Sitka, largely to serve the Japanese market.

The discoveries in 1968 of petroleum on lands fronting the Arctic Ocean gave promise of relief for Alaska's economic lag, but problems of transportation across the state and to the "South 48" held up exploitation of the finds. In 1969 a group of petroleum companies paid the state nearly $1,000,000,000 in oil-land revenues, but the proposed pipeline across the eastern Brooks Range, interior plains, and southern ranges to Valdez created heated controversies among industry, government, and conservationists. In November 1973 a bill passed Congress and was signed, making possible construction of the pipeline, which began in the following year. The completed 48-inch (122-centimetre) pipeline, 789 miles (1,270 kilometres) long, came into operation on June 20, 1977.

THE NATURAL AND HUMAN LANDSCAPE

Physiographic diversity

The immense area of Alaska has a great variety of physical characteristics. Nearly one-third of the state lies within the Arctic Circle and has permanent frost and treeless tundra. The southern coast and the Panhandle at sea level are fully temperate regions. In these latter and in the adjoining Canadian areas, however, lies the world's largest expanse of glacial ice outside Greenland and Antarctica. Off the extreme western end of the Seward Peninsula, Little Diomede Island, part of Alaska, lies in the Bering Strait only 2.5 miles from Soviet-owned Big Diomede; both nations have shown a tacit tolerance of unintentional airspace violations, which are common in bad weather.

Environmental regions. Alaska is composed of nine distinct environmental provinces.

The Panhandle. Much of the mainland Panhandle, a narrow strip of land 25 to 50 miles wide lying west and south of the St. Elias Mountains, is composed of the Boundary Ranges. There are several large ice fields, and the peaks include Mt. St. Elias, from whose summit the Alaska–Yukon border swings due north. The western extension of this mountain chain is the Chugach Range, a giant arc at the northernmost edge of the Gulf of Alaska. Many remote valleys and high ridges are still unexplored, and the relief and glaciation inhibit exploitation. The coast is characterized by frequent and intense oceanic storm systems that have produced dense rain forests on the coastal mountain flanks. In the valleys, rivers produce devastating annual floods.

South coast and gulf islands. The south coastal archipelago and the Gulf of Alaska islands include the Alexander Archipelago in the Panhandle region, with 11,000 islands, plus Kodiak Island and its satellites south of Cook Inlet. These islands, extensions of the southern region, are lower, less rugged, and less glaciated. All experience extraordinarily heavy rain and are affected by encircling waters warmed by the Kuroshio Current.

Aleutian Range and islands. This region includes the narrow Alaska Peninsula and the 1,100-mile-long Aleutian chain that separates the North Pacific from the Bering Sea. The chain includes 14 large islands, 55 significant but smaller ones, and thousands of islets. The largest and most inhabited are Unimak, Unalaska, and Umnak. On the occasionally clear summer days, active volcanoes and such glacier-covered peaks as symmetrical Shishaldin Volcano (9,372 feet, or 2,857 metres) on Unimak can be seen. Such magnificent views represent the Aleutians at their scenic best. Usually, however, the weather is wet and stormy, the winds cutting, and the fog all-pervading.

Shishaldin Volcano

Alaska Range. A curving interior cordillera, the broad Alaska Range connects the Aleutian Range across the southern third of mainland Alaska to the Wrangell Mountains, which abut against the vast complex of the St. Elias Mountains. The Wrangell Mountains have both large active volcanoes and high valley glaciers. The flanks of this sub-Arctic range are largely tundra-covered.

Interior basins. The low-lying region between the Alaska Range in the north and the Chugach–Wrangell–St. Elias mountains to the south enjoys a relatively temperate climate. The lower valleys contain good farmlands, such as the fertile Matanuska area, and it is there that most of the people of Alaska live.

Interior Alaska. The central plains and tablelands are a vast region west and north of the Alaska Range and reaching as far north as the Brooks Range. The area is rolling and dissected by numerous streams tributary to the Yukon River. The plains extend from the Canadian border to Norton Sound and the Yukon Delta on the Bering Sea. It is characterized by river flats and truncated upland tablelands. A region with abundant game, it is an important nesting ground for waterfowl, including great numbers of migrating birds.

Brooks Range. A major mountain chain north of the central plains and extending from the sea nearly to the Yukon border, the Brooks Range gradually slopes northward to a narrow linear coastal plain bordering the Arctic Ocean and westward to lower hills north of Kotzebue Sound. There are a few high Arctic glaciers, and the area is semi-arid. The lower flanks and valleys are tundra-covered, with permafrost features.

Arctic plains. The coastal lowland north of the Brooks Range, sometimes called the North Slope, is the home of great herds of caribou. The environment is truly polar, with the seacoast frozen eight months a year and the ground permanently frozen except for a thin zone of summer melting. It is treeless, though in summer grasses and high Arctic-alpine flowers abound. A large navy petroleum reserve is located here, and the Prudhoe Bay oil fields are found in the east.

Sea islands. The islands of the Bering Sea represent a small but unique Arctic maritime environment, typified by St. Lawrence, Nunivak, and St. Matthew islands and the Pribilof group. These tundra-covered islands are surrounded by sea ice in winter and serve as protected refuges for the world's largest herds of fur-bearing seal and sea otter, as well as sea lion and walrus. A large herd of domesticated reindeer is tended by Eskimos on Nunivak Island.

Climates. The wide-ranging geographical provinces and the great physiographic relief, extending from sea level to more than 20,000 feet (6,000 metres), provide Alaska with much climatic diversity. Summers are mild, and midwinter along the coast is often clear and dry. It is thought likely that an approximately 90-year cycle of climatic change may bring another cold interval toward the end of this century. Five general climatic zones may be delineated, excluding the great mountain ranges.

Temperature and rainfall

Temperate oceanic. Southern coastal and southeastern Alaska, the Gulf of Alaska islands, and the Aleutians

have average temperature ranges in the summer of 40° to 60° F (4° to 16° C), in the winter of 40° to 20° F (4° to −7° C). Rainfall varies locally from 60 to 160 inches (1,525 to 4,065 millimetres), and the Panhandle and southern islands are covered with Sitka spruce and other evergreens. The Cordova-Valdez region has the state's highest precipitation, 200 inches (5,100 millimetres) or more. At Valdez 200 inches of snow is not uncommon. Precipitation is less in the Aleutians, but even there about 250 rainy days occur annually.

Subtemperate. The interior basin ranges from 45° to 75° F (7° to 24° C) in summer, 20° to −10° F (−7° to −23° C) in winter. The region is drier than the coast and only slightly colder in winter, with Anchorage receiving about 25 inches (635 millimetres) of precipitation annually. The pleasant conditions and proximity to the sea have made the area the centre of the state's population.

Maritime Arctic. The islands and coast of the Bering Sea have summer temperatures of 40° to 60° F (4° to 16° C) and winter temperatures of 20° to −10° F (−7° to −23° C). Tempering influences of the Pacific dissipate north of the Pribilof Islands, and Arctic sea ice often reaches this area.

Interior continental. The central plains and uplands range from 45° to 75° F (7° to 24° C) in the summer, −10° to −30° F (−23° to −34° C) in the winter. Average rainfall is 10 to 20 inches (255 to 510 millimetres), though less than 10 is common.

High Arctic, or polar. The ameliorating effects of the Arctic Ocean keep temperatures of the North Slope—35° to 55° F (2° to 13° C) in the summer, −5° to −20° F (−21° to −29° C) in the winter—less severe than those of the interior plains. About five inches (130 millimetres) of precipitation nonetheless remain on the ground as snow about eight months a year. The 24-hour sunlight of summer can produce strong buildups of radiant energy, sending temperatures to 90° F (32° C).

THE PEOPLES OF ALASKA

English, Russian, Spanish, and French place-names reflect the early European exploration, but equally as prominent are dozens of names carried down from the pre-Western era. The name Alaska itself is derived from the Aleut *alaska* and the Eskimo *ālakshak*, both meaning "mainland."

Aboriginal peoples. Long before Bering's voyages the Tlingit Indians lived in the southern and southeastern coastal area; the Aleuts on the Aleutian Islands and the Alaska Peninsula; and the Eskimos on the Bering shore and the Arctic Ocean coast. The interior natives were the Tinneh Indians, whose language was Athabascan, that of the Plains Indians of the interior continent to the south.

Native population and cultures

The Indian groups are presumably descendants of the earliest immigrants from Asia, perhaps more than 15,000 years ago, with the Alaskan Indians reflecting the migratory wave that reached as far as the southern extremity of South America. The Eskimos and Aleuts appear to be much later immigrants, having arrived perhaps 3,000 to 8,000 years ago. Rather than going south, they remained in the islands and coastal regions of the far north. All groups are involved in the debates over public land grants that are helping delay the economic development of Alaska in the 1970s.

Tlingits. The coastal Tlingits, who lived by fishing, hunting, and gathering, were wealthy, artistic, and fierce. They built oceangoing war canoes and roamed as far south as Puget Sound in pursuit of conquest and slaves. Their grass basketry was superb, they were expert carvers, and they built large wooden houses with magnificent decorative panels. They developed a rigid clan and caste system and respected a hereditary monotheistic religion and a pantheon of lesser spirits. Their religious spirits were everywhere in nature, in the sea and in the forests. After death the souls of men were delivered to "the spirit's home" in the ice country of the mountains where the living dared not venture.

A smaller tribe of paler faced and occasionally red-haired Indians, the Haidas, lived on the Queen Charlotte Islands in British Columbia and Prince of Wales Island in the Panhandle. Their culture was similar to the Tlingit, but anthropologically they showed an infusion of Hawaiian or Polynesian blood. Today, only a few hundred natives still speak the Haida tongue.

Aleuts. Unlike the Tlingits, who placed their dead in trees, the Aleuts buried their dead in volcanic caves. Their culture was quite different in artistry, reflecting their comfortless and barren islands. They, too, wove fine grass baskets and wore beautiful hooded jackets of bird skins with feather inserts. They were a gentle people who lived in crude sod houses, called *barabaras,* that were built partly underground. Early Russian hunters massacred many Aleuts and made others veritable slaves. Their population, decimated by cruelty and greed and ravaged by disease, dropped from 20,000 to less than 2,000 between 1791 and 1867. Today, only about 4,000 natives live in the Aleutians, subsisting by fishing and seal hunting in the Bering Sea.

Eskimos. The term Eskimo means "eaters of meat." Like their Aleut cousins, Eskimos are believed to be descendants of migrant Asians. They probably crossed the Bering Strait after it ceased being a land bridge, between 3,000 and 8,000 years ago. Today's Eskimos live in such settlements as Kotzebue, Shishmaref, Wainwright, and Barrow, but their forebears were nomads, constantly roaming the Arctic in search of meat. In language and physiognomy they are similar to the Eskimos of Siberia, northern Canada, and Greenland but quite different from the darker skinned Athabascans and the coastal Tlingits.

About 18,000 Eskimos live in Arctic Alaska, compared with about 11,000 in Canada and 22,000 in Greenland, but only about 1,200 live in Siberia. Good humour and a hospitable attitude have helped them adjust to their rigorous environment. The Alaskan Eskimo formerly lived off the land and resided in tents in summer, settling into semipermanent driftwood timber and sod huts in winter. Most now live in permanent frame houses. Their early religion centred around the concept of a soul within a three-faced spirit system depicting future life, a basically optimistic view of present life, and the ever present possibility of evil. The Eskimo attributed souls to all animate objects, such as animals. Their beliefs were similar to those of the Aleuts and Tlingits, except that they believed that human souls lived in the sea after death. When an Eskimo died, his body was sewn into a caribou skin and left under a pile of stones. As in ancient Egypt, Eskimos put food and weapons by the body of their deceased for use in the afterworld. Today, as is true of the Aleut and Indian, most Alaskan Eskimos have become Christian, and aboriginal religious practices, except in a few rare instances, are extinct. Eskimos, who are skilled carvers of walrus ivory, wood, and steatite, or soapstone, have few equals as artistic craftsmen. Unlike other native groups in Alaska, they continue their native crafts vigorously, many making a living plying their artistry for the tourist trade.

Eskimo settlements

Tinneh. The Athabascan Tinneh were hunters and salmon fishermen along the Yukon River and its tributaries. Some work part time on construction crews in this economically developing area. The Tinneh tribes never had clans or much cohesiveness, and they were assimilated more easily into contemporary life. Many of them migrated to the coastal areas and intermingled with the Tlingits, bringing the two groups close together in character and culture.

Demography. The wave of youthful immigrants attracted from the "South 48" in the 1960s suggested a faith in the burgeoning opportunities in a still-frontier environment. The discovery of oil fields and the emergence of Alaska as an international air crossroads added impetus to the influx of the 1940s and 1950s, a period of new settlement and expansion that raised the population from 70,000 to 250,000. Prior immigration—the first wave of which occurred in the decade before World War I as an aftermath of the gold rush—was a response to Alaska's initial concentration on its mineral, fish, and timber resources. Of today's population about one-seventh, or 43,000, are Eskimos, Aleuts, and Indians. The remaining citizenry includes about 80,000 military per-

sonnel and descendants, and a melting pot of mixed American, Russian, Japanese, Chinese, Filipino, and other nationalities.

Population distribution

In addition to the 40 percent of Alaskans living in the southern interior basins around Anchorage, most of the remainder live in the Panhandle region, where Juneau is the major city and the administrative centre of the state, and in the interior plains around Fairbanks. Tiny pockets of people are scattered in small villages, the most sparsely occupied being the Arctic plains, the Bering shores, and the Aleutians. Only about 20 percent of the white population were born in Alaska, and about 100,000 persons have immigrated since 1950. Many frontier conditions persist: a male-to-female ratio of 5 to 1 in 1910 was reduced by 1970 to near equality, but bars are frequently as numerous as churches of all Christian denominations.

THE STATE'S ECONOMY

The Alaskan economy is conditioned strongly by the state's frontier stage of development and its inadequate tax base for state and municipal growth. High costs of labour and transportation tend to discourage outside investment. The beginning development of the state's natural resources can assist markedly in the transition from a federal military to a private economic base that will increase tax revenues.

Sectors of the economy. *Government.* From 1940 to 1960 the federal government invested nearly $2,000,000,000 in the development of military bases in Alaska. Nothing else in Alaska's history has produced such long-term results, bringing thousands of new residents into the territory and creating jobs and a vast array of transportation and communications facilities extending to remote corners of the state. The defense installations continued to add much to Alaska's economy.

Agriculture. More than 3,000,000 acres of tillable land are available for farming, but much clearing has yet to be done. Most acreage is near Anchorage and on the Kenai Peninsula, though there is some near Fairbanks, and stock ranching is practiced on Kodiak and Unimak islands. As a result, most foods must be imported, tremendously increasing costs of living. Closure of the Homestead Act, ending settlement of the native land claims issue, has further curtailed development of new land. In spite of a short growing season, the long hours of summer sunlight are adapted to the successful production of wheat, oats, rye, barley, and hay and all cool-climate vegetables.

Fishing. Alaska's most constant source of revenue, fishing is rivalled only by mining, timber, and petroleum. Its occurrence concentrated in waters off the southern coasts, salmon is the main product. The centre of the world's salmon-packing industry is at Ketchikan, and it flourishes also on Kodiak Island and at Bristol Bay ports in the southern Bering Sea. The industry since about 1950 has represented about $100,000,000 annually. Fleets also bring in quantities of halibut, herring, sablefish, Dungeness crab, king crab, and shrimp. A serious threat to fish conservation and a source of continuing international friction in the North Pacific has been the unregulated incursion of Soviet and Japanese fishing vessels into Alaska waters.

Forest products. Most of Alaska's timber resources are in the Tongass and Chugach national forests, in the Panhandle and the south coast, respectively. Pulp is an important industry in Ketchikan and Sitka. Lumbering and standard sawmill operations continue at about 40,000,000 board feet per year. In 1969, forest products accounted for about $101,000,000 and represented a great potential for further development.

Furs. Pribilof sealskins represent nearly 60 percent of the state's $8,000,000 annual fur production. Other furs, largely from controlled farms, are processed as well. A new industry in the 1960s was the production of reindeer hides from a herd on Nunivak Island that was managed by the Alaska Native Association.

Hydroelectric power. Alaska's immense waterpower reserve is virtually untapped. The largest project to date is at Eklutna, near Anchorage. A new hydroelectric development near Juneau will deliver power to the Panhandle area in the early 1970s. In many communities, diesel and coal plants produce much of the required municipal power.

Mining. Petroleum was first extracted and refined between 1917 and 1933, but the development of the Kenai oil field in 1961 brought the petroleum and natural gas industry to the top of Alaska's mineral production. Oil seeps were known as early as the 1880s in the North Slope region, which is expected to be of major economic importance.

Alaska's gold production declined drastically from the 1940s through the 1960s because of rising labour and transportation costs. In the 1970s only a few small operations still remained. Copper mining as a major industry petered out with the closing of the Kennecott Mine in 1938, though a few new prospects elsewhere show promise. Coal remained an important industry, with the mines yielding some 700,000 tons annually. Known coal reserves amount to more than 100,000,000,000 short tons, of which 70 percent is low-grade fuel. A new and important activity is the extraction of sand, gravel, and clay to serve the construction industry.

Since 1880 well over $1,000,000,000 in hard-rock ore minerals have been mined in Alaska, gold, copper, and silver accounting for 95 percent of this. Prospecting continues, with modern scientific technology and aerial exploration. The areas of maximum mineral potential lie in the Panhandle, the Chugach and Alaska ranges, and the Seward Peninsula.

Tourism and transportation. Alaska is experiencing an upsurge of tourism. Travellers can now cover large areas by airplane and road. The influx is, in part, the result of the 500-passenger, 100-car ferries that began operation as the Alaska Marine Highway. One ferry system connects Kodiak with mainland Seward and the Alaska Railroad; another, Cordova and Valdez.

High costs of transportation continue to sap Alaska's economic development, largely because the major transportation links, both internal and external, are by air, which provides the fastest way to cross Alaska's great distances and formidable terrain. Two dozen airlines serve Alaska, with daily service from the "South 48" and Canada, Europe, Hawaii, and Japan. Nearly 400 airfields, seaplane bases, and emergency strips are in use throughout the state, and few villages are without service at least by bush pilots. About 4,500 miles of roads, most of them surfaced, are currently in use. The Alcan Highway and its Haines cutoff connect Alaska's internal road network to the outside and provide relatively easy access for tourists. Some road construction was so severely damaged by the 1964 earthquake that it was abandoned. In 1969 a 356-mile haul road from Fairbanks to Prudhoe Bay was completed. It connects with the existing highway system to provide a winter-only overland route from the ice-free southern ports to the Arctic Ocean.

Transportation networks

The government-owned Alaska Railroad, the only operating railroad in the entire state, runs for 498 miles, linking Seward, Anchorage, and Fairbanks. Ocean shipping connects Seattle, Vancouver, and the trans-Canada railhead of Prince Rupert to towns in the Panhandle and westward to Cordova, Valdez, Seward, and Kodiak. Ocean vessels also run during the ice-free midsummer months to Nome and Barrow and the oil regions of the Arctic coast. A natural gas pipeline, completed in 1961, runs between the Kenai gas fields and Anchorage.

In the mid-1950s an Alaska Communication Cable was installed between Seattle and Alaska. Radio telephones now connect all interior communities.

ADMINISTRATION AND SOCIAL CONDITIONS

Governmental structure. The state constitution was adopted in 1956. A 1971 amendment set the minimum voting age at 18. The governor and secretary of state are the only executive officers and are elected by the same vote. The 40-member House of Representatives and 20-member Senate are elected for terms of two and four

years, respectively. The Supreme Court has a minimum of three members. There are four district courts. A single federal district court, replacing the territorial courts, sits alternately in Juneau, Anchorage, Fairbanks, and Nome.

Public financing is through various personal income and business taxes. As a part of the Act of Admission, Congress granted Alaska certain revenues from the sale of furs and of federal lands.

State and borough governments have difficulty in providing the usual range of services because of the limited extent of the economy, an above-average unemployment rate, and a tax base that is not yet viable. The vast area and the difficult terrain increase these problems, but it is expected that these problems will be eased during the 1970s as the utilization of natural resources increases.

Responsibilities for native peoples. Activities on behalf of the native population have turned largely on the U.S. Bureau of Indian Affairs (BIA), which assists Alaska's natives in achieving economic and social self-sufficiency. Despite a number of helpful programs, most of Alaska's natives have remained at the bottom of the economic and social ladder—suffering from unemployment, low income, and poverty. The native peoples were educated first by missionary groups, though by the time of statehood the BIA had assumed most of the responsibility. Funds are provided for vocational training and the development of job opportunities and for welfare, social work, and medical and health needs. The BIA also assists natives to organize their villages under federal and state laws, etc. Some oil revenues from native lands have been applied in self-help programs, and settlement of the native land claims issue should also improve their economic plight.

Education. Education is compulsory through the eighth grade or until age 16 and is administered by a state board and commissioner of education. Many federal schools operate on military bases. The University of Alaska, founded in 1917 and located at College, near Fairbanks, is the only state university, but there are community colleges in Anchorage, Juneau–Douglas, Kenai–Ketchikan, Palmer, and Sitka. Alaska Methodist University in Anchorage and Sheldon Jackson Junior College in Sitka, a two-year college for native peoples, are the only other institutions of higher learning.

Welfare and health. The elderly, dependent children, and the blind are aided by the Department of Public Welfare, and there is a special fishermen's fund. The state also operates a psychiatric hospital, a tuberculosis sanatorium, a youth camp, and a prison.

Medical and health clinics and hospitals available to the public at large are provided by municipal and borough governments or private agencies, or run as church-operated facilities. Health standards have been raised markedly since 1950 through visits by U.S. Public Health Service nurses and doctors to the remote villages. The large number of airfields, the radio communications network, and the extensive use of bush pilots operating through the state make it possible for most persons, even in the remote villages, to reach medical facilities when there is serious need.

CULTURAL LIFE AND NATURAL ENVIRONMENT

Cultural preservation. Alaska's past, including the arts and the crafts of its native peoples, figures heavily in Alaskan culture. Juneau has the state historical library and the state museum. The university has a large museum, as do other communities, including Sitka, Haines, Valdez, and Skagway. Eminent Alaskan artists have included both whites and Eskimos. Native carvings in ivory and wood are well known, and the nearly lost art of totem carving has been revived in part through private and public stimulus.

Natural preservation. Wildlife refuges and ranges abound throughout Alaska, with more than 19,000,000 acres managed by the United States Fish and Wildlife Service. The federal Bureau of Land Management also holds about 25,000,000 acres for waterpower development.

One national park and three national monuments comprise another 7,500,000 acres. Mt. McKinley National Park (established 1917) has a diverse abundance of wildlife, including brown and grizzly bears, caribou, and moose. Katmai National Monument (1918), on the Alaska Peninsula, includes the Valley of Ten Thousand Smokes, an area of active volcanoes that in 1912 produced one of the world's most violent eruptions. Glacier Bay National Monument (1925) features magnificent fjords, as well as glaciers that are retreating slowly. Tiny Sitka National Monument (1910), with a large totem pole collection, commemorates the stand of the Tlingits against early Russian settlers. The Tongass and Chugach national forests in the southeast and south, respectively, complete the picture of federal public land reserves. The Department of the Interior continues to study the need for withdrawing further regions from public domain into reserves, but the native land claims and strong conservationist pressure made the situation uncertain in the early 1970s. How such areas, though they are already part of the state's tourist attraction, will figure in the future cultural and recreational life of Alaska remains unclear.

National wilderness preserves

Environmental use and study. The sporting industry, including guide and outfitter services, boat charter, and the like, continues to be a colourful activity. Alaska provides America's only significant Arctic wilderness, and much research is done in glacier, mountain, tundra, and polar oceanography fields by federal, state, and private agencies. These projects, too, bring income to the state. The University of Alaska carries out extensive research on Arctic problems through its Geophysical Institute, Institute of Marine Sciences, Institute of Arctic Biology, and other groups. At Barrow the United States Naval Arctic Research Laboratory opened in the 1940s to conduct Arctic research, including sea-ice and oceanographic studies. The U.S. Bureau of Commercial Fisheries operates a research laboratory at Auke Bay. Since 1946 an international glaciological and environmental research and field-sciences training program has been conducted on the Juneau Icefield.

Prospects. To visitors from America's other states, Alaska's contrasts and uniqueness make it virtually a foreign land, for nature has endowed it with abundant natural resources and a special beauty of land and character of people. Among the factors that will influence the development of Alaska's resources in future years are the economics of national and international supply and demand, the re-evaluation of national ideals and priorities, and the shifting character of political and social institutions. The development of one resource can clearly affect, even destroy, the use of another. This type of competition among resources to satisfy the uncertain demands of a pluralistic public will require application of the best concepts in general systems management.

The balance between conservation and exploitation

Much of this northern land remains virgin territory, and the ideas of land use alongside conservation are not yet too late. What Alaska does will be watched with great interest, for it symbolizes the long-range concern in the environmental challenge to all nations. The success of Alaska's environmental task and the effective solution of related economic, political, and social problems have important implications for the balance between man and nature that will be required eventually in every corner of the globe.

BIBLIOGRAPHY. H.H. BANCROFT, *History of Alaska, 1730–1885* (1890, reprinted 1959), definitive narrative historical accounts of early explorations; JUDY BRADY, "Native Land Claims," *Institute of Social, Economic and Government Research*, vol. 4 (1967), a review of the land freeze of 1967 with map of the disputed areas; *The Alaska Survey and Report* (annual), an all-inclusive book on Alaska; S.C. CARRIGHAR, *Icebound Summer* (1953) and *Moonlight at Midday* (1958), books on Alaskan living and the nature of the land and its people; R.A. COOLEY, *Alaska: A Challenge in Conservation* (1966), overview of the status and future of land development and use and the need for a conservation ethic; GREGG ERICKSON, "Alaska's Petroleum Leasing Policy," *Institute of Social, Economic and Government Research*, vol. 7 (1970), discussion of petroleum policies and pertinent legislation and administration; W.E. GARRETT, "Alaska's Marine Highway," *Nat. Geog. Mag.*, 127:776–819 (1965); J.L. GIDDINGS, *Ancient Men of the Arctic* (1967), archaeology of Alaska and the Arctic; ERNEST GRUENING, *The State of Alaska* (1954), au-

thoritative text on the political and economic scene in Alaska in the decade before statehood; C.C. HULLEY, *Alaska: Past and Present*, rev. ed. (1970), concise general history from the Russian days to present; E.L. KEITHAHN, *Monuments in Cedar* (1963), professional account of Tlingit cultural heritage and Indian art of the Alaska Panhandle; M.M. MILLER, "Alaska's Mighty Rivers of Ice," *Nat. Geog. Mag.*, 131:194–217 (1967), survey of Alaska's spectacular glacier coast and explanation of the intriguing pattern of glacier variation in historic time; GEORGE W. ROGERS, "Alaska's Economy in the 1960's," *Institute of Social, Economic and Government Research*, vol. 7 (1970), a review of Alaska's economic picture; U.S. DEPT. OF THE INTERIOR, *Mid-Century Alaska* (1957) and *Natural Resources of Alaska* (1967).

(M.M.M.)

Alaskan Mountains

A northwestward continuation of the western cordillera, or Rocky Mountains, and the Pacific coastal ranges of North America, the Alaskan mountains give their state a rugged and beautiful terrain across nearly its entire expanse. They include the highest peak in North America. The ranges are characterized both by glaciers and by continuing volcanic activity. Little explored for vast stretches, the mountains are believed to contain, or lie close to, immense untapped mineral resources whose exploitation is hampered by the terrain and the climate. For details on related subjects, see the articles YUKON RIVER; ALASKA.

Brooks Range

Major groups. The most northerly of the three major Alaskan mountain groups are the Brooks Range and the Arctic foothills, extending the Rocky Mountains in an east–west arc from the Canadian border across northern Alaska. Central Alaska is characterized by highlands and basins drained by the great Yukon and Kuskokwim drainage systems. This area has been likened by some to a moister version of the arid region known as the Basin and Range Province (*q.v.*)—an area which takes in nearly all of Nevada and portions of all contiguous U.S. states —although here the precipitation and resulting vegetation have rounded the rugged topographic expression of mountains and valleys.

Southern Alaska is dominated by an arc of mountain ranges encircling the Gulf of Alaska and the northeastern Pacific in a broad sweep. This Pacific mountain province is subdivided into several groups. The large Alaska Range merges southwestward into the Aleutian Range and the Aleutian Islands. Separated from the Alaska Range by the Talkeetna and Wrangell mountains, the Kenai–Chugach mountains border the Gulf of Alaska and merge, to the south and east, with the St. Elias Mountains at the Canadian border. These, in turn, merge with the mountains of the coastal ranges, which form most of the Alaskan Panhandle.

Scientific study. Alaska has been mapped topographically from aerial photographs, at a relatively high level of detail: the underlying geology, however, has been mapped only sketchily. The U.S. Geological Survey, as part of the general reconnaissance mapping program of the late 19th century and the first half of the 20th century, did most of the geological work. Since the 1950s, areas of potential mineral and transportation value have been mapped on a more detailed basis by state and federal agencies. Various university research parties contributed to the data: much geological mapping work was also done by private mining and petroleum companies, but this work was rarely made public.

The physical environment. *Northern mountains.* The Arctic foothills, just north of the Brooks Range, consist of low east–west trending ridges and rolling plateaus with irregular isolated hills. They rise from 600 feet (200 metres) in the north to 3,500 feet (1,100 metres) in the south. Except for the east-flowing Colville River, most drainage is northward. The entire area is underlain by permafrost, a permanently frozen, rock-hard soil, and no glaciers are present. This region is as structurally complex as the higher Brooks Range immediately to the south, but formed of less resistant rocks. The youngest rocks, in the northern section, are sediments from the Cretaceous Period (136,000,000 to 65,000,000 years ago). They are folded, faulted, and overthrust toward the north. Sediments from the Devonian Period (beginning 395,000,000 years ago) to the Cretaceous form the southern section of the foothills. They are tightly folded and also overthrust northward.

The Brooks Range is the highest mountain range within the Arctic Circle. It includes groups of mountains extending about 600 miles (1,000 kilometres) from the Canadian border to the Chukchi Sea. These form several groups of mountains, individually named. Average altitudes range from 3,000–4,000 feet in the west to 5,000–6,000 feet in the east, with a high point of 9,239 feet (2,816 metres) in Mt. Michelson. The entire area has been glaciated, as is evidenced by the rugged topography. Several small glaciers are still present in the east, fewer in the west.

Average altitudes

The range forms the drainage divide between waters flowing northward into the Arctic Ocean, those flowing westward into Kotzebue Sound, and those flowing south into the Yukon and its tributaries and emptying into the Bering Sea. Several major rivers have cut back into the range to form low passes, the best known being Anaktuvak Pass, at an elevation of 2,200 feet in the central part of the range. The Dietrich River Pass has been suggested as a corridor connecting the oil-producing areas of the north slope with interior Alaska and the south.

The backbone of the range is composed of sedimentary and metamorphic rocks—respectively, rocks formed from deposits of various organic or inorganic materials and rocks heavily compacted and made more crystalline by the action of heat, pressure, or water. They were formed in the Paleozoic Era (from 570,000,000 to 225,000,000 years ago). Younger sedimentary rocks, of the Permian Period (from 280,000,000 to 225,000,000 years ago) and the Mesozoic Era (from 225,000,000 to 65,000,000 years ago) flank the range. The mountains were lifted by major upward foldings in the Earth's crust, called orogenies, beginning in the Late Jurassic Period (190,000,000 to 136,000,000 years ago). The process persisted in periodic uplifts throughout the Cretaceous and into the early part of the Tertiary Period (from 65,000,000 to 25,000,000 years ago). It was again strongly deformed and uplifted in the late Tertiary. The folding, faulting, and major overthrusting toward the north during these orogenies were modified by erosion and glaciation. Most of the area is underlain by permafrost. There are a few lakes in rock basins.

Central ranges. The mountains of central Alaska, extending from the Canadian border to the Bering Sea, are lower than the ranges to the north and south. They are drained almost entirely by two river systems, the Yukon and the Kuskokwim. The intricately dissected uplands are divided into three areas, the eastern highlands, the western highlands, and the Seward Peninsula.

The eastern highlands, consisting of several separate chains, have an average altitude of 4,000–5,000 feet, with a few mountains rising 1,000–2,000 feet above these uplands. Some peaks are as high as 6,800 feet (2,100 metres) in the east. No glaciers are present in the region, and permafrost is discontinuous. The underlying rocks are highly deformed metamorphic and sedimentary and of volcanic origin dating from the Precambrian (4,600,000,000 to 570,000,000 years ago) and Paleozoic eras. The higher parts are commonly composed of small, resistant segments of granite that were forced, or intruded, into the other rocks while in a molten state.

The western highlands, also subdivided into several smaller groups, are somewhat lower, rolling, and monotonous mountains, with northeast-trending ridges. Numerous isolated, nearly circular groups of mountains rise above these ridges. The rocks include tightly folded Paleozoic and Mesozoic sediments and volcanics and Tertiary intrusions. Although no glaciers are now present, the Ahklun Mountains are the largest glaciated area in interior Alaska, and the Wood River–Tikchik region has beautiful parallel glacial lakes and is considered one of the most scenic areas in the state.

Scenic glacial lakes

The upland area of the Seward Peninsula, which is only a few miles across the Bering Strait from the Soviet Union, is made up of broad, convex hills and ridges, with

an average altitude of about 2,000 feet (600 metres), and surmounted by more rugged mountain groups. A few peaks rise above 3,000 feet (900 metres), the highest 4,720 feet (1,440 metres), in the Kigluaik Mountains in the southwestern part of the peninsula. The entire area is underlain by permafrost. The rocks are early Paleozoic metamorphics, Cretaceous sediments, and intrusions of Mesozoic igneous, or once-molten, rock.

Pacific coastal ranges. The Alaska–Aleutian ranges, together with the Talkeetna and Wrangell mountains, the interior part of the St. Elias Mountains, and the coast ranges of southeastern Alaska, are geologically analogous in some respects to the Sierra-Cascade ranges of the western United States, of which they are regarded as the northwestern continuation.

The Alaska–Aleutian ranges extend in a great belt from the Canadian border, at first northwestward and then south and west to the Aleutian Islands. These islands comprise a chain of mountainous volcanic peaks, the crest of a huge submarine ridge 1,500 miles (2,400 kilometres) long and 20 to 60 miles wide, that rises 11,000 feet above the sea floor. Separating the Bering Sea from the Pacific Ocean, these islands stretch in a 1,100-mile-long arc from the Alaska Peninsula to Attu Island at the western end of the chain. With the Aleutian Range on the southeast side of the peninsula, they form a long, narrow belt of peaks for 1,600 miles from Attu to Mt. Spurr opposite the city of Anchorage. Nearly 80 volcanoes have been mapped in this chain, 47 of which have been reported active since 1760. Marine sedimentary rocks and granitic intrusions from the Mesozoic and Cenozoic eras (from 65,000,000 years ago), as well as Cenozoic volcanic rocks, form this range. Pavlof and Shishaldin volcanoes rival Japan's Fujiyama in the beauty of their symmetrical cones. Enormous calderas—collapsed craters marking past volcanic eruptions—are present. In 1912, one of the most spectacular of historic eruptions occurred when a volcano in the Aleutian Range erupted, forming the Valley of Ten Thousand Smokes (*q.v.*), in what is now Katmai National Monument. Mt. Spurr last erupted in 1953 and remains active, as do Mt. Redoubt and Mt. Augustine, both of which erupted during the 1960s.

The Alaska Range merges with the Aleutian Range between Iliamna Lake and Mt. Spurr and extends in an arc some 600 miles to the Canadian border at the southeast end. This great mountain wall is about 120 miles wide near Mt. Spurr and about 60 miles wide in the vicinity of Mt. McKinley. It gradually narrows to about 30 miles near the Canadian border. The range is extremely rugged and glaciated. The crest of most of the range averages between 7,000 and 9,000 feet with fewer than 20 peaks exceeding 10,000 feet. Mt. McKinley is 20,320 feet (6,194 metres) in altitude, the highest point in North America. Four great individual mountain masses dominate the group, divided by several low passes and river valleys, several of which provide routes of travel across the mountains. Great valley glaciers radiate from these mountain masses, except in the relatively low eastern sectors where the few glaciers are small. The Alaska Range is the drainage divide for rivers flowing north into the Yukon system and those flowing south into the Pacific Ocean.

Most of the Alaska Range is underlain by large deeply buried granitic intrusions (known as batholiths) into Paleozoic and Mesozoic sedimentary rocks that have been metamorphosed and deformed. The youngest deposits, of Cenozoic age, are only moderately deformed beds in which deposits of coal are present. Very long valleys and low passes, following the course of geological faults, lines of weakness in the earth's crust, extend almost parallel to the entire length of the range. The faults were probably formed near the close of the Mesozoic Era.

Between the Alaska Range and the coastal ranges lie the Talkeetna Mountains and, to the east of them, the Wrangell Mountains. The Talkeetnas, an oval, rugged area about 100 miles from north to south and 70 miles from east to west, is a compact group of radial ridges averaging from 6,000 to 8,800 feet in altitude. There are only a few low passes. The higher ridges are sharp-crested, but on the western flank, as the average height decreases, the ridge tops have been rounded and subdued by the glacial action of the Pleistocene Epoch (2,500,000 to 10,000 years ago). Small glaciers 5 to 15 miles long are still present at the heads of many valleys. The Talkeetnas consist of an igneous batholith intruded during the Mesozoic Era, with flanking Paleozoic and Mesozoic sediments and volcanics. Some Tertiary volcanic flows are present in the southeastern part.

The Wrangell Mountains, 100 by 70 miles, lie on the Pacific between the Alaska Range to the north and the Chugach Mountains to the south. They rise more than 10,000 feet above the Copper River Plateau. A large ice cap covers most of the high mountains and feeds large valley glaciers. This oval group is composed of composite volcanic mountains, which are resting on metamorphosed Paleozoic and Mesozoic sedimentary and igneous rocks. Some granitic masses intrude the Mesozoic rocks. Several volcanoes are at altitudes higher than 12,000 feet, the highest, Mt. Blackburn, at 16,390 feet (4,996 metres).

Encircling the Gulf of Alaska, from Kodiak Island in the west to Baranof Island in the southeast, is another bow-shaped, or arcuate, mountainous belt, 800 miles long and 20 to 110 miles wide. Kodiak Island and the Kenai and Chugach mountains form a rugged barrier along the northeast margin of the Pacific Ocean. The mountains of the Kodiak Island group, at average altitudes of 2,000 to 4,000 feet, trend northeasterly. Glaciation over most of the area has produced a very irregular coastline, but no glaciers are now present.

The average altitude of the Kenai Mountains is 3,000 to 5,000 feet, whereas the Chugach group to the north and east average 7,000 to 8,000 feet. The highest peaks are in the sharp bend of the arc, where Mt. Marcus Baker rises to 13,176 feet (4,016 metres). The mountains are extremely rugged and bear on their crests the last remnants of the ice fields from the Ice Age, the Sargent and Harding ice fields in the Kenai Mountains and the Bagley Ice Field in the eastern Chugach Mountains. Numerous long, spectacular glaciers descend from the crests of these mountains. The St. Elias Mountains and the Kenai–Chugach Mountains have the most extensive system of valley glaciers in North America.

Ice Age remnants

The arcuate form of the Kodiak–Kenai–Chugach belt is closely paralleled by the structural trends of the underlying rocks. The Mesozoic rocks, primarily marine sediments and volcanics, have been intensely metamorphosed, folded, faulted, and intruded by small-to-moderate-size igneous bodies. Along the coastal areas, in Kodiak and in the eastern part of the Chugach bordering the Gulf of Alaska, softer Tertiary marine sediments flank the higher mountains.

The Alaskan part of the St. Elias Mountains is topographically continuous with, and geologically similar to, the Chugach Mountains, and joins the southeastern panhandle to the rest of the state. The St. Elias chain, among the highest coastal mountains in the world, is about 200 miles long and has a maximum width of about 100 miles. The chain merges with the Chugach group near the Alaska–Yukon border and with the Wrangells and parts of the Alaska Range to the northwest.

The massive, blocklike St. Elias Mountains rise from a myriad of narrow ridges and peaks at 8,000 to 10,000 feet in height to 14,000 to 19,000 feet. Mt. Logan, 19,850 feet (6,050 metres), Mt. St. Elias, 18,008 feet (5,489 metres), and a dozen other peaks exceed in altitude the highest mountains elsewhere in the United States. The average altitude of the ice fields blanketing the lower regions is 3,000 to 7,000 feet. Because of the difficulty of access, the geology is little known. The high mountains are probably underlain by schist, a highly crystalline igneous rock, and by granitic masses. Elsewhere, the mountains consist mainly of sedimentary and volcanic rocks, from the Paleozoic and Mesozoic eras, that have been intensely metamorphosed, folded, and faulted. One great arcuate fault system, the Lynn Canal–Chatham Strait trough, bounds the St. Elias Mountains on the northeast and continues southward into southeastern Alaska.

Southeastern Alaska is a continuation of these Pacific coastal mountains, a rugged, scenic island archipelago and neighbouring ranges on the mainland. To the east the coast ranges rise 5,000 to 7,000 feet, with higher peaks at 8,000 to 10,000 feet. The summits of these mountains coincide in part with the Canadian boundary. Extensive ice caps feed valley glaciers, some of which still extend to the sea. These mountains are underlain by the massive granitic rocks of the Coast Range Batholith, intruded in late Mesozoic. The island archipelago ranges in altitude from 2,000 to 3,500 feet in the southern Prince of Wales Mountains, to more than 4,000 to 7,500 feet in the outer Chilkat–Baranof Mountains. The glaciated islands have a rugged coastline indented by many fjords. The island archipelago is composed of northwest-trending belts of Paleozoic and Mesozoic sedimentary and volcanic rocks and metamorphic rocks that have been intruded by large areas of igneous rocks and cut by many faults.

Vegetation and animal life. Vegetation in the higher altitudes of the mountain ranges, when they are not barren or covered by ice, may be a ridge type of tundra, or level, treeless plain, that consists of lichens, grasses, and weeds above the timberline, the upper limit of tree growth. Dwarf species of willows and alders are also found above the timberline, and grasslands occur in the Aleutians, on the Alaska Peninsula, and the south slopes of the Alaska Range. In the interior forests, the timberline in many places is less than 600 feet, although it is generally between 1,500 and 2,000 feet. The interior forests of white spruce and birch, with mixtures of balsam poplar, occur mostly on the lower slopes of the larger river valleys north of the Alaska Range. In the coastal forests, the timberline ranges from about 2,000 to 3,000 feet in southeastern Alaska to approximately 1,000 feet in the more northerly and westerly sections.

Tundras and timberlines

The Tongass National Forest encompasses southeastern Alaska, and the Chugach National Forest covers the lower parts of the Chugach, Kenai, and Kodiak mountain areas. The vegetation in the lower altitudes is a hemlock–spruce coastal forest. Western hemlock and Sitka spruce extend to about 1,500 feet in elevation, and subalpine species to at least 2,500 feet. The Tongass National Forest is composed largely of western hemlock, with a much smaller number of Sitka spruce and a sprinkling of western red cedar and Alaska cedar. The Chugach National Forest, in the Prince William Sound section, is predominantly a hemlock forest with some Sitka and white spruce and occasional cottonwoods and paper birches.

Several big-game animals are native to the Alaskan mountains. Both brown and grizzly bears inhabit the coastal ranges, the Alaska Range, and the interior. The black bear, absent from the Alaska Peninsula, is found in three-fifths of Alaska. Dall mountain sheep, the only wild, white sheep in existence, live high in the mountains from the Kenai Peninsula to the Brooks Range. The mountain goat is found in the Chugach, Kenai, and Talkeetna mountains. Moose, though living mostly in the lower altitudes, range into the higher country from the Alaska Peninsula to the Brooks Range. Barren-ground caribou range widely over high plateaus and mountain slopes from the Alaska Peninsula to the Arctic regions. Sitka deer are found in southeast Alaska and in parts of Prince William Sound. A number of smaller fur-bearing animals are also native to the mountains.

BIBLIOGRAPHY. H. WILLIAMS (ed.), *Landscapes of Alaska: Their Geologic Evolution* (1958), a well-written description of all parts of Alaska; C. WAHRHAFTIG, "Physiographic Divisions of Alaska," *Prof. Paper U.S. Geol. Surv. 482* (1965), a detailed summary description of all areas in Alaska, with good maps and illustrations; C.J. RHODE and W. BARKER, "Alaska's Fish and Wildlife," *Circ. Fish Wildl. Serv., Wash. 17* (1953), a handbook summary written for laymen; R.F. TAYLOR and E.L. LITTLE, JR., "Pocket Guide to Alaska Trees," *Agric. Handb. Forest Serv. U.S. 5* (1950), a good field guide to tree identification.

(R.Sc.)

Alba, Duke of

A Spanish soldier and statesman, Fernando Alvarez de Toledo, the 3rd duke of Alba, achieved fame for his conquest of Portugal and notoriety for his tyranny as governor general of the Netherlands. In Protestant countries Alba's name became a byword for cruelty and religious tyranny. Outside Spain he has never been forgiven for his disregard of legality, for his policy of terror, and for the outrages committed by his troops in the Netherlands and Portugal. Of the responsibility for these acts he cannot be exonerated, nor were they universally accepted even by Catholic opinion in the 16th century. In Spanish history Alba is important as a representative of the old nobility, independent and proud of its rights and privileges yet willing to serve as the champion and defender of an absolute monarchy. Alba was born at Piedrahita (Avila) on October 29, 1507, of a rich family with a long record of service to the kings of Castile. In 1524 he joined the Spanish forces fighting the French at Fuenterrabía and so distinguished himself that he was appointed governor of the town after its capture. Subsequent campaigns made him the most thoroughly professional military commander of his age. He insisted on rigorous training and discipline for his troops and developed the tactical use of firearms. He was a master of logistics, and his greatest asset was an unshakable self-confidence that enabled him to resist the rash counsels of his more impetuous officers. He commanded a part of Charles V's army in the successful expedition against Tunis in 1535, and in 1546–47 he commanded the imperial armies against the German Protestant princes of the Schmalkaldic League. By his victory at Mühlberg (April 24, 1547) Alba placed the emperor Charles V at the pinnacle of his power. Alba was made commander-in-chief of the imperial forces in Italy in 1552 and, after the succession of Philip II of Spain, viceroy of Naples (1556). In the last phase of the Franco-Spanish war in Italy, he outmanoeuvred the Duke of Guise and forced Pope Paul IV to come to terms with Spain (1557).

By courtesy of the Musees Royaux des Beaux-Arts, Brussels



Alba, oil painting by Sir Antony More, 1549. In the Musées Royaux des Beaux-Arts, Brussels.

Adviser to Philip II

After the Peace of Cateau-Cambrésis (1559), Alba became one of Philip II's two leading ministers. Charles V, an excellent judge of character, had recommended Alba to his son Philip in a secret testament of 1543 as a reliable adviser in all military matters and in affairs of state but otherwise not to be trusted, for he was enormously ambitious, wished to command everything, and would employ any means, including the use of women, to achieve his ambitions. Philip II therefore never fully trusted Alba. He did, however, summon him regularly to the Council of State, where, in opposition to the other of the King's chief advisers, Ruy Gómez de Silva, he pressed for a vigorous foreign policy.

As early as 1563 Alba advised the King to cut off the heads of the leaders of the aristocratic opposition in the Netherlands. But if this should not be immediately possible, he remarked, the King should dissemble now and

execute them at a more opportune moment. In 1565 Philip sent him, together with his queen, Elizabeth of Valois, to meet Elizabeth's mother, Catherine de Médicis, regent of France. Alba managed to hold his own against that virtuoso politician, blocking Catherine's efforts to arrange a Spanish marriage for her son to which Philip II did not wish to commit himself. Understandably, he did not succeed, in his turn, in committing Catherine to a more active anti-Huguenot policy with Spanish assistance. The later Protestant accusation (and Catholic boast) that at Bayonne he and Catherine planned the St. Bartholomew's Day Massacre of Protestants in 1572 has no basis in reality.

Following the popular movements of 1566, Philip sent Alba to the Netherlands with a large army to punish the rebels, root out heresy, and re-establish the King's shaken authority (August 1567). Alba arrested the Lamoraal, graaf van Egmond and Filips van Montmorency, graaf van Hoorne, the rather halfhearted leaders of the opposition, and set up a new court, the Council of Troubles (soon to be known as the Council of Blood). This court set aside all local laws and condemned some 12,000 persons for rebellion, many of whom, however, had fled the country. He allowed himself to be involved in a trade war with England that caused great damage to Netherlands commerce. Worst of all, he mishandled the plan to place his government on a stable financial basis, independent of the estates. He proposed a 10 percent tax on all sales (the "tenth penny") and a 1 percent property levy. But the States General would agree only to the property tax and made counter offers, in place of the "tenth penny." In the face of the opposition of the lower classes and the clergy, Alba had to modify the tax progressively. In the end it was never collected. While Alba's "tenth penny" certainly helped to rouse the country against Spain, it did not cause its economic ruin, as once was thought.

The Council of Blood

In 1572 the Gueux captured most of Holland and Zeeland, and the Prince of Orange and his brother Louis of Nassau invaded the Netherlands from Germany and France, respectively. Alba defeated the land invasions and recaptured part of Holland, where his troops committed terrible atrocities. Short of money and lacking adequate sea power to oppose the fleet of the Gueux, he failed to recapture the remainder of Holland and Zeeland (see LOW COUNTRIES, HISTORY OF).

Conquest of Portugal

Alba's failure and the intrigues of the Gómez party at court induced Philip to recall him (1573). In 1579 Alba was placed under house arrest on his estates after his son had married against the King's wishes. In 1580 Cardinal Granvelle persuaded Philip to let Alba command the invasion of Portugal. Within a few weeks, in one of his most brilliant campaigns, Alba took Lisbon. Yet he never regained Philip's favour. He died at Lisbon on December 11, 1582.

BIBLIOGRAPHY. There is no scholarly biography of Alba in any language, although much of the material for such a work is easily accessible, notably in the DUKE OF BERWICK AND ALBA, *Epistolario del III Duque de Alba*, 3 vol. (1952); and in the standard publications of 16th-century diplomatic correspondence. The same author's *The Great Duke of Alba As a Public Servant* (1947), is a useful but onesided corrective to the traditional Protestant view. Different aspects of Alba's career have been treated in the standard Spanish and Netherlands histories, but not generally from a biographical point of view. Only Alba's financial policy in the Netherlands has received fully adequate treatment in J. CRAEYBECKX, "Alva's Tiende Penning, een Mythe?," in *Bijdragen en Mededelingen van het Historisch Genootschap te Utrecht*, 76:10–42 (1962).

(H.G.K.)

Albania

A tiny Balkan country, with a fertile, mountain-rimmed coastal strip washed by the waters of the Mediterranean, the People's Republic of Albania, now the smallest member of the Socialist community, was long regarded as the most backward of European states. Its 11,100 square miles (28,748 square kilometres) are bounded by Yugoslavia on the north and east and Greece on the south, while Italy lies across the Adriatic Sea and the Strait of Otranto to the west. The capital is at Tiranë.

The Albania of the 1970s still bears the imprint of a long and dramatic past. Over the centuries, its seclusion cradled a distinctive and fiercely independent way of life, aptly suggested by the local name Shqipëri, meaning "eagles' country." Yet, Albania's strategic location at a crossroads of Europe also made it prey to its neighbours: it resisted Roman invasion until 168 BC and succumbed to Turkish rule in the Middle Ages only after a bitter struggle. Some 500 years of Turkish domination were followed by a brief period of independence from 1912–39 and then by the Italian invasion and occupation of World War II. Independence was subsequently restored under the Communist partisan leader Enver Hoxha, who steered the nation into a series of alliances: with Yugoslavia (1945–48), with the Soviet Union (1948–61), and then with China. The contemporary republic thus exhibits traces of Soviet economic aid, on the one hand, and a cultural revolution of the Chinese variety, on the other. In addition, outside observers have detected some signs of a distinctly Albanian approach to national and international affairs. Albanian leaders are making a strenuous attempt to modernize a country that still has many characteristics of a Third World nation—its birth rate, for example, is Europe's highest—and almost one-third of the 2,100,000 or so Albanians counted in estimates of the early 1970s lived in urban areas, a marked contrast to the rural past. A further 2,000,000 or so Albanians live in other Balkan countries, in southern Italy, Turkey, and overseas—a reflection of the adverse historical conditions that provoked massive emigration. (For related information, see ALBANIAN LANGUAGE; BALKANS, HISTORY OF THE.)

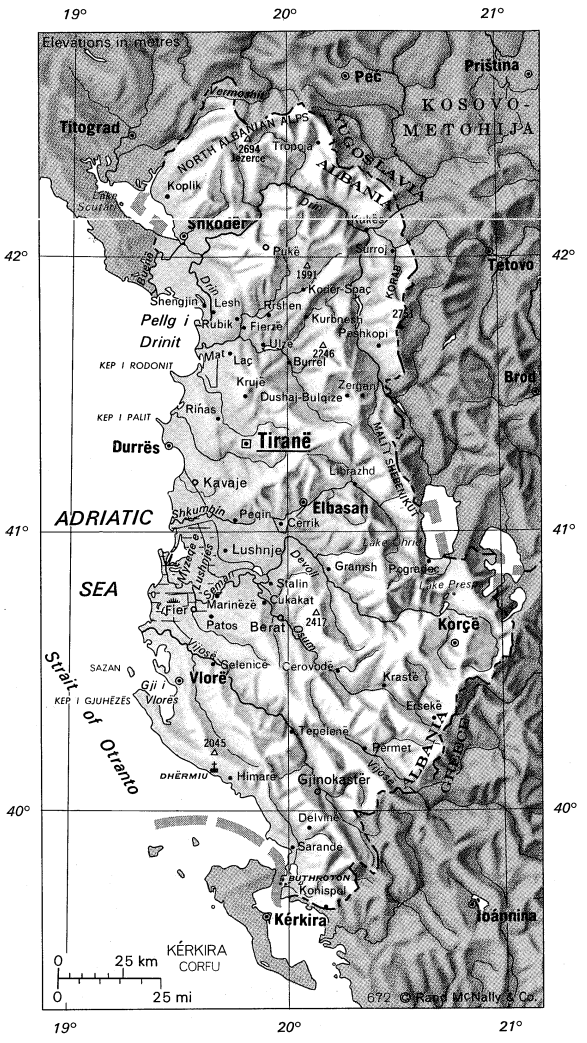
THE LAND

Relief, drainage, and climate. About 70 percent of Albania is mountainous. With few exceptions, notably the eastern Korab ranges, it owes its origin to the vast structural downwarp of the Dinaric mountain system, which dates from Mesozoic times, about 145,000,000 to 175,000,000 years ago, and which was uplifted into marked northwest to southeast relief belts in Tertiary times, about 50,000,000 years ago. The rocks are mainly limestone, often outcropping in starkly eroded forms.

The highlands. The mountain regions fall into three divisions. In the north, the North Albanian Alps run north–south, or slightly counter to the general trend. Their jagged limestone peaks soar to nearly 8,840 feet (2,694 metres), have been sculpted by the action of former glaciers, and are cut by deep, rugged, and inaccessible valleys. The domelike structure of the Krastë-Cukalit area forms a subregion.

The next division is the central uplands, where rounded formations characterize the areas of volcanic origin, while rugged limestone ranges predominate elsewhere. In the easternmost portion, the imposing gypsum block of Albania's highest peak, Korab, rises to 9,025 feet (2,751 metres). The central ranges rise to almost 7,400 feet (2,253 metres) in Mali i Shebenikut, while the westernmost range is broken up by rivers into several blocks of relief. Finally, the southern highlands also consist of three subsidiary ranges, with peaks ranging from about 7,000 to 8,000 feet. Limestone again predominates. Stretching out in front of all the highlands is a region of hills, forming tight belts in the north but gradually turning into scattered island-like clusters on the central plains. Their generally rounded profiles become rougher where limestone rocks emerge.

The lowlands. The clays and sands of the Albanian lowlands are of geologically recent origin, dating from the Miocene Epoch (less than 25,000,000 years ago) or later. The lowland belt lies in western Albania, along the irregular coast, but also thrusts finger-like inland along some of the river valleys, especially in the central regions. It is penetrated by low mountain ranges and chains of hills, and the coast is generally marked by bluffs and strings of sandy beaches: the extreme southern section is known as the Ionian Riviera.



ALBANIA

Rivers and lakes. Albanian rivers and streams form a pattern strongly marked by the orientation of the mountain ranges. With the Vermoshë (Lim), which is in the north and ultimately flows into the Danube, as the only major exception, most of the rivers drain to the Adriatic. High water levels occur from November to April, but many beds are almost dry in summer. The greatest vertical drop along the rivers occurs in the middle courses, and it is here that the nation's energy potential is located. The longest river (175 miles in Albania) is the Drin, which originates in Yugoslavia. Of the other major rivers, only the Shkumbin and the Mat flow through broad valleys: the remainder flow alternately through small basins and narrow, inaccessible gorges. Albania has three major lakes, all straddling its borders: Lake Scutari (Shkodër) lies in the extreme northwest, not far from the coast, while Ohrid and Prespa lie near each other to the east, with the last named marking the Yugoslavia–Greece–Albania frontier junction.

Mediterranean and continental climatic influences

Climate. Albania is characterized by a marked climatic variety. Inland, extremes of temperature reflect the presence of continental influences, with winds blowing from deep in the European landmass. Conditions in coastal regions are ameliorated by sea winds from the Mediterranean. Transitional regions experience sea winds in summer and winds from the interior for the rest of the year. The average rainfall is 55 inches, with almost double this amount in the north and with minimal amounts (less than half the average) in the central lowlands. In summer, the whole country is very dry and hot, with July temperatures ranging from 75° F to 81° F (24° C to 27° C), but, in winter, cyclones bring changeable, cloudy weather, with rain. Regionally, the northern interior climate resembles that of central Europe, with summer showers and harsh, cold winters, while central Albania has a typical continental climate with a wide temperature range. There are unique local winds: the bora (a strong coastal slope wind) and the sirocco (a hot, hazy wind from the North African interior) prevail in winter and the mistral (a cooler, mountain-gap wind) in summer.

Vegetation and animal life. Only a small part of Albania is completely without vegetation, and forests, some of them stunted, cover over 40 percent of the total area. The forests have been heavily cut into for logging purposes in the more accessible regions, although reforestation is a major goal of the Albanian authorities. The maqui, the typical shrub of the Mediterranean lowlands, is found in regions up to about 1,300 feet. Citrus fruits are found in the south, with figs and olives wherever there is sufficient irrigation: bananas are even found in the Albanian riviera. Among the trees, the oak predominates up to about 3,000 feet, followed by beeches and pines. The Tropojë and Pogradec regions are noted for their chestnut forests. Alpine pastures lie above the timberline, but the vegetation generally has both Mediterranean and central European characteristics: the flora includes over 3,000 different kinds of plants and herbs, which are used for industry, medical purposes, and the extraction of essences.

Old customs of weapon carrying and unrestricted shooting have taken a heavy toll of Albanian wildlife, but hunting laws have now been introduced in an attempt to preserve the remnants, which include jackals, wolves, and foxes and the rarer wild boars, bears, and chamois. The mild coastal climate attracts great numbers of migratory birds for wintering and nesting. The local livestock has degenerated in the harsh environment, and it is thought that only a slow process of crossbreeding will develop a sturdy race of cattle again: the Albanian water buffalo is slowly being replaced by an Italian counterpart. New methods of growing grapes, grain, and rice have been introduced from both East and West, and during World War II the Albanian silkworm was crossed with its Japanese equivalent.

The human imprint. The relative isolation of many parts of Albania, in conjunction with the changing currents of history, has produced a remarkable variety of regional life-styles and patterns of settlement. Apart from the long-lasting foreign rule of the Ottomans, large sections of the country were united only briefly and under such gifted sovereigns as Skanderbeg in the 15th century and the Bushatis dynasties and Ali Pasha in the 18th century. The differing Albanian tribes remained relatively independent, especially in the north. Historically important regions therefore developed, including Mirditë, a semi-sovereign Catholic dukedom; the province of Mat, the power base of former King Ahmed Zogu, who ruled from 1928–39; and the Chamuria (Çamëria), a region that generated friction between Albania and Greece and that is now largely Hellenized. The swamps of part of the coastal belt, meanwhile, had limited settlement to a few urban centres, notably Shkodër, Elbasan, and Korçë; the towns generally were small and without industries, and local trade and crafts, as a result of ineffectual practices carried over from centuries of Turkish domination, could not hold their own against European competition.

The historic regions

Since 1945, there has been a dramatic change in settlement patterns: the urban population had, by 1970, increased to over one-third of the total from the pre-World War II figure of barely 15 percent, and new mining and industrial towns have been set up. New agricultural centres in the form of huge collective farms and government-owned estates were also founded on such reclaimed coastal swamplands as the Myzeqe region, some 40 miles long and 30 miles wide. The housing shortage in industrial towns (see below *Demographic trends*) forced many labourers to live in the country and commute to towns, but recent efforts have been aimed at construction of a national road network and a develop-

Modern changes in settlement patterns

ment of population and settlement more in keeping with economic factors. Burgeoning industrial towns, such as Tiranë and Durrës, are thus becoming more and more important, while the historic centres of Shkodër and Korçë, in spite of the development of small industries, are losing much of their former significance. The increasing industrial population and the introduction of compulsory education is thus obliterating regional differences in settlement and customs, and the authorities insist that national unity can best be achieved through radical attempts to abolish all differences of tribe, religion, and even costume.

Albania, Area and Population

	area		population	
	sq mi	sq km	1960 census	1970 estimate
Provinces (*rrethi*)				
Berat	412	1,066	85,000	115,000
Dibër	606	1,569	78,000	99,000
Durrës	332	861	127,000	169,000
Elbasan	581	1,505	105,000	142,000
Fier	460	1,191	112,000	158,000
Gjirokastër	439	1,137	44,000	51,000
Gramsh	270	699	20,000	27,000
Kolonjë	310	804	17,000	19,000
Korçë	842	2,181	139,000	168,000
Krujë	236	611	43,000	69,000
Kukës	604	1,564	48,000	65,000
Lezhë	182	472	27,000	37,000
Librazhd	391	1,013	36,000	44,000
Lushnjë	275	712	66,000	91,000
Mat	397	1,028	39,000	50,000
Mirditë	269	698	17,000	27,000
Përmet	362	938	27,000	30,000
Pogradec	280	725	36,000	46,000
Pukë	374	969	23,000	29,000
Sarandë	424	1,097	48,000	62,000
Shkodër	978	2,533	128,000	167,000
Skrapar	278	720	20,000	28,000
Tepelenë	315	817	26,000	35,000
Tiranë*	458	1,186	194,000	254,000
Tropojë	403	1,043	21,000	28,000
Vlorë	621	1,609	101,000	126,000
Total Albania	11,100†	28,748	1,626,000‡	2,136,000‡

*Includes the urban entity of Tiranë. †Converted area figures do not add to total given because of rounding. ‡Figures do not add to total given because of rounding.
Source: Official government figures; UN.

THE PEOPLE

Ethnic and social background. The ultimate origin of the Albanians remains a mystery: they are generally considered to be descendants of the ancient Illyrians (Albanoi) and thus members of the Indo-European race, a hypothesis for which comparative linguistics offers some supporting evidence. The strongest evidence is the continuity of settlement and of local names within the boundaries of contemporary Albania. Strangely enough, rivers seem to have been more of a cultural barrier than mountain passes and ridges, for the boundary between the two major language groups, Ghegs in the north and Tosks in the south, roughly follows the Shkumbin River. These two groups, which also differ in dress, musical tradition, and other customs, are further subdivided into the Dukagjin and Malësia in the north and the Myzeqe, Camövia, and Labëria in the south, although differences are relatively minor. Differences in script add to the variety: Cyrillic and Arabian as well as Latin letters were traditionally used. Between the world wars some attempts were made to create a unified language, but it may be that present attempts to formulate and impose a language based on the Tosk dialects have been more successful—they have even been tolerated by the Ghegs in neighbouring Yugoslavia.

Religious divisions were perhaps deeper rooted: in the mid-1960s, 70 percent of the populace were Muslims, 20 percent Greek Orthodox, and 10 percent Roman Catholics, a breakdown reflecting various currents of history. The Bektash, a liberal Muslim sect, had many Albanian followers and made their headquarters in the country following their expulsion from Turkey early in the present century. Under the present government, places of worship have all been closed down (even though the constitution promises equality to all religions), and traditional observances—never zealously followed by the mass of people—have generally suffered a diminished status.

It may be noted that only half the Albanian people live in the contemporary state: the miserable traditional economic conditions generated a mass exodus over the centuries. The Albanian minority in Greece (of 16th-century origin) has now been completely assimilated, but the minority in Kosovo-Metohija, Yugoslavia, have achieved autonomy. Other emigrants are found in Turkey, Italy, and the United States. Within Albania, on the other hand, the minority groups of Greeks, Romanians, and Bulgarians make up about 5 percent of the population and, as is customary in Socialist countries, are allowed to keep their language and national customs.

Demographic trends. With a population density of 192 persons per square mile, Albania is the most sparsely settled country of the Balkans, although its rate of growth is Europe's highest. Major population centres are the provinces of Durrës, Tiranë, and Lushjnë, while the mountain districts of Tropojë, Pukë, Mirditë, and Kolonjë are very thinly populated. The rate of growth has varied in recent decades: from 19.9 per thousand people annually in pre-World War II years, it soared to an average of 29 per thousand for the 1951–65 period. The postwar period saw a reduction in the birthrate from a high of 43.8 per thousand (1955) to 35.3 per thousand at the end of the 1960s, while the death rate decreased from 11 to eight per thousand. The current government forbids birth control, and, as the majority of the population comes from farming groups, a population surplus has been possible on farms and among industrial workers: measures have been introduced to protect agriculture against a loss of manpower to industry and to prevent a mass migration to urban areas. The overall age composition of the population is demographically sound, although life expectancy has, with increased health facilities, gone up to 65 years for men and 67 years for women. Polygamy and the pre-arranged marriages of children officially have been abolished—though traces of both still remain—and modern forms of courtship and marriage are slowly replacing older customs.

Government control of migration

Albania has some 60 towns, most of them with fewer than 5,000 inhabitants. Three towns have a population of less than 1,000, and only 15 are larger than 20,000. Only in recent years have cattle raising and farming been abolished in urban areas. Since 1945, however, there has been a marked demographic shift toward industrial occupations, accompanied by a strong growth in the intelligentsia and the gradual rise of the Socialist equivalent of a middle class technocracy.

THE NATIONAL ECONOMY

Natural resources. Albania's sedimentary rocks contain large deposits of coal (in low quality lignite form), as well as petroleum and bitumen. There are also large deposits of chromium, copper, iron, and nickel. The bitumen—mined long ago under the Ottomans—is still extracted in Selenicë and Resulair, but, as reserves are being depleted rapidly, it is necessary to enrich large quantities of inferior ore. Crude-oil production is more important, with wells at Stalin (Kuçovë), Patos, and Marinëzë. The rich natural-gas deposits are being developed as power-plant fuel sources, and there is a coke plant at Qyteti Stalin.

Albanian chromium (found in Dushaj-Bulqizë, Badër Thekën, Tropojë, and Surroj) is of excellent quality with a high metal content, and enrichment techniques have been introduced recently. Copper ore, of poorer quality, is found in the Mirditë, Pukë, and Kukës districts, with metallurgical industries centred on Kurbnesh, Kodër-Spaç, and Rubik. Iron and nickel production, begun as late as 1958, has now greatly expanded, and sulfur deposits, as well as lead, bauxite, and zinc, are being exploited. The 1970–75 plan stressed phosphorite exploitation and the production of phosphate fertilizers. Albania

is also rich in mineral springs of various kinds, although exploitation for health purposes has so far been rather sparse.

Economic planning and sources of national income. In addition to natural resources, Albania has a good agricultural and cattle-raising potential and great tourist prospects, but late industrialization and the lack of investment resources are major handicaps. Economic development has only been possible with assistance from the Soviet Union and other Socialist countries and, latterly, from China alone. The introduction of a planned economy resulted in the division of the country into six regions. There are: the entire north, characterized by mines, forests, and waterpower; the industrial and agricultural area around Tiranë and Durrës; the crude-oil region around Berat; a mining district around Elbasan; the subtropical area around Himarë, marked by pastures, forests, hydroelectric plants, and some mining; and the southeastern region of rich agricultural production centred on Korçë. Economic development is by means of plans, with the 1970–75 plan molding much of the economy of the present decade. Heavy industry predominates in the state industrial sector, while agriculture, trade, and crafts are partly organized into cooperatives. Alone among the Socialist nations of the 1970s, Albania clings to strong central planning and strict wage and price controls.

Agriculture. Up to 1955, agricultural collectivization proceeded slowly, and only 85 percent of farms had been affected by 1959, although collectivization was complete a decade later. State farms occupy some 20 percent of agricultural land, with the remainder given over to agricultural cooperatives of various sizes, with an average of 1,700–2,500 acres each: the area of private plots has been reduced to a bare minimum. The lack of farm implements (the whole country had but 6,700 tractors at the start of the present decade), insufficient fertilization, and innate rural conservatism have resulted in constant failure to meet production quotas, and the loan of volunteers from urban areas has but temporarily alleviated the situation. In 1955 a program was introduced that aimed at achieving national self-sufficiency in grain production; although output per acre has almost doubled over the pre-World War II figure, the aim has not been achieved, and there has been a great increase in potato planting in an attempt to bridge the ensuing gap. The situation in cattle raising is also bad, and there are plans to replace traditional sheep and goat raising with an emphasis on cows, pigs, and poultry and beasts of burden. Attempts are being made to increase milk and wool production by cross-breeding, but shortages of milk and meat are far from being solved. Fruit and vegetable production could—thanks to the favourable southern location—dominate the entire central European market, though realization of this potential is largely a question of economic planning (especially in the transportation field). Forests (now more accessible) are being depleted to keep up with demand, although there is a re-afforestation program. In addition to exports, wood is also used domestically for industry and as fuel.

Problems in agricultural policy

Industry. Albania's industrial foundations were laid with Soviet assistance, while with later Chinese help the country has expanded into metallurgical industries (copper, chromium, iron, and nickel) and has begun to manufacture machinery, some quite complex. Steel production was initiated under the 1970–75 plan. Western observers feel that planning has been overcentralized and inflexible, and the lack of a qualified labour force (and lack of knowledge of recent scientific developments) has resulted in a small range of generally poor-quality products. More positively, electrification of all Albanian villages was completed in 1970, and in November 1971 the nation's largest power plant (capacity 200,000 kilowatts) was opened in Va i Dejës. A larger (capacity 400,000 kilowatts) plant was planned for Fierzë.

Albania's mining products are best suited for export; chromium, dressed copper, and iron–nickel are increasingly important. Other exports include copper cable, wood, textiles, and canned goods. Imports are restricted to equipment requirements, chiefly supplied by China, but the trade balance remains negative.

Transportation. Albania has possessed railroads only since 1948, and the network now stretches from Tiranë north to the industrial centre of Laç, south to Fier, and east to Librazhd and beyond. In 1944 there were only 1,390 miles of roads, some 250 miles of which were asphalted, but, during the next 25 years, some additional miles of national highway and some miles of local roads were built. This construction had high priority not only to make agricultural regions accessible but also to facilitate commuter traffic to industrial centres and to facilitate mining and wood-cutting operations. Trucks have to be imported to meet distribution needs. Except on the Bunë River, there is no inland water traffic, and seagoing freighters are used to convey exports to distant markets. Durrës is the major port, processing four-fifths of total tonnage, although the smaller ports of Vlorë, Sarandë, and Shëngjin are expanding. The single Albanian airport, at Rinas, is small, operating flights to neighbouring countries only when weather conditions permit.

Labour laws, social insurance, and wage–price structures. The social-insurance system is based on the Soviet prototype, and there is free medical assistance, together with pension plans (even for farmers) and universal health insurance. Low pensions, however, follow the low wage levels, and the relative scarcity of medical personnel (1,808 physicians and dentists by 1970) impedes comprehensive medical services. Under government decree the mass migration to the cities has been halted, and office workers are required to provide a minimum of physical labour or to move to the countryside. Unhygienic working conditions are openly criticized, although remedial measures, including the as-yet-unfulfilled aim of installing public medical centres in every agricultural collective, have been initiated.

Two-thirds of Albania's net national income is placed in a social-consumption fund, the remainder in capital-investment programs. The tax base is strongly dependent on indirect taxation, with an added-value tax contributing 48 percent and excess-profit tax 28 percent of the national budget: direct taxes have been abolished. The authorities attempt to justify the admittedly low standard of living by pointing to the absence of direct taxation and noting that the prices of basic necessities have been reduced on several occasions. Non-essential consumer goods, however, have a high markup, and wages are held low. Housing and public transportation are relatively inexpensive. The Albanian life-style, with regard to food and housing, is generally extremely frugal, but the traditional lavish entertainment of guests still takes place, in spite of official disapproval. As long-lasting homemade goods are no longer produced, demand for consumer goods is mushrooming, placing pressures on an economy that is, as yet, ill equipped to satisfy them.

Taxation

ADMINISTRATION AND SOCIAL CONDITIONS

The constitutional framework. The Albanian constitution of 1946, based on that of Yugoslavia, was revised along Soviet lines in 1950; as it is manifestly unsuited to the ensuing Albanian brand of Socialism, a total reform was contemplated in the early 1970s. The Albanian Party of Labour, or APL (Partia e Punës së Shqipërisë, or PPSh) is the sole party represented in the government and has constitutional priority over all government agencies: observers in both the West and in some Socialist countries have noted tendencies toward a centralized, unresponsive, and sprawling bureaucratic structure, a characterization Albanian authorities and their international allies have fiercely rejected. The highest authority within the government is the People's Assembly, elected every fourth year. It has over 260 representatives (one for each 8,000 voters) and normally meets twice a year; between sessions, the 14-member Presidium acts as a law-making body. The first decade of APL rule saw Enver Hoxha (still first secretary of the APL in the early 1970s) performing the function of both party and state leadership, but in 1954 he delegated

affairs of state to Mehmed Shehu, another powerful personality and former Communist guerrilla leader. The government has maintained a relatively stable composition with the exception of the Ministry of Justice, which was abolished in 1966. A more decentralized judicial system, with the highest court directly responsible to the parliamentary body, seemed to have introduced a more effective method of dispensing justice: village and factory courts are now firmly established.

Local administration comprises 26 small districts, with dominant villages in agricultural cooperatives functioning as administrative centres, as well as the usual small towns and larger cities. The system reflects the structure of the central government.

The Albanian Party of Labour. The APL of the early 1970s was unique among the Communist parties of the Socialist community. It was founded in 1941 by young partisan leaders aided significantly by Yugoslav Communists. Its subsequent development was strongly molded by Enver Hoxha, who was confirmed in his leadership over 30 years later, at the 6th Party Congress of November 1971. A university scholar who studied law in Belgium and a former secondary school teacher, he was typical of the many members of the 17-man Politburo who studied abroad. Although the Politburo leadership shows as much continuity as that of the central administration, the membership of the party's large Central Committee changes more often: over one-fifth of the 110 members at the 1971 congress were newly elected. The break with Belgrade in 1948 and with Moscow in 1961, together with the new Chinese foreign policy, have caused some problems for the APL as it has made the necessary sharp policy changes. Its pronouncements have also achieved international significance as a political barometer reflecting the wider disagreements in the Socialist bloc. Partly as a result, the country's diplomatic representation has expanded in recent years, embracing over 50 countries by the start of the 1970s. Developing-world and Socialist-bloc missions were particularly active.

Changing party membership

Social organizations. The APL permits the existence of mass social organizations; thus, the Democratic Front is regarded as a rallying point for those Albanians who are not party oriented. It carries out mass economic and cultural programs and also arranges the elections to the various administrative agencies. The Albanian Trade Unions (Bashkimet Profesionale të Shqipërisë) represent, according to official statistics, all but 4 percent of Albanian industrial workers. Half of the members are women. Major trade-union efforts are directed at attempts to meet production quotas and to enforce labour discipline. The Albanian League of the Working Youth (Bashkimi i Rinisë së Punës së Shqipërisë), formed almost simultaneously with the APL, includes about half the young people in the 7–18-year age group. It, too, is engaged in mass cultural and economic programs. Youth camps for building railways and for implementing soil-enrichment projects act as ideological schools, and the youth group functions as a reservoir for coming generations of APL members. The Albanian League of Women (Bashkimi i Grave të Shqipërisë) concerns itself primarily with women's emancipation, a task begun largely since 1945. Its task is often difficult, especially in the mountain regions, where chauvinist traditions hold sway.

The armed forces. Military service—two years in the army, three in the navy, air force, or special units—is compulsory. Albania abolished military ranks in 1966 and reinstated the APL political commissars. Other revived partisan principles included the idea of a self-sufficient army, producing its own food and technical equipment. Naval and infantry equipment remains primarily Soviet, and somewhat un-Socialist property rights were invoked during the Albanian seizure of Soviet submarines following the 1961 split. The air force's Soviet MIGs are gradually being replaced by Chinese aircraft. During the Albanian version of a cultural revolution in the 1960s, existing paramilitary organizations were given greatly increased emphasis and soon permeated the entire society.

CULTURAL LIFE

The Cultural Revolution. Albania's Cultural Revolution, although carried out virtually simultaneously with China's and with similar radical consequences for both the traditional and the pre-existing Socialist cultural structure, has nevertheless been quite unique. In March 1966, as the first steps of an administrative reform, many party and state functionaries, hitherto assigned to the central offices, found themselves rapidly and unexpectedly transferred to remote villages. The Cultural Revolution, however, did not unleash bands of Red Guards, as in China, but was strictly controlled throughout by the party leadership. The 5th Party Congress in 1966 marked the high point of the battle against tradition and "religious prejudices," and a reduction of salaries and an increased work load was coupled with a strong emphasis on national culture. Classic Western literature and drama vanished from bookstores and theatres. Leadership in many state bodies was also handed over to younger, often ill-qualified persons.

The Albanian and Chinese experiences

Art and science. The long Ottoman rule precluded any great development of Albanian culture, although many talents found expression elsewhere in the Turkish empire: a notable example was Sinan, the most famous Constantinople architect of his day, who fled his native Albania to escape Turkish massacres in 1464. Kemal Atatürk, the founder of modern Turkey, was also of Albanian descent, and the Muḥammad ʿAlī dynasty of Egypt, which lasted from 1805 to 1952, was founded by an Albanian. A modern example is provided by the famous Yugoslav writer and former aide to Tito, Milovan Djilas, who is also of Albanian origin.

Within Albania, the major cultural institution is the National Library in Tiranë, which contains over 450,000 volumes, some of them rare, and averages 300 visitors a day. There are also nearly 40 additional public libraries and some 80 smaller library branches. Films are very popular, and the Albanian film studios produce documentaries, newsreels, and feature films. There are some 160,000 radio receivers in the country, their distribution being facilitated by the recent electrification of all the villages. The radio network has one main transmitter and a number of additional stations. Television programming began on January 1, 1971. Newspapers and magazines have a combined circulation of about 330,000. The Albanian Telegraph Agency, ATA, functions as the official news agency.

Collective education. Illiteracy was widespread before the introduction of a 1946 school reform, making eight years of elementary education compulsory. At the same time, technical courses were introduced, as was co-education and the emancipation of women in higher education. Later reforms of 1963 and, especially, 1970 showed a strong deviation from the Yugoslavian–Soviet models, and the 1971 Party Congress called for the graduation of some 20,000 specialists (notably those in the technical and economic fields) from higher education bodies by 1975.

Party leaders have constantly demanded a trend to collective education in their speeches concerning current reforms, and, by revolutionizing their educational methods, the Albanians hope to produce a "new man" (perhaps not unrelated to the Chinese model). Young people therefore have to perform a threefold task, in their capacities as students, temporary workers, and military trainees. They thus carry an unusually large burden. These activities take place in a framework containing almost 60 general secondary schools, some 36 intermediate engineering schools, and some 20 technical schools. The University of Tiranë, founded in 1957, had affiliated bodies in nine cities by 1970. Over 8,000 students are enrolled in evening courses, correspondence courses, or as full-time students at the university, with almost 2,000 enrolled at the affiliates. There are also advanced schools of agriculture, sport, education, and fine art, and a school of engineering was under construction in the early 1970s. The major educational problem continues to be access to and utilization of the accumulating world body of scientific knowledge.

Educational goals

THE OUTLOOK

Albania's modernization has been accomplished through a Socialist system, but the foundations of the development already had been laid by Italian and Austrian cultural influence; and it could not have taken place without Yugoslav, Soviet, and Chinese aid. An industrial base has been created, and the APL has provided a measure of ideological guidance and incentive. Isolation, however, has been encouraged by the peculiar forms taken by Albanian Communist thought, a late and isolationist blossoming of nationalism, and the strange and perhaps temporary alliance with far-off China. Outside observers noted that, if anything, the slogans of radically revolutionary class conflict—abolition of national customs, prohibitions of all religious manifestations, and the shunning of co-existence, even with the Soviet Union—were heard as loudly as ever at the 1971 Congress. Under these conditions, it seems that future generations alone can fulfill the task of bringing Albania up to the European level of development; prospects of achieving this goal are not entirely unfavourable, given a pool of motivated young people willing to work and make sacrifices.

BIBLIOGRAPHY. STAVRO SKENDI (ed.), *Albania* (1956), is the best general survey in English up to the year 1955. For recent years, see the section on Albania in the annual *Statesman's Year-Book*. Recent books of special interest are W.E. GRIFFITH, *Albania and the Sino-Soviet Rift* (1963), which describes the role of Albania in the quarrel between the Soviet Union and China; N.C. PANO, *The People's Republic of Albania* (1968), a study on political developments including the Cultural Revolution; J.I. THOMAS, *Education for Communism: School and State in the People's Republic of Albania* (1969), a basic study but not including the last school reform. There are no recent general books giving the Western viewpoint. Useful and informative works, as well as propaganda publications, are available from Albania in English translation. They include *History of the Party of Labor of Albania* (1969–70); ENVER HOXHA, *Speeches* (1967–); *Reports Submitted by Hoxha and Shehu at the Fifth and 6th Congress of the Party of Labor of Albania* (1966 and 1971); and MEHMET SHEHU, *On the Experience of the National Liberation War and on the Development of the National Army* (1963), an authoritative statement of Albanian partisan warfare. KRISTO FRASHERI, *The History of Albania* (1964); and KOCO BIHIKU, *An Outline of Albanian Literature* (1964), are brief surveys from official publications.

(R.R.Sc./O.Li.)

Albanian Language

Albanian is an Indo-European language spoken by about 2,500,000 inhabitants of the eastern Adriatic coast in Albania and also in neighbouring Yugoslavia, principally in Kosovo and Makedonija (Macedonia), west of a line from near Leskovac to Lake Ohrid. There are perhaps 300,000 more speakers in isolated villages in southern Italy (Abruzzi e Molise, Basilicata, Puglia, and Calabria), and Sicily, and southern Greece (in Voiotía, Attica, Évvoia, Ándros, and the Pelopónnesos).

The origins of the general name Albanian, which traditionally referred to a restricted area in central Albania, and of the current official name Shqip or Shqipëri, which may well be derived from a term meaning "pronounce clearly, intelligibly," are still disputed. The name Albanian has been found in records since the time of Ptolemy. In Calabrian Albanian the name is Arbresh, in Modern Greek Arvanítis, and in Turkish Arnaut; the name must have been transmitted early through Greek speech.

Gheg and Tosk dialects

Dialects. The two principal dialects, Gheg in the north and Tosk in the south, are separated roughly by the Shkumbin River. Gheg and Tosk have been diverging for at least a millennium, and their less extreme forms are mutually intelligible. Gheg has the more marked subvarieties, the most striking of which are the northernmost and eastern types, which include those of the city of Shkodër (Scutari), the neighbouring mountains along the Crna Gora border, Kosovo, Makedonija, and the isolated village of Arbanasi (formerly Borgo Erizzo) on the Croatian coast of Dalmatia outside Zadar. Arbanasi, founded in the early 18th century by refugees from near Bar (formerly Antivari), has about 2,000 speakers.

All of the Albanian dialects spoken in Italian and Greek enclaves are of the Tosk variety, and seem to be related most closely to the dialect of Çamërija in the extreme south of Albania. These dialects resulted from incompletely understood population movements of the 13th and 15th centuries. The Italian enclaves—nearly 50 scattered villages—probably were founded by emigrants from Turkish rule in Greece. A few isolated outlying dialects of south Tosk origin are spoken in Bulgaria and Turkish Thrace but are of unclear date. The language is still in use in Mandritsa, Bulgaria, at the border near Edirne, and in an offshoot of this village surviving in Mándres, near Kilkís in Greece, that dates from the Balkan Wars. A Tosk enclave near Melitopol in the Ukraine appears to be of moderately recent settlement from Bulgaria. The Albanian dialects of Istria, for which a text exists, and of Sirmio (Srem), for which there is none, have become extinct.

History. The official language, written in a standard roman-style orthography adopted in 1909, was based on the south Gheg dialect of Elbasan from the beginning of the Albanian state until World War II, and since has been modelled on Tosk. The region of Kosovo, officially bilingual in Serbian and Albanian, uses a standardized form of the Gheg dialect of the region. Before 1909, the little literature that was preserved, was written in local makeshift Italianate or Hellenizing orthographies, or even in Turko-Arabic characters.

The first written records

A few brief written records are preserved from the 15th century, the first being a baptismal formula from 1462. The scattering of books produced in the 16th and 17th centuries originated largely in the Gheg area (often in Scutarene north Gheg) and reflect Roman Catholic missionary activities. Much of the small stream of literature in the 19th century was produced by exiles. Perhaps the earliest purely literary work of any extent is the 18th-century poetry of Gjul Variboba, of the enclave at S. Giorgio, in Calabria. Some literary production continued through the 19th century in the Italian enclaves, but no similar activity is recorded in the Greek areas. All these early historical documents show a language that differs little from the current language. Because they exhibit marked dialect peculiarities, however, they often have a value disproportionate to their literary merit.

Classification. That Albanian is of clearly Indo-European origin was recognized by the German philologist, Franz Bopp, in 1854; the details of the main correspondences of Albanian with Indo-European languages were elaborated by another German philologist, Gustav Meyer, in the 1880s and 1890s. Further linguistic refinements were presented by the Danish linguist Holger Pedersen and the Austrian Norbert Jokl. The following etymologies illustrate the relationship of Albanian to Indo-European (an asterisk preceding a word denotes an unattested, hypothetical Indo-European parent word, which is written in a conventionalized orthography): *pesë* "five" (from **pénk^{w}e*); *zjarm* "fire" (from **g^{w}hermos*); *natë* "night" (from **nok^{w}t-*); *dhëndër* "son-in-law" (from **ǵemə ter-*); *gjarpër* "snake" (from **sérpŏn-*); *bjer* "bring!" (from **bhere*); *djeg* "I burn" (from **dheg^{w}hō*); *kam* "I have" (from **kapmi*); *pata* "I had" (from **pot-*); *pjek* "I roast" (from **pek^{w}ō*); *thom*, *thotë* "I say, he says" (from **k'ēmi*, **k'ēt* . . .).

The verb system includes many archaic traits, such as the retention of distinct active and middle personal endings (as in Greek) and the change of a stem vowel *e* in the present to *o* (from **ē*) in the past tense, a feature shared with the Baltic languages. For example, there is *mbledh* "gathers (transitive)" as well as *mblidhet* "gathers (intransitive), is gathered" in the present tense, and *mblodha* "I gathered" with an *o* in the past. Because of the superficial changes in the phonetic shape of the language over 2,000 years and because of the borrowing of words from neighbouring cultures, the continuity of the Indo-European heritage in Albanian has been underrated.

Albanian shows no obvious close affinity to any other Indo-European language; it is plainly the sole modern survivor of its own subgroup. It seems likely, however, that in very early times the Balto-Slavic group was its

nearest of kin. Of ancient languages, both Dacian (or Daco-Mysian) and Illyrian have been tentatively considered its ancestor or nearest relative.

Grammar. The grammatical categories of Albanian are much like those of other European languages. Nouns show overt gender, number, and three or four cases. An unusual feature is that nouns are further inflected obligatorily with suffixes to show definite or indefinite meaning; *e.g.*, *bukë* "bread," *buka* "the bread." Adjectives—except numerals and certain quantifying expressions—and dependent nouns follow the noun they modify; and they are remarkable in requiring a particle preceding them that agrees with the noun. Thus, in *një burrë i madh*, meaning "a big man," *burrë* "man" is modified by *madh* "big," which is preceded by *i*, which agrees with the term for "man"; likewise, in *dy burra të mëdhenj* "two big men," *mëdhenj*, the plural masculine form for "big," follows the noun *burra* "men" and is preceded by a particle *të* that agrees with the noun. Verbs have roughly the number and variety of forms found in French or Italian and are quite irregular in forming their stems. Noun plurals are also notable for the irregularity of a large number of them. When a definite noun or one taken as already known is the direct object of the sentence, a pronoun in the objective case that repeats this information must also be inserted in the verb phrase; *e.g.*, *i-a dhashë librin atij* is literally "him-it I-gave the-book to-him," which in standard English would be "I gave the book to him." In general, the grammar and formal distinctions of Albanian are reminiscent of Modern Greek and the Romance languages, especially of Romanian. The sounds suggest Hungarian or Greek, but Gheg with its nasal vowels strikes the ear as distinctive.

Vocabulary and contacts. Although Albanian has a host of borrowings from its neighbours, it shows exceedingly few evidences of contact with ancient Greek; one such is the Gheg *mokën* (Tosk *mokër*) "millstone," from the Greek *mēkhanḗ*. Obviously close contacts with the Romans gave many Latin loans; *e.g.*, *mik* "friend," from Latin *amicus; këndoj* "sing, read" from *cantāre*. Furthermore, such loanwords in Albanian attest to the similarities in development of the Latin spoken in the Balkans and of Romanian, a Balkan Romance tongue. For example, Latin *palūdem* "swamp" became *padūlem*, and then *pădure* in Romanian and *pyll* in Albanian, both with a modified meaning, "forest."

Early contacts between Albanian and Romanian

Conversely, Romanian also shares some apparently non-Latin indigenous terms with Albanian; *e.g.*, Romanian *brad*, Albanian *bredh* "fir." Thus these two languages reflect special historical contacts of early date. Early communication with the Goths presumably contributed *tirq* "trousers, breeches" (from an old compound "thigh-breech"), while early Slavic contacts gave *gozhdë* "nail." Many Italian, Turkish, Modern Greek, Serbian, and Macedonian-Slav loans can be attributed to cultural contacts of the past 500 years with Venetians, Ottomans, Greeks (to the south), and Slavs (to the east).

A fair number of features—*e.g.*, the formation of the future tense and of the noun phrase—are shared with other languages of the Balkans but are of obscure origin and development; Albanian or its earlier kin could easily be the source for at least some of these. The study of such regional features in the Balkans has become a classic case for research on the phenomena of linguistic diffusion.

BIBLIOGRAPHY. P.L. HORECKY (ed.), *Southeastern Europe: A Guide to Basic Publications*, pp. 102–111 (1969), includes a compilation by E.P. HAMP of about 50 major annotated items on Albanian scholarship, language, literature, folklore, ethnography, and folk music, with references to other supporting work; E.P. HAMP, "Albanian," in T.A. SEBEOK (ed.), *Current Trends in Linguistics*, vol. 9, pp. 1626–92 (1972), a review of work since 1918, with copious bibliography.

(E.P.H.)

Alberta

Alberta, sparsely populated but possessing rich resources and beautiful scenery, is the most westerly of the three Prairie Provinces occupying the continental heartland of southern Canada. With an area of 255,285 square miles (661,188 square kilometres), it remains one of the most desirable of the underpopulated areas of the globe.

Location and general character

Following the earliest explorations of Indians and fur traders, the settlement of its plains and parklands led to the rapid development of agriculture. Subsequent exploitation of rich oil, gas, coal, and timber resources has led to further population growth, with steadily increasing urbanization and industrialization, while the scenery of the mountain parks in the west is internationally renowned, with tourism a major factor in the provincial economy.

Relative isolation from the more populous eastern regions of the nation has slowed the development of industries needing mass markets. As communications improve, rapid changes are taking place. Notably, natural routes to the north have made the Alberta of the 1970s the major jumping-off point for development of the Arctic.

Alberta is 756 miles (1,216 kilometres) in extent from north to south, and 404 miles (650 kilometres) across at greatest width. To the east, the 110th meridian is the boundary with its prairie neighbour Saskatchewan. Its western boundary with British Columbia is formed by the 120th meridian and the crest of the Rocky Mountains. North, beyond the 60th parallel, the Mackenzie District of the Northwest Territories slopes down to the Arctic Ocean, while the 49th parallel in the south forms the international boundary between Alberta and the state of Montana in the United States. Freshwater surfaces cover 6,485 square miles (16,796 square kilometres) of the province. Alberta was established as a district of the Northwest Territories in 1882 and was enlarged to its present boundaries on becoming a province in 1905. The provincial government has its seat in Edmonton.

(R.O.H./D.A.E.S.)

THE HISTORY OF ALBERTA

Exploration and settlement. The area now known as Alberta has been inhabited by various Indian groups for at least 10,000 years. European explorers first appeared in the 1750s to extend the fur trade. Two rival trading companies, the Hudson's Bay Company and the North West Company, began building trading posts in the last quarter of the 18th century along the major northern rivers, most notably the Athabasca, North Saskatchewan, and Peace. From 1821, when the companies merged, until 1870, when this region was transferred to the Dominion of Canada, the Hudson's Bay Company controlled and governed the area, which was populated by Indians, the half-breed Métis, and a few fur traders and missionaries.

After 1870, settlement in southern Alberta began, based on a ranching economy. The Indian tribes had been decimated by European diseases; the disappearance of the buffalo, his main source of livelihood; and a devastating proclivity for alcohol. The signing of treaties placed most of the remaining natives on reservations, but not before the abuses of unscrupulous whisky traders had brought the North West Mounted Police westward. In 1874 they established Ft. Macleod and laid the bases of Canadian law enforcement in Alberta.

A railroad that reached Calgary in 1883, cheap or free land, and vigorous publicity campaigns brought an influx of settlers from eastern Canada, the United States, and Europe. The population grew from 73,000 in 1901 to 374,000 in 1911. The development of an earlier-ripening variety of wheat, in 1904, made crop farming less risky and, in northerly areas, newly feasible. Subsequently a wheat-based economy expanded throughout most of the arable parts of the province.

Provincial status. Responsible government in the territories developed gradually from 1875, when the North West Territories Act went into effect, until 1897, when a fully responsible legislative assembly was elected. Made a district of the Northwest Territories in 1882, Alberta was enlarged to its present boundaries in 1905 and made a province of Canada; the Liberal Party formed the first government. Edmonton, as the capital and distribution centre, grew rapidly, as did other urban centres. Calgary boomed with the discovery of oil at Turner Valley in 1914. Medicine Hat and Lethbridge, the latter a coal-

Aspects of economic growth and instability

mining area since 1869, developed into large distribution centres. Railways spread over most of the province, increasing agricultural development and providing a ready market for the province's vast coal deposits.

During World War I the growth of population and economy slowed. A postwar slump in wheat prices was a major contributing factor to agrarian political agitation, which led to the United Farmers Party of Alberta forming the government in 1921. Until the Depression of the 1930s, Alberta's growth in population and economy continued. In 1935, the Social Credit Party, espousing a new monetary policy, was elected and retained power for 36 years. Not until World War II, however, did the economy begin to recover. Since the discovery of oil at Leduc in 1947, and subsequent discoveries of other major oil and natural gas deposits, Alberta has experienced rapid economic expansion. (R.B.D.)

THE NATURAL AND HUMAN LANDSCAPE

Relief and drainage. Along the southwest border, Mt. Columbia (12,294 feet [3,747 metres]) is the highest of 26 magnificent peaks reaching above 11,000 feet. Beyond the narrow foothills, soft rocks form an undulating plateau falling from over 3,000 feet to under 1,000 feet in the northeast, where the ancient rocks of the Precambrian shield outcrop. Outliers of higher ground include the Cypress and Swan Hills, and the Caribou Mountains.

The Columbia Icefield (*q.v.*) is the source of two major rivers, the Athabasca and the North Saskatchewan. The first flows northeast to the Lake Athabasca–Lake Claire complex, where it is joined by the Peace River, which crosses the north of the province from British Columbia. The Slave River drains the area north toward the Arctic Ocean, and the lowest point in the province (573 feet) occurs in its valley.

Apart from a small area drained by the Milk River into the Missouri, the south of the province is dominated by the Saskatchewan River system. The Oldman, Bow, and Red Deer rivers combine to form the South Saskatchewan, which flows east with the North Saskatchewan toward Hudson Bay. Many of the rivers flow through deeply incised valleys, in which the eroded, barren landscapes known as badlands may develop. Those of the Red Deer are famous for their rich deposits of dinosaur remains.

Climate. Alberta has a continental climate, with more sunshine—2,000 to 2,200 hours a year—than any other Canadian province. Winters are dry, sunny, and cold, though in the south Chinook winds, which comprise warm, dry airflows of Pacific origin descending the western slopes of the Rockies, may raise temperatures by 40° F (22° C) in ten minutes. Summers are warm and wetter (except in drought years), with occasional destructive hail storms. Edmonton's mean and extreme temperatures are, in January, 6° F and −57° F (−15° C and −49° C), while in July they are 62° F and 99° F (17° C and 37° C). Annual precipitation averages 18 inches, with half falling between June and August.

Soils and vegetation. Chernozems—rich brown and black soils with a deep humus layer—underlie the prairies and parklands. Blue grama and other grasses dominate the prairie area, though cacti, tumbleweeds, and sage brush are more conspicuous; few trees grow naturally outside river valleys. The transitional parklands have aspen bluffs, increasing in size until forest dominates. The less fertile gray-wooded soils known as podzols underlie the extensive mixed and coniferous forests of the foothills and the north, where various combinations of aspen, white spruce, jack and lodgepole pines, and balsam fir are found. Black spruce and tamarack grow over accumulations of organic peat in extensive bogs. Engelmann spruce and alpine fir are important trees in the mountain forests. The wild rose, Alberta's provincial flower, is widespread.

Animal life. Meadowlarks and pronghorn antelopes are conspicuous on the prairies, and the badlands support rattlesnakes, horned lizards, and scorpions. Whitetailed deer and coyotes are familiar in parkland areas, while the northern forests house a wealth of fur bearers (wolverines) and big game, including the moose, wapiti, caribou, and black bear. Gray jays visit campsites, and loons are heard on the many lakes. Mountain goats, bighorn sheep, and grizzly bears, together with Clark's nutcrackers and golden eagles, are characteristic of the mountains, while rivers and lakes throughout the province support trout, whitefish, and pike, and ducks nest by every slough. Examples of most habitats (except the prairies) and some threatened species receive partial protection in national and provincial parks, wilderness areas, and other reserves. Resource development in the reserve regions, and the growing pollution of other areas, gave rise to much controversy in the early 1970s.

Threatened habitats

Traditional regions. The agricultural southeast is fairly uniform, though with considerable variations in landscape, farming practice, and density of settlement, while the isolated Peace River block forms a basically agricultural area to the northwest. The broad strip of forested foothills is little settled outside of mining areas, but much visited for hunting and other recreational purposes, utilizing the network of forestry roads. The northern forested area, away from major roads, is little changed from fur trading days, with a scattered population of trappers and fishermen. The mountain national parks each have small permanent populations and many visitors.

Rural settlement. The square pattern of the Dominion land survey dominates much of the landscape. The survey established townships with 36 sections, each one mile square, and an appropriate road allowance. From the air, the resulting great square fields can be seen spreading from horizon to horizon, while on the ground the straight roads seem to stretch endlessly.

Homesteading introduced some individuality into this pattern: log huts, elegant frame buildings, and modern houses were often occupied successively on the same farm and may still survive, although rural depopulation has caused abandonment of many dwellings. Prairie tree plantings often show characteristic "stags-heads" where drought and disease have killed the upper branches, and a network of dugouts and irrigation canals provides water in dry areas. In addition to the basic pattern of settlement, Indian reserves, pre-survey settlements, and the communal farms of some religious groups make their own distinctive contributions to the landscape.

Typical villages

Groups of highly visible grain elevators, set at intervals along the railways, are characteristic sights in the province. Most villages have grown up around such a group, often only on one side of the railway track. A grid street plan, unpaved roads, and one- and two-storey frame houses are typical, and the buildings will include shops, a hotel, and several churches.

Urban centres. Most urban centres resemble the smaller villages and gain character from their often splendid settings, rather than from outstanding buildings. Such settlements as Peace River, in a spectacular river valley, and Drumheller, rambling among extensive badlands, are noteworthy but not untypical. Most towns have a basically rectangular grid pattern, perhaps elongated along river frontage or a railway. Jasper and Banff are laid out in attractive, less regular patterns, focussing on administration buildings in a park setting.

Only six of Alberta's cities had achieved populations of more than 10,000 by the 1970s. Each of these serves a large rural area: Lethbridge and Medicine Hat, the southwest and southeast, respectively; Grande Prairie, the Peace River country; and Red Deer, the central region. Edmonton and Calgary, both with populations over 400,000, have a much wider influence. Each is developing the high-rise buildings and expressways characteristic of cities elsewhere, while retaining enough parkland to remain attractive places to live.

THE PEOPLE OF ALBERTA

Ethnic groups and religion. In terms of national and cultural origins, Alberta is one of the most varied of the Canadian provinces. The early period of settlement around the turn of the century, and immigration following World War II, have contributed to this diversity. The 1961 census designated 18 specific ethnic origins for its purposes, and showed the British Isles group (45%)

to be the largest, while other major groups included German (13.8%); Ukrainian (8%); Scandinavian (7.2%); French (6.3%); and Dutch (4.2%). The Indian population comprised 2.2%, and there were also Polish, Austrian, Russian, Chinese, and Japanese elements. By contrast, 94% of the population spoke only English, although 45% claimed their mother tongue as other than English. Twenty religious denominations were designated for census purposes, with the United Church of Canada (31.5%), the Roman Catholic (22.5%), and the Anglican Church of Canada (11.8%) predominating.

Folk traditions

This human mosaic has provided a valuable cultural legacy for Alberta. This has found expression in festive occasions, costumes, the performing and fine arts, and in handicrafts. In domestic and religious architecture, this heritage is evident in areas of ethnic concentration such as the regions north and east of Edmonton; the irrigated areas of southern Alberta (under Mormon influence); the Nikka Yuko centennial garden in Lethbridge; and the Hutterite settlements around the province.

The population today. Total population given by the 1951 census was 939,501, and the 1961 census showed an increase to 1,331,944. The 1971 census population figure was 1,627,874.

The most densely populated area of Alberta is the central and southwest portion, east of the foothills. The lowest density, averaging about one person per square mile, is in the four census districts comprising the northern portion of Alberta, the foothills and the mountain region, which, together, occupy about three-quarters of the province. By contrast, the two census districts including the cities of Edmonton and Calgary have densities of 85 and 75 persons per square mile, respectively.

Alberta entered the 1970s with 10 cities, 105 towns, 137 villages, and 32 summer villages. The populations of principal cities in 1971 were Edmonton, 438,152; Calgary, 403,319; Lethbridge, 41,217; Red Deer, 27,674; Medicine Hat, 26,518; and Grande Prairie, 13,079. The largest town was St. Albert (11,800), while the smallest village was Munson, with only 43 inhabitants. The predominantly urban provincial living pattern has long been one of single family dwellings, but the proportion of multiple dwelling units has risen sharply.

The birth rate in Alberta, which reached a 1954 peak of 31.8 births per thousand population, subsequently evidenced a continued downward trend, reaching less than 20.0 births per thousand in the late 1960s. By the early 1970s, the death rate was 6.5 per thousand, and the rate of natural increase 13.3 per thousand. Alberta has a young population, entering the 1970s with nearly 60 percent of the population under 25 years of age.

The rapid increase in population has been paralleled by important shifts in many other aspects of life. Whereas, in 1941, 38.5% of the population was urban, by 1970 the figure had reached over 70%. The rural population declined from 489,583 to under 400,000 in the same period. A significant shift was also evident in the occupational distribution of the labour force. Between 1941 and the beginning of the 1970s, the proportion taken up by forestry, agriculture, and fishing declined from 50 to 18%, while the mining, manufacturing, and construction sectors showed an increase for the same period, from 13 to 23%, paralleled in the service sector by an increase from 47 to 59%.

Changes in the labour force

Movement also has occurred in the relative participation of males and females in the labour force. While the proportion of males in the labour force remains at around 80% of the male population, within the female population there has been a rise, from a 1951 proportion of about 20%, to over 30% by 1961, and over 40% ten years later. In the decade 1951–61, the number of males working in agriculture declined by 20,000, while it increased by 24,000 in mining, construction, and manufacturing, and by 68,000 in service industries. The number of females increased in these areas by 10,000, 5,000, and 50,000, respectively. Because of the economic buoyancy of the province, and the available educational and training facilities, workers have been able to move readily into alternative occupations.

Alberta's population will probably continue to increase, especially in the north, as a result of resource development. Rural population will continue to decrease, while urban populations will increase. Authorities project a total population of 2,000,000 for 1981, half of them residing in Edmonton and Calgary.

THE PROVINCIAL ECONOMY

By the early 1970s, the significance of the contribution of various industries to the province, indicated by percentage of total net value, was as follows: mining 34%, construction 22%, manufacturing 20%, and agriculture 19%. Electric power, forestry, fisheries, and trapping together amounted to only 4%. Agriculture dominated the provincial economy before 1945, but subsequently a more balanced economy was produced by the development of industries based on oil, gas, and coal, together with food and wood products.

Extent and distribution of resources. Alberta holds the bulk of Canada's known fossil fuels. Estimated recoverable reserves by the early 1970s included 7,600,000,000 barrels of oil, 100,000,000,000 tons of coal, and enormous quantities of gas. The Athabasca oil sand deposit, which outcrops over 21,000 square miles, holds an additional 300,000,000,000 recoverable barrels. Oil and gas occur widely; and coal is found both in low grade, easily worked beds in the plains, and high grade, but difficult to extract deposits in the foothills. Other mineral resources include sands, gravels, bentonite, clay, limestone, salt, phosphates, and sodium sulphate; sulphur and helium are produced along with the natural gas.

Vastness of mineral deposits

Forests, containing no less than 55,000,000,000 cubic feet of wood, cover 62% of the surface of the province. Aspen and birch (40% by volume), white spruce, and pines (each 27% by volume) are the principal economic species.

Apart from a few areas in the south, there is abundant freshwater in lakes and rivers, and the hydroelectric power potential of the rivers is over 2,000,000 kilowatts. Wildlife and scenery in all parts of the province are an important economic resource, especially when considered from the viewpoint of tourism.

Exploitation of resources. Although fisheries and furs remain important in the undeveloped areas of the province, over 18,400,000 acres (some irrigated) are under crops, with wheat as the major crop, followed by barley, hay, and oats. There is increasing diversification with rapeseed, flaxseed, sugar beets, potatoes, rye, and mustard becoming important. Cattle, numbering some 3,540,000, including 200,000 dairy cattle, are also significant, together with swine, sheep, and poultry.

Forest products were worth over $35,000,000 annually by the early 1970s, and consisted mainly of lumber, plywood, and pulp.

The Leduc oil discovery of 1947 stimulated the development of more than 40 oil fields, which, 20 years later, produced over 325,000,000 barrels annually. The production of natural gas and of sulphur are also important. The Great Canadian Oil Sands Plant, which began production in 1967, yields 45,000 barrels per day of synthetic crude oil. Coal production declined as the steam locomotive went out of circulation, but nearly 4,000,000 tons were produced annually by the early 1970s, over one-fifth of which was exported to Japan. Coal produces over two-thirds of the province's electricity supply, the rest coming from hydroelectric plants.

Major manufacturing industries are foods and beverages, petroleum and chemical products, metals, and wood and paper. Construction is linked to the development of gas and oil resources but also includes a proportion of residential building.

Management of the economy

Natural resources are largely developed by private industry, controlled by the Canadian and provincial governments through legislation and regulations. Much of the industry is managed and financed outside Canada, notably in the United States. Extensive government research aids resource development, and practical assistance is provided to agriculture. Alberta's record of industrial peace has been the best in Canada.

TRANSPORTATION

Southern Alberta communications are dominated by east–west routes that thread the few mountain passes and spread over the plains. There are good communications south into the U.S., and from Edmonton air and surface routes fan out to serve the Arctic from Alaska to the northern Canadian islands. After the fur traders' canoes had opened up the area, steamers used the Saskatchewan River until the development of a railway network. Today the Athabasca and Slave rivers remain active in the carrying of freight to the Arctic. The east-west Canadian Pacific and Canadian National railways have a network of branch lines serving agricultural areas, while the Northern Alberta and Great Slave Lake railways serve northern areas. The Alberta Resources Railway was completed in 1969. Oil and gas pipelines link Alberta to markets as far away as Quebec and California.

Alberta has 64,868 miles of paved and gravelled highways, including the Trans-Canada and Yellowhead east–west routes, and the Mackenzie and Alaska highways to the north and northwest. All-terrain vehicles are of growing importance in the north.

Air transport has been used in development of northern resources since 1921; half a century later, air charter services were provided by 37 firms, using 47 licensed airports, and scheduled regional, national, and international services were provided by five airlines.

ADMINISTRATION AND SOCIAL CONDITIONS

Provincial government and politics. The British North America Act (1867) and the Alberta Act (1905) provide the constitutional framework for the province. In line with other Canadian provinces, Alberta has a lieutenant governor (functioning as representative of the Queen), an elected legislative assembly, and an executive council. The lieutenant governor is appointed by the federal government for a five-year term. The legislative body comprises a member from each of 75 constituencies elected by Canadian citizens and British subjects of 18 years of age and over. The political party that commands support of the majority of members forms the government, and the party leader generally becomes the premier. The term of office of a government is five years. The premier heads the executive council of 22 cabinet members which is responsible for administering laws and appropriations approved by the legislature. From 1935 the Social Credit Party was dominant, but in 1971 the Progressive Conservative Party swept into power.

In 1970 the legislature passed annual appropriations for the government's first billion dollar budget, at $1,146,000,000, about 34% of which was earmarked for education, with 20% going to health services. Expenditure a decade earlier had totalled only $307,900,000. Sources of revenue are diversified with about 26% emanating from natural resources, 21.5% from the Alberta income tax, 17.4% from the federal government, and the remainder from the fuel oil tax and liquor profits. Other sources include financing of $112,000,000 from provincial reserves and from borrowing.

Law. The highest judicial authority in Alberta is the supreme court, comprised of appellate and trial divisions. There are two district courts, one each for northern and southern Alberta, and there are also juvenile courts and family courts. The governor general of Canada appoints superior, district, and county judges, while family court judges, police magistrates, and justices of the peace are appointed provincially.

The Mounties

There is no provincial police force. The Royal Canadian Mounted Police is under contract to the province and to some municipalities, while in others there are municipal police forces. Correctional and rehabilitation services are provided under federal and provincial jurisdiction, while defense is a federal responsibility.

Municipal administration. Municipal administration takes a variety of forms, appearing as cities, towns, villages, counties, municipal districts, improvement districts, and special areas. Mayors, councils, and school and hospital boards are elected, while the major sources of municipal revenue are the property tax, the local-improvement tax, and the business tax. Substantial grants are nevertheless made by the provincial authorities, although some municipalities derive additional revenue from the operation of utilities.

Education. Education is on a 12-grade system, with over 420,000 pupils enrolled annually at elementary and secondary levels by the early 1970s, including over 70,000 in the separate Roman Catholic school system, also tax supported. There were also over 5,000 students in institutes of technology and nearly 27,000 at the university level. The province outlines the school curriculum and maintains general supervision of the elementary and secondary school systems, but administration is through elected school boards. The School Foundation Program Fund, pooled from provincial grants and local property taxes, provides funds—distributed back to the school boards—for the school systems.

The two school systems

Institutes of technology in both Edmonton and Calgary provide extensive vocational training. Universities are located in Edmonton, Calgary, and Lethbridge, and a fourth may be built in St. Albert. There are also eight junior colleges with university affiliations and three agricultural colleges. Financial assistance for students is obtainable through the Alberta students assistance board. Extensive credit or noncredit adult evening programs are also available, notably in the larger cities.

Health and welfare. The Alberta health-care insurance plan and the Alberta hospitalization benefit plan provide for all basic health services through payment of an annual premium but with subsidies for low income families. The Alberta Blue Cross plan and private insurance plans provide for certain supplementary hospital, drug, and other services, while mental hospitals are provided under a special provincial act. The costs of social assistance programs are borne jointly by federal, provincial, and municipal authorities.

Income. Although the minimum wage under provincial regulations was raised in 1970 from $1.25 to $1.55 per hour for employees 18 years of age and over, the cost of living, as in other parts of Canada, has continued to rise. Alberta entered the 1970s with the Edmonton consumers' price index at 126.1, with 1961 equalling 100.

CULTURAL LIFE AND INSTITUTIONS

Museums and historic sites. Although the province is young, interest in preserving its heritage and fine arts is strong. By 1971 there were over 85 museums, including nine art galleries, and over 100 provincial historic site markers. The new Provincial Museum and Archives of Alberta, in Edmonton, was opened in 1967; it interprets the province to visitors through colourful displays, collections, and educational and reference services. The Glenbow-Alberta Institute Museum, Calgary, has collections and displays of art, history, and ethnology covering western Canadiana and many other parts of the world. Heritage Park, Calgary, has preserved many historic buildings, and the Edmonton Art Gallery has a permanent collection of Canadian art. Art galleries in Alberta cooperate in circulating exhibits, with the Ace Foundation organizing "museumobiles" that tour the province. Edmonton and Calgary are the locations of two new planetariums, as well as three zoos.

Libraries. Public libraries were an early development in Alberta, and there were 73 municipal, 83 community, and 2 regional library systems by the early 1970s, with 9 and 12 branch libraries, respectively, in the Edmonton and Calgary systems. The legislative library in the Legislature Building, Edmonton, together with school, college, and university libraries (including the University of Alberta extension library, available to all parts of the province), provide supplementary facilities.

Performing and creative arts. Two modern provincial auditoriums (built in 1955 in Edmonton and Calgary) and a new art centre in Lethbridge help to attract touring artists and to develop Alberta talent. There are three symphony orchestras and many amateur and several professional performing art groups, while the Banff School of Fine Arts is an outstanding complex for instruction in both the performing and the fine arts.

A significant aspect of Alberta's cultural life, however, is the growth and diversity of individual and group creative activities—painting, music, dancing, theatre, and writing. The cultural development branch of the provincial government gives practical aid and encouragement in all fields through such programs as performing-arts tours and exhibitions, leadership training programs and workshops in all phases of performing and expressive arts, consultative services, and financial assistance.

Traditional cultural and sporting events

Further diversity is evident in ethnic arts and Indian crafts; in annual events such as the Banff Indian Days, the Calgary Stampede, and the Edmonton Klondike Days; in local rodeos and fairs; and in activities such as skiing, mountaineering, camping, hiking, and ice skating, together with curling, ice hockey, and football.

Radio broadcasts are made through 26 privately owned stations, the Alberta Government Telephones station, and the transmitters of the Canadian Broadcasting Corporation. There is one French-language radio station, in Edmonton, which with other stations also broadcasts programs in languages other than French and English. Television stations are located in six cities, and there are numerous satellite transmitters serving small communities. There is one French-language television station in Edmonton, which also provides time for educational television programs in English. There are seven daily newspapers and 87 weekly newspapers in the province.

PROBLEMS AND PROSPECTS

The continued prosperity of Alberta seems likely as long as the world needs food and fuels, but prosperity has brought problems: a stable society based on a declining agricultural population must assimilate both its underprivileged native peoples and many newcomers from other provinces and countries without diminishing a rich heritage. Alberta, still developing, has the chance to plan for, and achieve, an optimum population. It can learn from urbanization problems elsewhere and plan cities for future needs, taking into account the hazards of pollution. Resource development presents the problem of balancing long-term social and environmental needs against short-term economic gains, and it raises the question of control of largely foreign investment and markets. The development of nonrenewable resources with a minimum of irretrievable damage to the environment is of particular importance, so that beauty will remain as a provincial resource when the minerals have gone.

(R.O.H./D.A.E.S.)

BIBLIOGRAPHY. R. KROETSCH, *Alberta* (1968), a general account of the province, by a local novelist; W.C. WONDERS *et al.* (eds.), *Atlas of Alberta* (1969), a comprehensive survey of the province; W.G. HARDY (ed.), *Alberta: A Natural History* (1967), covers the geology and ecology, and has an extensive bibliography. For current information, see the annual reports of the government of Alberta: BUREAU OF STATISTICS, *Alberta Industry and Resources*, for economic information; and CULTURAL DEVELOPMENT BRANCH, *Government of Alberta*, a useful source on cultural affairs. Three basic sources, also published by the government, are: W.R. SALT and A.L. WILK, *The Birds of Alberta* (1958); J.D. SOPER, *The Mammals of Alberta* (1964); and M.J. PAETZ and J.S. NELSON, *The Fishes of Alberta* (1970). See also J.P. GILLESE (comp.), *Chinook Arch*, 2nd ed. (1968), an anthology of Alberta writing. Historical materials include: W.G. HARDY (ed.), *Alberta: A Natural History* (1967), see especially ch. 14, "The Impact of the White Man," by JAMES G. MACGREGOR; GRANT MACEWAN, *Between the Red and the Rockies* (1953), a popular history of agricultural development on the Canadian prairies; C.M. MACINNES, *In the Shadow of the Rockies* (1930), a popular history of southern Alberta to 1930; and E.C. WATERTON, *Indians in Alberta, Tribal Locations in the Historic Period*, Provincial Museum and Archives of Alberta Notes, No. 10 (1971). The following, part of the "Canadian Frontiers of Settlement Series," contain detailed scientific accounts of settlement: C.A. DAWSON and R.W. MURCHIE, *The Settlement of the Peace River Country: A Study of a Pioneer Area* (1934); W.A. MACKINTOSH, *Prairie Settlement, the Geographical Setting* (1934); and ARTHUR S. MORTON and CHESTER MARTIN, *History of Prairie Settlement and Dominion Lands Policy* (1938).

(R.O.H./D.A.E.S./R.B.D.)

Alberti, Leon Battista

Humanist, poet and scholar, architect and art theorist, engineer, mathematician, cartographer and cryptographer, Leon Battista Alberti is the prototype of the Renaissance "universal man." He was in the vanguard of almost every major development in the cultural life of early Renaissance Italy. He has been admired for his many-sided nature from his day to our own, as has Leonardo da Vinci (1452–1519), who followed him by half a century and very much resembles him in this respect. Yet in Alberti's case, unity as much as versatility typifies the man and his accomplishments. Leonardo's genius carried him further than Alberti: he saw more and saw more deeply. But Leonardo's vision has a disturbingly "modern," fragmentary character, whereas Alberti attained a completeness in thought and life that fulfilled the Renaissance ideals of measure and harmony. His intellectual and artistic pursuits were all of a piece; and he struck a unique balance between theory and practice, realizing this dominant aspiration of his age at the very moment social and political events had begun to cause it to fade.



Leon Battista Alberti, self-portrait plaque, bronze, *c.* 1435. In the National Gallery of Art, Washington, D.C., 20.1 cm × 13.5 cm.
By courtesy of the National Gallery of Art, Washington, D.C., Samuel H. Kress Collection

Childhood and education. The society and class into which Alberti was born endowed him with the intellectual and moral tendencies he was to articulate and develop over a lifetime. He belonged to one of the wealthy merchant-banker families of Florence. At the time of his birth in 1404, the Alberti were in exile, expelled from Florence by the oligarchical government then dominated by the Albizzi family. Alberti's father, Lorenzo, was managing the family's concerns in Genoa, where Battista was born. Shortly thereafter he moved to Venice, where he raised Battista (Leo or Leon was a name adopted in later life) and his elder brother, Carlo. Both sons were illegitimate, the natural offspring of Lorenzo and a Bolognese widow, but they were to be Lorenzo's only children and his heirs. An affectionate and responsible father, Lorenzo provided his sons with a Florentine stepmother (whom he married in 1408), and he attended carefully to their education.

Mathematical training

It was from his father that Battista received his mathematical training. The useful intellectual tools of the businessman inspired in him a lifelong love for the regular, for rational order, and a lasting delight in the practical application of mathematical principles. "Nothing pleases me so much," Alberti was to have a figure in one of his

dialogues remark, "as mathematical investigations and demonstrations, especially when I can turn them to some useful practice as Battista here did, who drew from mathematics the principles of painting [perspective] and also his amazing propositions on the moving of weights." As in Leonardo's case, mathematics led Alberti into several seemingly disparate fields of learning and practice. At one stroke, it resolved a diversity of problems and awakened an appreciation of the rational structure and processes of the physical world.

His early formal education was Humanistic. At the age of ten or 11, Alberti was sent to boarding school in Padua. Here he was given the classical Latin training that was to be denied to Leonardo, illegitimate son of a poor notary in a rustic village of Tuscany. The "new learning" was largely literary, and Alberti emerged from the school an accomplished Latinist and literary stylist. Relishing his skill as a classicist, he wrote a Latin comedy at the age of 20 that was acclaimed as the "discovered" work of a Roman playwright—and was still published as a Roman work in 1588 by the famous Venetian press of Aldus Manutius. But it was the content rather than the form of the classical authors that absorbed Alberti as a youth and throughout his life. As for most Humanists, the literature of ancient Rome opened up for him the vision of an urbane, secular, and rational world that seemed remarkably similar to the emerging life of the Italian cities and met its cultural needs. He brought his own emotional and intellectual tendencies to "the ancients," but from them he drew the conceptual substance of his thought.

Alberti completed his formal education at the University of Bologna in an apparently joyless study of law. His father's death and the unexpected seizure of his legacy by certain members of the family brought him grief and impoverishment during his seven-year stay at Bologna, but he persisted in his studies. After receiving his doctorate in canon law in 1428, he chose to accept a "literary" position as a secretary rather than pursue a legal career. By 1432 he was a secretary in the Papal Chancery in Rome (which supported several Humanists), and he had a commission from a highly placed ecclesiastical patron to rewrite the traditional lives of the saints and martyrs in elegant "classical" Latin. From this point on, the church was to provide him with his livelihood. He took holy orders, thus receiving in addition to his stipend as a papal secretary an ecclesiastical benefice, the priory of Gangalandi in the diocese of Florence, and some years later Nicholas V conferred upon him as well the rectory of Borgo San Lorenzo in Mugello. Although he led an exemplary, and apparently a celibate, life, there is almost nothing in his subsequent career to remind one of the fact that Alberti was a churchman. His interests and activities were wholly secular and began to issue in an impressive series of Humanistic and technical writings.

Contribution to philosophy, science, and the arts. The treatise "Della famiglia" ("On the Family"), which he began in Rome in 1432, is the first of several dialogues on moral philosophy upon which his reputation as an ethical thinker and literary stylist largely rests. He wrote these dialogues in the vernacular, expressly for a broad urban public that would not be skilled in Latin: for the *non litteratissimi cittadini*, as he called them. Based upon classical models, chiefly Cicero and Seneca, these works brought to the day-to-day concerns of a bourgeois society the reasonable counsel of the ancients—on the fickleness of fortune, on meeting adversity and prosperity, on husbandry, on friendship and family, on education and obligation to the common good. They are didactic and derivative, yet fresh with the tone and life-style of the quattrocento (the 1400s). In Alberti's dialogues, the ethical ideals of the ancient world are made to foster a distinctively modern outlook: a morality founded upon the idea of work. Virtue has become a matter of action, not of right thinking. It arises not out of serene detachment but out of striving, labouring, producing.

Moral philosophy

This ethic of achievement, which corresponds to the social reality of his youth, found ready acceptance in the urban society of central and north Italy in which Alberti moved after 1434. Travelling with the papal court of Eugenius IV to Florence (the ban of exile against his family was lifted with the restoration of Medici influence), Bologna, and Ferrara, Alberti made several congenial and fruitful contacts. The writings, both the Latin and vernacular ones, that he dedicated to his new associates are imbued with his characteristic notions of work, practice, productive activity; and he took upon himself in turn the technical and practical problems that were absorbing his friends and patrons. In Florence his close associations with the sculptor Donatello and the architect Brunelleschi led to one of his major achievements: the systematization of painter's perspective. The book *On Painting*, which he wrote in 1435, set forth for the first time the rules for drawing a picture of a three-dimensional scene upon the two-dimensional plane of a panel or wall. It had an immediate and profound effect upon Italian painting and relief work, giving rise to the correct, ample, geometrically ordered space of the perspectival Renaissance style. Later perspectival theorists, such as the painter Piero della Francesca and Leonardo, elaborated upon Alberti's work, but his principles remain as basic to the projective science of perspective as Euclid's do to plane geometry.

Achievements in perspective and cartography

His friendship with the Florentine cosmographer Paolo Toscanelli was of comparable practical and scientific importance. It was Toscanelli who provided Columbus with the map that guided him on his first voyage. Alberti seems to have collaborated with him in astronomy rather than geography, but the two sciences were closely bound at the time (and bound to perspective) by the conceptions and methods of geometric mapping rediscovered in the writings of the ancient astronomer and geographer Ptolemy. Alberti's distinctive contribution to this current of thought took the form of a small treatise on geography, the first work of its kind since antiquity. It sets forth the rules for surveying and mapping a land area, in this case the city of Rome, and it was probably as influential as his earlier treatise on painting. Although it is difficult to trace the historical connections, the methods of surveying and mapping and the instruments described by Alberti are precisely those that were responsible for the new scientific accuracy of the depictions of towns and land areas that date from the late 15th and early 16th centuries.

At the Este court in Ferrara, where Alberti was first made a welcome guest in 1438, the Marchese Leonello encouraged (and commissioned) him to direct his talents toward another field of endeavour: architecture. Alberti's earliest effort at reviving classical forms of building still stands in Ferrara, a miniature triumphal arch that supports an equestrian statue of Leonello's father. Leonello inspired a great Humanistic undertaking as well as a mode of artistic practice on Alberti's part by urging him to restore the classic text of Vitruvius, architect and architectural theorist of the age of the Roman emperor Augustus. With customary thoroughness, Alberti embarked upon a study of the architectural and engineering practices of antiquity that he continued when he returned to Rome in 1443 with the papal court. By the time Nicholas V became pope in 1447, Alberti was knowledgeable enough to become the Pope's architectural adviser. The collaboration between Alberti and Nicholas V gave rise to the first grandiose building projects of Renaissance Rome, initiating among other works the reconstruction of St. Peter's and the Vatican Palace. As the Este prince was now dead, it was to Nicholas V that Alberti dedicated in 1452 the monumental theoretical result of his long study of Vitruvius. This was his *De re aedificatoria* (*Ten Books on Architecture*), not a restored text of Vitruvius but a wholly new work, that won him his reputation as the "Florentine Vitruvius." It became a bible of Renaissance architecture, for it incorporated and made advances upon the engineering knowledge of antiquity, and it grounded the stylistic principles of classical art in a fully developed aesthetic theory of proportionality and harmony.

Architect and architectural theorist

During the final 20 years of his life, Alberti carried out his architectural ideas in several outstanding buildings.

The facades of Sta. Maria Novella and the Palazzo Rucellai, both executed in Florence for the merchant Giovanni Rucellai, are noted for their proportionality, their perfect sense of measure. They are worthy successors to the art of Brunelleschi, initiator of the Florentine quattrocento style of architecture. Other buildings look forward to the 16th century, particularly to Bramante, the architect of St. Peter's. The classical severity of Alberti's Tempio Malatestiano, commissioned by Sigismondo Malatesta, the ruler of Rimini, and the new sense of volume and amplitude of the majestic church of S. Andrea, which he designed for Lodovico Gonzaga, the Humanist prince of Mantua, announce the fullness of the High Renaissance style. Alberti was not only the foremost theorist of Renaissance architecture: he had become one of its great practitioners as well.

Architecture preoccupied him during the 1450s and 1460s and he travelled a great deal to the various cities and courts of Renaissance Italy, but Rome and Florence remained his intellectual homes and he continued to cultivate the interests they had always stimulated. In Rome, where republican life was precluded by the papal government, he was absorbed by technical and scientific matters. His response to certain problems entertained by members of the Papal Chancery led to two highly original works in this category. One is a grammar book, the first Italian grammar, by which he sought to demonstrate that the Tuscan vernacular was as "regular" a language as Latin and hence worthy of literary use. The other is a pioneer work in cryptography: it contains the first known frequency table and the first polyalphabetic system of coding by means of what seems to be Alberti's invention, the cipher wheel. Although he had been dismissed from the Papal Chancery in 1464 because of the retrenchment ordered by Pope Paul II, Alberti undertook this study, of obvious importance to the papacy, at the request of a friend who stayed on as a papal secretary. In this as in all his projects, his generous disposition, practical attitude, and ethical conviction combined to make him employ his intellectual gifts in some "useful" work—useful to the artistic, cultivated, and courtly circles in which he moved: to painters and builders, map makers and astronomers, Humanists, princes, and popes; and useful to the bourgeois society from which he came, his *non litteratissimi cittadini*. In all of his work, his versatility remained bound to the social outlook that characterized the "civic Humanism" of Florence.

Grammarian and cryptographer

It is fitting that his final and finest dialogue should be set in Florence and be written in the clear Tuscan prose he had helped to regularize and refine. Although the republicanism of Florence was now eclipsed, and Alberti now moved as a familiar in the circle of the princely Lorenzo de'Medici, *De iciarchia* ("On the Man of Excellence and Ruler of His Family") represents in full flower the public-spirited Humanism of the earlier bourgeois age to which he belonged. Alberti is its chief protagonist, and no more appropriate figure is conceivable. For this dialogue, more than any other, celebrates the union of theory and practice Florentine Humanism had attained and the ethic of achievement and public service that he himself had come to exemplify. *De iciarchia* was completed just a few years before his death. He died in Rome in 1472, "content and tranquil," according to the 16th-century biography of Giorgio Vasari, leaving behind "a most honorable name."

MAJOR WORKS

ARCHITECTURE: Palazzo Rucellai, Florence (*c.* 1445–51); S. Francesco, Tempio Malatestiano, Rimini (1447–50); facade, Sta. Maria Novella, Florence (1456–70); S. Sebastiano, Mantua (1460–70); Rucellai Chapel, Florence (1467); S. Andrea, Mantua (*c.* 1470); SS. Annunziata, Florence (designed *c.* 1470).

WRITINGS: *Della famiglia* (written 1435–44); *De pictura praestantissima* (1435); *Della pittura* (1436; *On Painting*, 1956); *De re aedificatoria* (promulgated in 1452; published in 1485; *Ten Books on Architecture*, 1955); *De iciarchia* (1468).

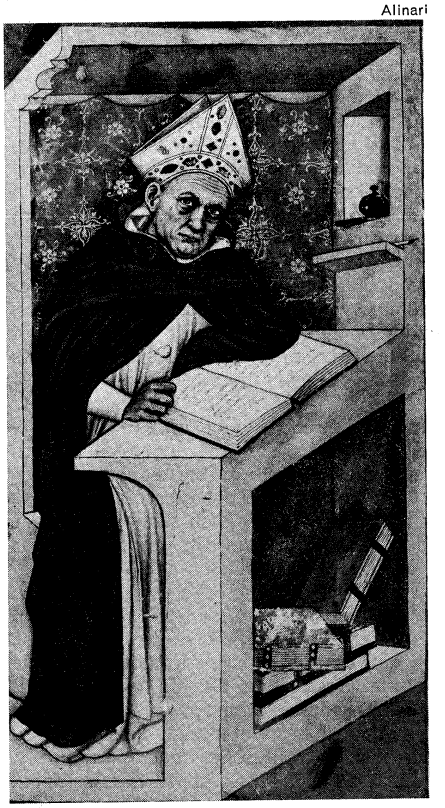
BIBLIOGRAPHY. For biography, bibliography, and analysis of Alberti's Humanistic, scientific, and artistic work, see JOAN GADOL, *Leon Battista Alberti: Universal Man of the Early Renaissance* (1969). CECIL GRAYSON gives a succinct biography, bibliography, and résumé of Alberti's writings in the *Dizionario biografico degli Italiani*, vol. 1, pp. 702–709 (1960). Among the many special studies of different aspects of Alberti's work, those of most general interest are the articles by BRUNO ZEVI *et al.*, in the *Encyclopedia of World Art*, vol. 1, col. 188–216 (1959); SIR KENNETH CLARK, "Leon Battista Alberti on Painting," *Proceedings of the British Academy*, 30:283–302 (1944); CECIL GRAYSON, "The Humanism of Alberti," *Italian Studies*, 12:37–56 (1957); and "Universal Man," in PHILIP WIENER (ed.), *Dictionary of the History of Ideas* (1972). The following studies of some of the fields in which Alberti worked give interesting appraisals of the contributions he made to them: DAVID KAHN, *The Codebreakers: The Story of Secret Writing* (1967); ERWIN PANOFSKY, *Renaissance and Renascences in Western Art* (1960); RUDOLF WITTKOWER, *Architectural Principles in the Age of Humanism*, 3rd ed. rev. (1962); WILLIAM HARRISON WOODWARD, *Studies in Education During the Age of the Renaissance, 1400–1600* (1906, reprinted 1965).

(J.Ga.)

Albertus Magnus, Saint

Albertus Magnus (variously known as Albert the Great or Albert of Cologne), a 13th-century philosopher, scientist, and theologian who is called *doctor universalis* for the great breadth of his knowledge and intellectual interest, is best known as a proponent of Aristotelianism at the University of Paris and as the teacher of Thomas Aquinas. He established the study of nature as a legitimate science within the Christian tradition. The most prolific writer of his century, he was the only scholar of his age to be called "the Great," and this title was used even before his death.

Alinari



Albertus Magnus, detail of a fresco by Tommaso da Modena, *c.* 1352. In the Church of S. Nicolo, Treviso.

Albertus, born about 1200 at Lauingen on the Danube, was the eldest son of a wealthy German lord. After his early schooling, he went to the University of Padua, where he studied the liberal arts. In the summer of 1223, Jordan of Saxony, master general of the recently founded Dominican Order, arrived in Padua hoping to attract new members into the order. Ten students soon sought admission; one of these seems to have been Albertus.

After overcoming the strong opposition of his family, Albertus joined the order. He continued his studies at Padua, Bologna, and in Germany and then taught theology at several convents in Germany, lastly at Cologne.

Studies in Paris

Sometime before 1245 he was sent to the Dominican convent of Saint-Jacques at the University of Paris, where he came into contact with the works of Aristotle, newly translated from Greek and Arabic, and with the commentaries on Aristotle's works by Averroës, a 12th-century Spanish-Arabian philosopher. At Saint-Jacques he lectured on the Bible for two years and then for another two years on Peter Lombard's *Sentences*, the theological textbook of the medieval universities. In 1245 he was graduated master in the theological faculty and obtained the Dominican chair "for foreigners."

It was probably at Paris that Albertus began working on a monumental presentation of the entire body of knowledge of his time. He wrote commentaries on the Bible and on the *Sentences;* he alone among medieval scholars made commentaries on all the known works of Aristotle, both genuine and spurious, paraphrasing the original but frequently adding "digressions" in which he expressed his own observations, "experiments," and speculations. The term experiment for Albertus indicates a careful process of observing, describing, and classifying. His speculations gave wide room to Neoplatonic thought. Apparently in response to a request that he explain Aristotle's *Physics*, Albertus undertook—as he states at the beginning of his *Physica*—"to make . . . intelligible to the Latins" all the branches of natural science, logic, rhetoric, mathematics, astronomy, ethics, economics, politics, and metaphysics. While he was working on this project, which took about 20 years to complete, he probably had among his disciples Thomas Aquinas, who arrived at Paris late in 1245.

Albertus distinguished the way to knowledge by revelation and faith from the way of philosophy and of science; the latter follows the authorities of the past according to their competence, but it also makes use of observation and proceeds by means of reason and intellect to the highest degrees of abstraction. For Albertus these two ways are not opposed; there is no "double truth"—one truth for faith and a contradictory truth for reason. All that is really true is joined in harmony. Although there are mysteries accessible only to faith, other points of Christian doctrine are recognizable both by faith and by reason—*e.g.*, the doctrine of the immortality of the individual soul. He defended this doctrine in several works against the teaching of the Averroists (Latin followers of Averroës), who held that only one intellect, which is common to all human beings, remains after the death of man.

Albertus' lectures and publications gained him great renown. He came to be quoted on an equal footing with such other recognized authorities as the Arabian philosophers Avicenna and Averroës and even Aristotle himself. Roger Bacon, a contemporary English scholar who was by no means friendly toward Albertus, spoke of him as "the most noted of Christian scholars."

Years in Germany and Italy

In the summer of 1248, Albertus was sent to Cologne to organize the first Dominican *studium generale* ("general house of studies") in Germany. He presided over the house until 1254 and devoted himself to a full schedule of studying, teaching, and writing. During this period his chief disciple was Thomas Aquinas, who returned to Paris in 1252. The two men maintained a close relationship even though doctrinal differences began to appear. From 1254 to 1257 Albertus was provincial of "Teutonia," the German province of the Dominicans. Although burdened with added administrative duties, he continued his writing and scientific observation and research.

In 1256 Albertus was ordered by Pope Alexander IV to the papal court at Anagni to join Thomas and Bonaventure, a noted Franciscan scholar, in defending the mendicant orders against the masters of the University of Paris. The mendicant orders had introduced a new form of religious life in which the members renounced all personal and community property and depended upon their work —especially preaching and teaching—to subsist. Because the friars, as the mendicants were called, were not attached to any one monastery or diocese, their superiors had great freedom to send them wherever they were needed.

The Dominicans and the Franciscans, the first models of this kind of religious life, had both established at Paris influential communities that aroused the antagonism of William of Saint-Amour and other university masters who were members of the secular, or diocesan, clergy. The three friars were successful in defending the right of the mendicants to teach at the university.

Albertus resigned the office of provincial in 1257 and resumed teaching in Cologne. In 1259 he was appointed by the Pope to succeed the Bishop of Regensburg, who had been removed from office because of irregularities in the diocese. Despite his own reluctance and the objections of Humbert of Romans, general of the Dominican order, he was installed as bishop in January 1260. After Alexander IV died in 1261, Albertus was able to resign his episcopal see. He then returned to his order and to teaching at Cologne, but, as a bishop, he was in some part exempt from the order's control and could dispose of his time and even of his revenues more as he wished. From 1263 to 1264 he was legate of Pope Urban IV, preaching the crusade throughout Germany and Bohemia; subsequently, he lectured at Würzburg and at Strasbourg. In 1270 he settled definitively at Cologne, where, as he had done previously in 1252 and in 1258, he made peace between the archbishop and his city.

During his final years he made two long journeys from Cologne. In 1274 he attended the second Council of Lyons, France, and spoke in favour of acknowledging Rudolf of Habsburg as German king. In 1277 he travelled to Paris to uphold the recently condemned good name and writings of Thomas Aquinas, who had died a few years before, and to defend certain Aristotelian doctrines that both he and Thomas held to be true.

Albertus died in Cologne on November 15, 1280. In 1931 he was canonized a saint and proclaimed a doctor of the church; in 1941 he was named patron of those who study the natural sciences.

BIBLIOGRAPHY. The complete edition of Albert's works, undertaken by the Albertus-Magnus-Institut of Cologne, includes for the first time his *De bono; Super Isaiam; Liber de natura et origine animae, Liber de principiis motus processivi, Quaestiones super de animalibus; De sacramentis, De incarnatione, De resurrectione; Metaphysica; Super Ethica; De anima; De caelo et mundo;* and *De divinis nominibus* (1951–71). For detailed studies see FRIEDRICH UEBERWEG and BERNHARD GEYER (eds.), *Die patristische und scholastische Philosophie* (1928, reprinted 1951); THOMAS M. SCHWERTNER, *St. Albert the Great* (1932); LYNN THORNDIKE, *A History of Magic and Experimental Science*, vol. 2 (1923); and H.C. SCHEEBEN, *Albertus Magnus* (1955), in German. For bibliography see the *Revue Thomiste*, vol. 36 (1931); and MICHEL SCHOOYANS, *Bibliographie philosophique de Saint Albert le Grand (1931–1960)* (1961).

Albuquerque, Afonso de

The second governor of the Portuguese colonies in India and the conqueror of Goa and Malacca, Afonso de Albuquerque was the organizer of the Portuguese hegemony in the Orient.

He was born at Alhandra, near Lisbon, in 1453, the second son of the senhor of Vila Verde. His paternal great-grandfather and grandfather had been confidential secretaries to Kings John I and Edward (Duarte); his maternal grandfather had been admiral of Portugal. Albuquerque served ten years in North Africa, where he gained early military experience crusading against Muslims. He was present at Afonso V's conquest of Arzila and Tangier in 1471. King John II (ruled 1481–95) made him master of the horse, a post Albuquerque held throughout the reign. In 1489 he again served in North Africa at the defense of Graciosa. Under John's successor, Manuel I, Albuquerque was less prominent at court but again served in Morocco.

Although Albuquerque made his mark under the stern John II and gained his experience in Africa, his reputa-

tion rests on his service in the East. When Vasco da Gama returned to Portugal in 1499 from his pioneering voyage around the Cape of Good Hope to India, King Manuel straightway sent a second fleet under Pedro Álvares Cabral to open relations and trade with the Indian rulers. The Muslim traders who had monopolized the distribution of spices turned the zamorin, or Hindu prince of Calicut, against the Portuguese. His dependency, Cochin, on the southwestern Indian coast, however, welcomed them. In 1503 Albuquerque arrived with his cousin Francisco to protect the ruler of Cochin, where he built the first Portuguese fortress in Asia and placed a garrison. After setting up a trading post at Quilon, he returned to Lisbon in July 1504, where he was well received by Manuel and participated in the formulation of policy. In 1505 Manuel appointed Dom Francisco de Almeida first governor in India, with the rank of viceroy. Almeida's object was to develop trade and aid the allies of the Portuguese. Albuquerque left Lisbon with Tristão da Cunha in April 1506 to explore the east coast of Africa and build a fortress on the island of Socotra to block the mouth of the Red Sea and cut off Arab trade with India. This done (August 1507), Albuquerque captured Hormuz (Ormuz), an island in the channel between the Persian Gulf and the Gulf of Oman, to open Persian trade with Europe. His project of building a fortress at Hormuz had to be abandoned because of differences with his captains, who departed for India. Albuquerque, though left with only two ships, continued to raid the Persian and Arabian coasts.

Campaign against the Arabs

King Manuel appointed Albuquerque to succeed Almeida at the end of his term, though without the rank of viceroy. When Albuquerque reached India in December 1508, Almeida had crushed the improvised sea force of Calicut, but a navy from Egypt had defeated and killed his son. Insisting on retaining power until he had avenged his son's death, Almeida had Albuquerque imprisoned. Almeida defeated the Muslims off Diu in February 1509, and it was only in the following November, with the arrival of a fleet from Portugal, that he finally turned his office over to Albuquerque.

Albuquerque's plan was to assume active control over all the main maritime trade routes of the East and to establish permanent fortresses with settled populations. His attempt to seize Cochin in January 1510 was unsuccessful. By February Albuquerque had realized that it was better to try to supplant the Muslims; assisted by a powerful corsair named Timoja, he took 23 ships to attack Goa, long ruled by Muslim princes. He occupied it in March 1510, was forced out of the citadel by a Muslim army in May, and was finally able to carry it by assault in November. The Muslim defenders were put to the sword.

Capture of Goa

After this victory over the Muslims, the Hindu rulers accepted the Portuguese presence in India. Albuquerque planned to use Goa as a naval base against the Muslims, to divert the spice trade to it, and to use it to supply Persian horses to the Hindu princes. By marrying his men to the widows of his victims he would give Goa its own population, and its supplies would be assured by the village communities under a special regime. After providing for the government of Goa, Albuquerque embarked on the conquest of Malacca, on the Malay Peninsula, the immediate point of distribution for the Spice Islands and points east. He took that port in July 1511, garrisoned it, and sent ships in search of spices.

In the meantime Goa was again under heavy attack. He left in January 1512 and relieved Goa. Having established himself there and having gained control over the movement of goods by a licensing system, Albuquerque again turned to the Red Sea, taking a force of Portuguese and Indians. Because Socotra was inadequate as a base, he attempted to take Aden, but his forces proved insufficient. He thereupon explored the Arabian and Abyssinian coasts. Returning to India, he finally subdued Calicut, hitherto the main seat of opposition to the Portuguese.

In February 1515 he again left Goa with 26 ships for Hormuz, gaining control of part of the island. He was taken ill in September and returned to Goa. On the way he learned that he had been superseded by his personal enemy, Lope Soares; he died embittered on shipboard off Goa on December 15.

Albuquerque's plans derived from the crusading spirit of John II and others. He did not allow himself to be diverted from his schemes by considerations of mercantile gain. His boldest concepts, such as turning the Persians against the Turks or ruining Egypt by diverting the course of the Nile, were perhaps superhuman, but so perhaps was his achievement.

BIBLIOGRAPHY. Biographies of Albuquerque in English are: H. MORSE STEPHENS, *Albuquerque*, in "Rulers of India Series" (1892); E. PRESTAGE, *Afonso de Albuquerque, Governor of India* (1929), a brief, scholarly account; and E. SANCEAU, *Indies Adventure* (1938). Albuquerque's remarkable series of reports and letters have been published by the Lisbon Academy of Sciences, *Cartas*, 7 vol. (1884–1935). They were drawn on by his son Brás (later called Afonso) de Albuquerque for the *Commentarios de Afonso Dalboquerque*, completed in 1557, revised in 1576, and reported in 1774. The Hakluyt Society published an English translation by W. DE G. BIRCH, *The Commentaries of the Great Afonso Dalboquerque*, 4 vol. (1875–84).

(H.V.L.)

Alchemy

Alchemy was the name given, in Latin Europe, from the 12th century, to an aspect of thought that corresponds to astrology, which is apparently an older ideology. Both represent attempts to discover the relationship of man to the cosmos and to exploit that relationship to his benefit. The first of these objectives may be called scientific, the second technological. Astrology is concerned with man's relationship to "the stars" (including the members of the solar system); alchemy, with terrestrial nature. But the distinction is far from absolute, since both are interested in the influence of the stars on terrestrial events. Moreover, both have always been pursued in the belief that the processes men witness in heaven and on earth manifest the will of the Creator and, if correctly understood, will yield the key to his intentions.

Relationship of alchemy and astrology

NATURE AND SIGNIFICANCE

That both astrology and alchemy may be regarded as fundamental aspects of thought is indicated by their apparent universality. It is notable, however, that the evidence is not equally substantial in all times and places. Evidence from ancient Middle America (Aztecs, Mayans) is still almost nonexistent; evidence from India is tenuous and from ancient China, Greece, and Islāmic lands is only relatively more plentiful. A single manuscript of some 80,000 words is the principal source for the history of Greek alchemy. Chinese alchemy is largely recorded in about 100 "books" that are part of the Taoist canonical writings. Neither Indian nor Islāmic alchemy has ever been collected, and scholars are thus dependent for their knowledge of the subject on occasional allusions in works of natural philosophy and medicine, plus a few specifically alchemical works.

Nor is it really clear what alchemy was (or is). The word is a European one, derived from Arabic, but the origin of the root word, *chem*, is uncertain. Words similar to it have been found in most ancient languages, with different meanings, but conceivably somehow related to alchemy. In fact, the Greeks, Chinese, and Indians usually referred to what Westerners call alchemy as "The Art," or by terms denoting change or transmutation.

The chemistry of alchemy. Superficially, the chemistry involved in alchemy appears a hopelessly complicated succession of heatings of multiple mixtures of obscurely named materials; but it seems likely that a relative simplicity underlies this complexity. The metals gold, silver, copper, lead, iron, and tin were all known before the rise of alchemy. Mercury, the liquid metal, which seems to have been discovered about 300 BC, when it appears in both Eastern and Western sources, was crucial to alchemy, and its discovery coincided with the first appearance of the art. Sulfur, "the stone that burns," was also crucial. It was known from prehistoric times in native

deposits and was also given off in metallurgic processes (the "roasting" of sulfide ores). Mercury united with most of the other metals, and the amalgam formed coloured powders (the sulfides) when treated with sulfur. Mercury itself occurs in nature in a red sulfide, cinnabar; and cinnabar can be artificially made. All of these, except possibly the last, were operations known to the metallurgist and were adopted by the alchemist.

The alchemist added the action on metals of a number of corrosive salts, mainly the vitriols (copper and iron sulfates), alums (the aluminum sulfates of potassium and ammonium), and the chlorides of sodium and ammonium. And he made much of the property of arsenic (itself made by heating the sulfide pigments orpiment and realgar) of colouring metals. All of these materials, except the chloride of ammonia, were known in ancient times. Known as sal ammoniac in the West, *nao sha* in China, *nao sadar* in Sanskrit, and *nushādir* in Persia and Arabic lands, the chloride of ammonia first became known to the West in the *Chou i ts'an t'ang ch'i*, a Chinese treatise of AD 142. It was to be crucial to alchemy, for on sublimation (a favourite alchemical process) it dissociates into antagonistic corrosive materials, ammonia and hydrochloric acid, which readily attack the metals. Until the 9th century it seems to have come from a single source, the Fire Mountain (Huo yen shan) near Turfan, in Central Asia.

Finally, the manipulation of these materials was to lead to the discovery of the mineral acids, the history of which began in Europe in the 13th century. The first was probably nitric acid, made by distilling together saltpetre (potassium nitrate) and vitriol or alum. More difficult to discover was sulfuric acid, which was distilled from vitriol or alum alone but required apparatus resistant to corrosion and heat. And most difficult was hydrochloric acid, distilled from common salt or sal ammoniac and vitriol or alum, for the vapours of this acid cannot be simply condensed but must be dissolved in water.

Goals. "Transmutation" is the key word characterizing alchemy, and it may be understood in several ways: in the changes that are called chemical, in physiological changes such as passing from sickness to health, in a hoped-for transformation from old age to youth, or even in passing from an earthly to a supernatural existence. Alchemical changes seem always to have been positive, never involving degradation except as an intermediate stage in a process having a "happy ending." Alchemy aimed at the great human "goods": wealth, longevity, and immortality.

Wealth, longevity, and immortality

Alchemy was not original in seeking these goals, for it had been preceded by religion, medicine, and metallurgy. The first chemists were metallurgists, who were perhaps the most successful practitioners of the arts in antiquity. Their theories seem to have come not from science but from folklore and religion. The miner and metallurgist, like the agriculturalist, in this view, accelerate the normal maturation of the fruits of the earth, in a magico-religious relationship with nature. In primitive societies the metallurgist is often a member of an occult religious society.

But the first ventures into natural philosophy, the primitive beginnings of what is called the scientific explanation, also preceded alchemy. Systems of five almost identical elements were postulated in China, India, and Greece. They were supposed to form the world of things under a view in which nature comprised antagonistic, opposite forces—hot and cold, positive and negative, and male and female; *i.e.*, primitive versions of the modern conception of energy. Under their equally similar astrological heritage, philosophers found correspondences among the elements, planets, and metals. In short, both the chemical arts and the theories of the philosophers of nature had become complex before alchemy appeared.

REGIONAL VARIATIONS

Chinese alchemy. Neither in China nor in the West can scholars approach with certitude the origins of alchemy, but the evidences in China appear to be slightly older. Indeed, Chinese alchemy was connected with an enterprise older than metallurgy—*i.e.*, medicine. Belief in physical immortality among the Chinese seems to go back to the 8th century BC, and belief in the possibility of attaining it through drugs to the 4th century BC. The magical drug, namely the "elixir of life" (elixir is the European word), is mentioned about that time, and that most potent elixir, "drinkable gold" which was a solution (usually imaginary) of this corrosion-resistant metal, as early as the 1st century BC—many centuries before it is heard of in the West.

Although non-Chinese influences (especially Indian) are possible, the genesis of alchemy in China may have been a purely domestic affair. It emerged during a period of political turmoil, the "Period of the Warring States" (from the 5th to the 3rd century BC); and it came to be associated with Taoism—a mystical religion founded by the 6th-century-BC sage Lao-tzu—and its sacred book, the *Tao-te Ching* ("Classic of the Way of Power"). The Taoists were a miscellaneous collection of "outsiders"—in relation to the prevailing Confucianists—and such mystical doctrines as alchemy were soon grafted onto the Taoist canon. What is known of Chinese alchemy is mainly owing to that graft, and especially to a collection known as *Yün chi ch'i ch'ien* ("Seven Tablets in a Cloudy Satchel"), which is dated 1023. Thus, sources on alchemy in China (as elsewhere) are compilations made long after the supposed composition of the writings were collected.

The oldest known Chinese alchemical treatise is the *Chou i ts'an t'ang ch'i* ("commentary on the *I Ching*"), now thought to date from AD 142. In the main it is an apocryphal interpretation of the "Classic of Changes" (*I Ching*), an ancient classic especially esteemed by the Confucianists, relating alchemy to the mystical mathematics of the 64 hexagrams (six-line figures used for divination purposes). Its relationship to chemical practice is tenuous, but it mentions materials (including sal ammoniac) and implies chemical operations. The first Chinese alchemist who is reasonably well known was Ko Hung (AD 283–343), whose book *Pao p'u tzu* (pseudonym of Ko Hung) contains two chapters with obscure recipes for elixirs, mostly based on mercury or arsenic compounds. The most famous Chinese alchemical book is the *Tan chin yao chüeh* ("Great Secrets of Alchemy"), probably by Sun Ssu Miao (AD 581–after 673). It is a practical treatise for elixirs of the same kind (mercury, sulfur, and the salts of mercury and arsenic are prominent), for the attainment of immortality, plus a few for specific cures for disease and such other purposes as the fabrication of precious stones.

Altogether, the similarities between the materials used and the elixirs made in China, India, and the West are more remarkable than are their differences. Nonetheless, Chinese alchemy differed from that of the West in its objective. Whereas in the West the objective seems to have evolved from gold making to elixirs of immortality to simply superior medicines, neither the first nor the last of these objectives seems ever to have been very important in China. Chinese alchemy was consistent from first to last, and there was relatively little controversy among its practitioners, who seem to have varied only in their prescriptions for the elixir of immortality or perhaps only over their names for it, of which one American Sinologist has counted about 1,000. In the West there were conflicts between advocates of herbal and "chemical" (*i.e.*, mineral) pharmacy, but in China mineral remedies were always accepted. There were, in Europe, conflicts between alchemists who favoured the gold-making objective and those who thought medicine the proper goal, but the Chinese always favoured the latter. Since alchemy rarely (probably never) achieved any of these goals, it was an advantage to the Western alchemist to have the situation obscured, and the art survived in Europe long after Chinese alchemy had simply faded away.

Objective of Chinese alchemy

Chinese alchemy followed a strange path. Whereas the Western world, with its numerous religious promises of immortality, never seriously expected alchemy to fulfill that goal, the deficiencies of Chinese religions in respect to promises of immortality left that goal open to the al-

chemist. And he sought a *physical* immortality. A serious reliance on medical elixirs that were in varying degrees poisonous led the alchemist into permanent exertions to mollify those poisons, either through variation of the ingredients (although the Chinese always favoured cinnabar [mercury sulfide]) or through chemical manipulations, such as what they called "repeated cyclical transformation." The fact that immortality was so desirable and the alchemist thus correspondingly valued has enabled the British historian of science Joseph Needham to tabulate a series of Chinese emperors who probably died of elixir poisoning. Royal victims made alchemists more cautious, and apparently emperors as well, and Chinese alchemy vanished (probably as the Chinese adopted Buddhism, which offered other, less dangerous avenues to immortality), leaving its literary manifestations imbedded in the Taoist canons.

Indian alchemy. The oldest Indian writings, the Vedas (Hindu sacred scriptures), contain the same hints of alchemy that are found in evidences from ancient China, namely vague references to a connection between gold and long life. Mercury, which was so vital to alchemy everywhere, is first mentioned in the 4th- to 3rd-century-BC *Arthaśāstra*, about the same time it is encountered in China and in the West. Evidence of the idea of transmuting base metals to gold appears in 2nd- to 5th-century-AD Buddhist texts, about the same time as in the West. Since Alexander the Great had invaded India in 325 BC, leaving a Greek state (Gandhāra) that long endured, the possibility exists that the Indians received this objective from the Greeks, but it could have been the other way round.

Medical and religious concerns of Indian alchemy

It is also possible that the alchemy of medicine and immortality came to India from China, or vice versa; in any case, gold making appears to have been a minor concern of Indian alchemy, as it was of Chinese. Medicine was the chief objective of alchemy in both cultures, and the Indians were as receptive as the Chinese to mineral remedies. But the elixir of immortality was of little importance in India (which had other avenues to immortality). The Indian elixirs were for specific diseases or, at the most, for long life.

As in China and the West, alchemy in India came to be associated with religious mysticism, but much later—not until the rise of Tantrism (an esoteric, occultic, meditative system), AD 1100–1300. To Tantrism is owed writings that are clearly alchemical (such as the 12th-century *Rasārṇava*, or "Treatise on metallic preparations"), writings that show more clearly what the Indian alchemist thought and practiced.

From the earliest records of Indian natural philosophy, 5th–3rd centuries BC, theories of nature were based on conceptions of material elements (fire, wind, water, earth, and space), vitalism ("animated atoms"), and dualisms of love and hate or action and reaction. But theory is a very dim background to Indian alchemy. The alchemist coloured metals and on occasion "made" gold, but he gave little importance to that. His six metals (gold, silver, tin, iron, lead, and copper), each further subdivided (five kinds of gold, etc.), were "killed" (*i.e.*, corroded) but not "resurrected," as was the custom of Western alchemy. Rather, they were killed to make medicines. Although "the secrets of mercurial lore" became part of the Tantric rite, mercury seems to have been much less important than in China. The Indians exploited metal reactions more widely, but, although they possessed from an early date not only vitriol and sal ammoniac but also saltpetre, they nevertheless failed to discover the mineral acids. This is the more remarkable because India was long the principal source of saltpetre, which occurs as an efflorescence on the soil, especially in populous tropical countries. But it lacks the high degree of corrosivity of metals possessed by the vitriols and chlorides and played a small part in early alchemy. Saltpetre appears particularly in 9th- to 11th-century-AD Indian and Chinese recipes for fireworks, one of which—a mixture of saltpetre, sulfur, and charcoal—is gunpowder. Saltpetre first appears in Europe in the 13th century, along with the modern formula for gunpowder and the recipe for nitric acid.

Hellenistic alchemy. Western alchemy may go back to the beginnings of the Hellenistic period (*c.* 300 BC–*c.* AD 300), although the earliest alchemist whom authorities have regarded as certainly authentic is Zosimos of Panopolis (Egypt), who lived near the end of the period. He is one of about 40 authors represented in a compendium of alchemical writings that was probably put together in Byzantium (Constantinople) in the 7th or 8th century AD and that exists in manuscripts in Venice and Paris. Synesius, the latest author represented, lived in Byzantium in the 4th century. The earliest is the author designated Democritus but identified by scholars with Bolos of Mende, a Hellenized Egyptian who lived in the Nile Delta about 200 BC. He is represented by a treatise called *Physica et mystica* ("Natural and Mystical Things"), a kind of recipe book for dyeing and colouring but particularly for the making of gold and silver. The recipes are stated obscurely and are justified with references to the Greek theory of elements and to astrological theory. Most end with the phrase "One nature rejoices in another nature; one nature triumphs over another nature; one nature masters another nature," a phrase that authorities variously trace to the Magi (Zoroastrian priests), Stoic pantheism (a Greek philosophy concerned with nature), or to the 4th-century-BC Greek philosopher Aristotle. It was the first of a number of such aphorisms over which alchemists were to speculate for many centuries.

The importance of the *Physica et mystica*

In 1828 a group of ancient papyrus manuscripts written in Greek was purchased in Thebes (Egypt), and about a half-century later it was noticed that among them, divided between libraries in Leyden (The Netherlands) and Stockholm, was a tract very like the *Physica et mystica*. It differed from the *Physica et mystica*, however, in that it lacked the theoretical embellishments and stated in some recipes that only fraudulent imitation of gold and silver was intended. Scholars believe that this kind of work was the ancestor both of the *Physica et mystica* and of the ordinary artist's recipe book. The techniques were ancient. Archaeology has revealed metal objects inlaid with colours obtained by grinding metals with sulfur, and Homer's description (8th century BC) of the shield of Achilles gives the impression that the artist in his time was virtually able to paint in metal.

Democritus is praised by most of the other authors in the Venice–Paris manuscript, and he is much commented upon. But only Zosimos shows what had become of alchemy after Bolos of Mende. His theory is luxurious in imagery, beginning with a discussion of "the composition of waters, movement, growth, embodying and disembodying, drawing the spirits from bodies and binding the spirits within bodies" and continuing in the same vein. The "base" metals are to be "ennobled" (to gold) by killing and resurrecting them, but his practice is full of distillation and sublimation, and he is obsessed with "spirits." Theory and practice are joined in the concept that success depends upon the production of a series of colours, usually black, white, yellow, and purple, and that the colours are to be obtained through *Theion hydōr* (divine or sulfur water—it could mean either).

Zosimos credits these innovations mainly to Maria (sometimes called "the Jewess"), who invented the apparatus, and to Agathodaimon, probably a pseudonym. Neither is represented (beyond Zosimos' references) in the Venice–Paris manuscript, but a tract attributed to Agathodaimon published in 1953 shows him to be preoccupied with the colour sequence and complicating it by using arsenic instead of sulphur. Thus, the colour-producing potentialities of chemistry were considerable by the time of Zosimos.

Zosimos also shows that alchemical theory came to focus on the idea that there exists a substance that can bring about the desired transformation instantly, magically, or, as a modern chemist might say, catalytically. He called it "the tincture," and had several. It was also sometimes called "the powder" (*xērion*), which was to pass through Arabic into Latin as elixir and finally (signifying its inorganic nature) as the "philosopher's stone," "a stone which is not a stone," as the alchemists were wont to say. It was sometimes called a medicine for the

rectification of "base" or "sick" metals, and from this it was a short step to view it as a drug for the rectification of human maladies. Zosimos notes the possibility, in passing. Another step and the objective of alchemy became human salvation, and the material constitution of the elixir became less important than the incantations that accompanied its production. Synesius, the last author in the Venice–Paris manuscript, already defined alchemy as a mental operation, independent of the science of matter.

Thus, Greek alchemy came to resemble, in both theory and practice, that of China and India. But in its objective it adhered to gold making and thus remained fundamentally different.

Arabic alchemy. Arabic alchemy is as mysterious as Greek in its origins, and the two seem to have been significantly different. The respect in which *Physica et mystica* was held by the Greek alchemists was bestowed by the Arabs on a different work, the *Emerald Tablet* of Hermes Trismegistos, the reputed Hellenistic author of various alchemical, occultic, and theological works. Beginning "That which is above is like to that which is below, and that which is below is like to that which is above," it is brief, theoretical, and astrological. Hermes "the thrice great" (Trismegistos) was a Greek version of the Egyptian god Thoth and the supposed founder of an astrological philosophy that is first noted in 150 BC. The *Emerald Tablet*, however, comes from a larger work called *Book of the Secret of Creation*, which exists in Latin and Arabic manuscripts and is attributed to Apollonius of Tyana (called Balīnūs by the Arabs), an authentic pagan mystic of the first Christian century.

The significance of Hermetic literature

Some scholars have suggested that Arabic alchemy descended from a western Asiatic school and that Greek alchemy was derived from an Egyptian school. As far as is known, the Asiatic school was not Chinese or Indian. What is known is that Arabic alchemy was associated with a specific city in Syria, Ḥarrān, which in Arabic times seems to have been a fountainhead of alchemical notions. And it is possible that the distillation ideology and its spokeswoman, Maria—as well as Agathodaimon—represented the alchemy of Ḥarrān, which presumably migrated to Alexandria and was incorporated into the alchemy of Zosimos.

The existing versions of the *Book of the Secret of Creation* have been carried back only to the 7th or 6th century but are believed to represent much earlier writings, although not necessarily those of Appollonius himself. He is the subject of an ancient biography that says nothing about alchemy, but neither does the *Emerald Tablet* nor the *Book of the Secret of Creation*, of which it is a part. On the other hand, their theories of nature have an alchemical ring, and the *Book* mentions the characteristic materials of alchemy, including, for the first time in the West, sal ammoniac. It was clearly an important book to the Arabs, most of whose eminent philosophers mentioned alchemy, although some of them disapproved of it. Those who practiced it were even more interested in literal gold making than had been the Greeks. The most well-attested and probably the greatest Arabic alchemist was ar-Rāzī (*c.* 850–923 or 924), a Persian physician who lived in Baghdad. The most famous was Jābir ibn Ḥayyān, but this is now believed to be a name applied to a collection of "underground writings" produced in Baghdad after the theological reaction against science. In any case, the Jābirian writings are very similar to those of ar-Rāzī.

Ar-Rāzī classified the materials used by the alchemist into "bodies" (the metals), stones, vitriols, boraxes, salts, and "spirits," putting into the latter those vital (and sublimable) materials, mercury, sulfur, orpiment and realgar (the arsenic sulfides), and sal ammoniac. Much is made of sal ammoniac, the reactive powers of which seem to have given Western alchemy a new lease on life. Ar-Rāzī and the Jābirian writers were really trying to make gold, through the catalytic action of the elixir, and they concentrated their effort on its production. Both wrote much on the compounding of "strong waters," an enterprise that was ultimately to lead to the discovery of the mineral acids; but students have been no more able to find this discovery in the writings of the Arabic alchemists than in those of China and India. The Arabic strong waters were merely corrosive salt solutions.

Ar-Rāzī's writing represents the apogee of Arabic alchemy, so much so that students of alchemy have little evidence of its later reorientation toward mystical or quasi-religious objectives. Nor does it seem to have turned to medicine, which remained independent. But there was a tendency in Arabic medicine to give greater emphasis to mineral remedies and less to the herbs that had been the chief medicines of the earlier Greek and Arabic physicians. The result was a pharmacopoeia not of elixirs but of specific remedies that are inorganic in origin and not very different from the elixirs of ar-Rāzī. This new pharmacopoeia was taken to Europe by Constantine of Africa, a Baghdad-educated Muslim who died in 1087 as a Christian monk at Monte Cassino (Italy). The pharmacopoeia also appeared in Spain in the 11th century and passed from there to Latin Europe, along with the Arabic alchemical writings, which were translated into Latin in the 12th century.

Latin alchemy. In the 12th century the Christian West began to shed its habit of indifference or hostility to the secular literature of the ancient and alien worlds. Christian scholars were particularly attracted to Muslim Spain and Sicily and there pursued the arduous task of translation from both Arabic and Greek works. Much of what they translated was in some degree familiar, but some, including the literature of alchemy, was new.

The Greek alchemy of the Venice–Paris manuscript had much less impact than the work of ar-Rāzī and other Arabs, which emerged among the voluminous translations made in Spain about 1150 by Gerard of Cremona. By 1250 alchemy was familiar enough to enable such encyclopaedists as Vincent of Beauvais to discuss it fairly intelligibly, and before 1300 the subject was under discussion by the English philosopher and scientist Roger Bacon and the German philosopher, scientist, and theologian Albertus Magnus. To learn about alchemy was to learn about chemistry, for Europe had no independent word to describe the science of matter. It had been touched upon in works concerned with other forms of change—*e.g.*, the motion of projectiles, the aging of man, and similar concepts found in Aristotelian commentaries. On the practical side there were also artist's recipe books; but for the first time in the works of Roger and Albertus change was discussed in a truly chemical sense, using evidences from the newly translated alchemy. Roger wrote as though alchemy were a general science of matter, and he had great hopes for its utility.

But the more familiar alchemy became, the clearer it was understood that gold making was not merely the main but was almost the exclusive objective of alchemy, and Europeans proved no more resistant to this objective than had their Arabic predecessors. By 1350, alchemical tracts were pouring out of the scriptoria (monastic copying rooms), and the Europeans had even taken over the tradition of anonymity and false attribution. One authority wrote at length about supposed disagreements between two Arabs, Iahiae Abindinon and Geber Abinhaen, who were probably two versions of the name of Jābir ibn Ḥayyān. The most famous Jābirian work in Europe, *The Sum of Perfection*, is now thought to have been an original European composition.

The objective of European alchemy

By 1300 alchemists had begun the discovery of the mineral acids, a discovery that occupied about three centuries between the first evidences of the new strong water (aqua fortis—*i.e.*, nitric acid) and the clear differentiation of the acids into three kinds: nitric, hydrochloric, and sulfuric. These three centuries saw prodigous efforts in European alchemy, for these spontaneously reactive and highly corrosive substances opened a whole new world of research. And yet, it was of little profit to chemistry, for the experiments were controlled by the old objectives of separating the base metals into their "elements," concocting elixirs, and other traditional procedures.

The "water of life" (aqua vitae; *i.e.*, alcohol) was probably discovered a little earlier than nitric acid (suggest-

ing that the condensing apparatus was crucial), and some physicians and a few alchemists turned to the elixir of life as an objective. John of Rupescissa, a Catalonian monk who wrote about 1350, prescribed virtually the same elixirs for metal ennoblement and for the preservation of health. His successors multiplied elixirs, which lost their uniqueness and finally simply became new medicines, often for specific ailments. Medical chemistry may have been conceived in Islām, but it was born in Europe. It only awaited christening by its great publicist, Paracelsus (1493 or 1494–1541), who was the sworn enemy of the malpractices of 16th-century medicine, the vigorous advocate of "folk" and "chemical" remedies, and a vociferous advertiser of himself. By the end of the 16th century, medicine was divided into warring camps of Paracelsians and anti-Paracelsians, and the alchemists began to move en masse into pharmacy.

Paracelsian pharmacy was to lead, by a devious path, to modern chemistry, but gold making still persisted although overshadowed by this development. In the 15th century it entered its most picturesque period, for then the personal reminiscences of alchemists began to appear. Most famous was the Paris notary Nicolas Flamel (1330–1418), who dreamed of an occult book, subsequently found it, and succeeded in deciphering it with the aid of a Jewish scholar learned in the mystic Hebrew writings known as the Kabbala. In 1382 Flamel succeeded in the "Great Work" (gold making), evidence of which is the fact that he became rich and made donations to churches.

Alchemy and kabbalistic speculation

Few alchemists are so well attested. Salomen Trismosin, purported author of the *Splendor solis*, or "Splendour of the Sun" (published 1598), engaged in extensive visits to alchemical adepts (a common practice) and found success through "kabbalistic and magical books in the Egyptian language." The impression given is that many had the secret of gold making but that most of them received it from someone else and not from personal experimentation. Illustrations became particularly important. The illustrations of *Splendor solis* are as bizarre as the text, but they clearly had more appeal, even to modern students.



Saturn, or *Mercurius senex*, being cooked in the bath until the spirit or white dove *(pneuma)* ascends. From Salomen Trismosin's *Splendor solis*, 1598. In the British Museum (MS. Harley 3469).
By courtesy of the trustees of the British Museum

MODERN ALCHEMY

The possibility of chemical gold making was not, until the 19th century, conclusively contradicted by scientific evidence. As rational a scientist as Isaac Newton (1642–1727) had thought it worthwhile to experiment with it. The official attitude toward alchemy in the 17th and 18th centuries was ambivalent. On the one hand, The Art posed a threat to the control of precious metal and was often outlawed; on the other hand, there were obvious advantages to any sovereign who could control gold making by controlling the gold maker. The latter view generally prevailed. Queen Elizabeth I of England (1533–1603) is said to have given John Dee a license to perform alchemy. In "the metropolis of alchemy," Prague, the emperors Maximilian II (1564–76) and Rudolf II (1576–1612) of the Holy Roman Empire proved ever-hopeful sponsors and entertained most of the leading alchemists of Europe.

This was not altogether to the alchemist's advantage. The elector of Saxony, Christian II, imprisoned and tortured the Scotsman Alexander Sethon, who had been travelling about Europe performing well-publicized transmutations, and Edward Kelly, an English alchemist, lost his life in an attempt to escape from Rudolf II's prison. The situation was complicated by the fact that some alchemists were turning from gold making, not to medicine but to a quasi-religious alchemy reminiscent of the Greek Synesius. Rudolf II made the German alchemist Michael Maier a count and his private secretary, although Maier's mystical and allegorical writings were, in the words of a modern authority, "distinguished for the extraordinary obscurity of his style" and made no claim to gold making. Neither did the German alchemist Heinrich Khunrath (*c.* 1560–1601), whose works have long been esteemed for their illustrations, make such a claim.

Conventional gold making was not dead and probably never will be. Adolf Hitler is said to have kept gold makers, on the chance of improving the treasury of Nazi Germany. But in the 18th century the art turned conclusively to religious aims. The appearance of modern chemistry gave rise to general skepticism as to the possibility of gold making, but it also created a generally widespread dissatisfaction with the objectives of modern science, which were viewed as too limited. Unlike the scientists of the Middle Ages and Renaissance, the descendants of Newton and Antoine Lavoisier (the great 18th-century French chemist), limited their objectives in a way that amounted to a renunciation of what many had supposed to be the most important question of science, the relation of man to the cosmos. Those who persisted in asking these questions came to feel an affinity with the alchemists. Thus arose a new "esoteric" alchemy, distinguished from the "exoteric" alchemy of the gold makers and seeking roots in Synesius and other late Greek alchemists of the Venice–Paris manuscript.

This revival of alchemy, or Hermeticism, as its practitioners often prefer to call it, crystallized around the Rosicrucian brotherhood, which appeared in Germany in the early 17th century. It attracted the favourable attention not only of such reforming alchemists as Michael Maier, but of many prominent philosophers who were disquieted by the mechanistic character of the new science. In modern times alchemy has become a focal point for various kinds of mysticism and not merely formally. The old alchemical literature continues to be scrutinized for evidence, for alchemy, modern as well as ancient, professes to deal with doctrine that is not new but that has on more than one occasion come into the possession of man but has always again been lost. Nor is its association with chemistry considered accidental. In the words of a famous 19th-century alchemist, Mary Anne Atwood,

Association of alchemy with chemistry

> Alchemy is an universal art of vital chemistry which by fermenting the human spirit purifies and finally dissolves it. . . . Alchemy *is* philosophy; it is *the* philosophy, the finding of the Sophia in the mind."

ASSESSMENTS OF ALCHEMY

Accomplishments. The most consistent goals of alchemy have been the prolongation of life and the trans-

mutation of the base metals into gold. It would appear that neither was accomplished, unless one would choose to credit to alchemy the consequences of modern chemotherapy and of the cyclotron. These, however, are generally credited to modern science, the essence of which is the rejection of ideologies typified by alchemy.

Can alchemy be credited with the development of the science of chemistry, a keystone of modern science? During the alchemical period the repertoire of known substances was enlarged (*e.g.*, by the addition of sal ammoniac and saltpetre), alcohol and the mineral acids were discovered, and the basis was laid from which modern chemistry was to rise. Historians of chemistry, whose science lacks the respectable antiquity of astronomy, physics, and biology, have been tempted to credit alchemy with laying this base while at the same time paradoxically regarding alchemy as mostly "wrong." But it is far from clear that the basis of chemistry was in fact laid by alchemy rather than medicine. During the crucial period of Arabic and early Latin alchemy, the weight of evidence tends toward the conclusion that innovation owed more to the nascent medical chemistry than it did to alchemy.

But those who seek the history of the science of matter as understood more widely than the modern chemist chooses to admit may find alchemy more rewarding. Numerous Hermetic writers of the last century offer an assurance that the misunderstood goal of the art can yet be found—indeed, that it has been repeatedly found and repeatedly lost. This is a matter of judgment, but it can certainly be said that the modern chemist has not found the goal sought by the alchemist. For those who seek the goal of scientific chemistry, alchemy can have no further interest. For those who seek the wider goal, which was also that of the natural philosopher before the advent of "mechanical," "Newtonian," or "modern" science, the search is still on.

Interpretations. Charlatanism was a prominent feature of European alchemy during the 16th century, and such monarchs as Rudolf II—even if they had mainly themselves to blame—were not entirely without reason in incarcerating some of their resident adepts. This picturesque era and its coincidence with the period when the modern science of chemistry appeared in embryonic form has led many historians of chemistry to view alchemy in general as a fraud.

Other historians of chemistry have attempted to differentiate the good from the bad in alchemy, citing as good the discovery of new substances and new processes and the invention of apparatus. Some of this was certainly accomplished by alchemists (*e.g.*, Maria), but most of it is more probably credited to the early pharmacist.

Modern Hermetic and psychological views

And the question remains as to how much alchemy had to do with what is now called the science of chemistry. Scholars generally agree that alchemy had something to do with chemistry, but the modern Hermetic holds that chemistry was the handmaiden of alchemy, not the reverse. From this point of view the development of modern chemistry involved the abandonment of the true goal of the art, and it is still to be sought by further perusal of the ancient literature.

And, finally, a new interpretation was brought forth in the 1920s by the Swiss psychoanalyst Carl Jung, who judged alchemical literature to be explicable in psychological terms. Noticing the similarities between alchemical literature, particularly in its reliance on bizarre symbolic illustrations, and the dreams and fantasies of his patients, Jung viewed them as manifestations of a "collective unconscious" (inherited disposition). His theory remains largely undeveloped and can at best be described as the latest word, though certainly not the last. It remains a challenge rather than an explanation.

BIBLIOGRAPHY. ROBERT P. MULTHAUF, *The Origins of Chemistry* (1967), a discussion of Greek, Arabic, and European alchemy, with an extensive bibliography of the sources; E.J. HOLMYARD, *Alchemy* (1957), a general history; F.S. TAYLOR, *The Alchemists* (1949, reprinted 1962), a general history, by a leading authority on Greek alchemy; MIRCEA ELIADE, *Forgerons et alchimistes* (1956; Eng. trans., *The Forge and the Crucible*, 1962), a comparison of alchemy with the practices of the primitive metallurgist; JOSEPH NEEDHAM, *Science and Civilisation in China* (1954–), a multivolume history of science in China, briefly dealing with alchemy; NATHAN SIVIN, *Chinese Alchemy: Preliminary Studies* (1968); P.C. RAY, *A History of Hindu Chemistry*, 2 vol. (1902–09), a valuable study not yet replaced; F.S. TAYLOR, "A Survey of Greek Alchemy," *Journal of Hellenic Studies*, 50:109–139 (1930), an excellent English language summary of the subject; JACK LINDSAY, *The Origins of Alchemy in Graeco-Roman Egypt* (1970), notable for its detailed consideration of the mysticism of the Hellenistic period; A.J. HOPKINS, *Alchemy: Child of Greek Philosophy* (1934), a pioneer attempt at the understanding of the chemistry of the Greek alchemical writings; M.P. BERTHELOT and C.E. ROUELLE, *Collection des anciens alchimistes grecs*, 3 vol. (1887–88, reprinted 1963), although its accuracy has often been impugned, it remains the principal source of all studies of alchemy in Hellenistic Greece; H.E. STAPLETON, "The Antiquity of Alchemy," *Ambix*, 5:1–43 (1953), a discussion of the relationship between Greek and Arabic alchemy; GEBER, *Works*, Eng. trans. by RICHARD RUSSELL (1678, new ed. with an introduction by E.J. HOLMYARD, 1928), the most famous and important medieval Latin work on alchemy; A.E. WAITE, *Alchemists Through the Ages* (1888, reprinted 1970), a collection of biographies of alchemists, originally published by a reputable scholar who was also a "modern" alchemist; C.G. JUNG, *Psychologie und Alchemie* (1944; Eng. trans., *Psychology and Alchemy*, 2nd rev. ed., 1968); *Ambix* (quarterly), the journal for the study of alchemy and early chemistry, contains a major portion of the important new research in the history of alchemy.

(R.P.M.)

Alcibiades

Well-born and wealthy, a brilliant but unscrupulous Athenian politician, Alcibiades provoked the sharp political antagonisms that were the main causes of Athens' defeat in the Peloponnesian War (431–404 BC). Born in about 450 BC, he was only a small boy when his father was killed in command of the Athenian army, in 447 or 446 BC, at Coronea, Boeotia, and his guardian, the statesman Pericles, a distant relation, was too preoccupied with political leadership to provide the guidance and affection he needed. As he grew up, Alcibiades was strikingly handsome and keen witted, but he was extravagant, irresponsible, and self-centred as well. He was, however, impressed by the moral strength and keen mind of the philosopher Socrates, who, in turn, was strongly attracted by Alcibiades' beauty and intellectual promise. They served together at Potidaea (432) in the Chalcidice region, where Alcibiades was defended by Socrates when he was wounded, a debt repaid when he stayed to protect Socrates in the flight from the Battle of Delium (424), north of Athens. Yet, before he was 30 he had abandoned the intellectual integrity that Socrates demanded, in favour of the rewards of the kind of politics that Socrates despised.

During the 420s he was best known for his personal extravagance and courage in battle, but he had become a recognized speaker in the Assembly (Ecclesia) and, as Athens moved toward peace, he hoped that the ties that had once existed between his family and Sparta would enable him to secure the credit for bringing peace to Athens. According to the historian Thucydides, who knew Alcibiades well and judged him dispassionately, it was the fact that the Spartans chose to negotiate through established political leaders that dictated Alcibiades' choice of policies.

General for the first time in 420, he opposed the aristocratic leader Nicias, who had negotiated peace, and steered Athens into an anti-Spartan alliance with Argos, Elis, and Mantineia, three city-states of the Peloponnese. This alliance was defeated by Sparta at the Battle of Mantineia (418). Alcibiades, however, escaped ostracism, a form of banishment, by joining forces with Nicias against Hyperbolus, the successor of the demagogue politician Cleon as champion of the common people. In 416 he restored his reputation by entering seven chariots at Olympia and taking first, second, and fourth places. This made it easier for him, in 415, to persuade the people to send a major expedition to Sicily against Syracuse. He was appointed to share the command, but shortly be-

Battle of Mantineia



Alcibiades, Roman copy of a Greek sculpture of about the second quarter of the 4th century BC. In the Museo Chiaramonti, Rome.
By courtesy of the Direzione Generale dei Monumenti, Musei e Gallerie Pontificie

fore the expedition was due to sail the Hermae (busts of Hermes, messenger of Zeus and patron of all who use the roads, set up in public places throughout the city) were found to have been mutilated. In the ensuing panic Alcibiades was accused of being the originator of the sacrilege as well as of having profaned the Eleusinian mysteries. He demanded an immediate inquiry, but his enemies, led by Androcles (the successor of Hyperbolus), ensured that he sailed with the charge still hanging over him. Shortly after reaching Sicily, he was recalled; but on the journey home he escaped and, learning that he had been condemned to death in absence, went to Sparta. There he advised the Spartans to send a general to help the Syracusans and to fortify Decelea in Attica, two serious blows to Athens. He also confirmed his reputation with women (which the rich Athenian whom he had married appreciated only too well) by seducing the wife of the Spartan king Agis II, who was at Decelea with his army.

Accusation of sacrilege and defection to Sparta

In 412 Alcibiades helped to stir up revolt among Athenian allies in Ionia, on the west coast of Asia Minor; but Sparta now turned against him, and he moved to Sardis to exercise his charm on the Persian governor. When some Athenian officers in the fleet began to plan an oligarchic coup, he held out hopes that if the democracy was overthrown he could secure financial support from Persia. In this he failed and, discarded by the oligarchs who had seized power, he was recalled by the Athenian fleet, which remained loyal to the democracy and needed his abilities. From 411 to 408 he helped Athens to a spectacular recovery, defeating the Spartan fleet in the Hellespont at Abydos (411) and Cyzicus (410) and regaining control over the vital grain route from the Black Sea. These successes encouraged him to return in 407 to Athens, where he was welcomed with enthusiasm and given supreme control of the conduct of the war. In a typically bold gesture he led the procession to the Eleusinian festival by road in spite of the danger from the Spartan force at Decelea, but, in the same year, after a minor naval defeat in his absence, his political enemies persuaded the people to reject him, and he retired to a castle in Thrace. He remained, however, a disturbing influence on Athenian politics and destroyed any hopes of a political consensus. When the Athenians at Aegospotami (405) facing the Spartans in the Hellespont grew increasingly careless, he warned them of their danger. But he was ignored, and, when the Athenians lost their whole fleet in a surprise attack by the Spartan admiral Lysander, Alcibiades was no longer safe in his Thracian castle. He took refuge in Phrygia in northwestern Asia Minor with the Persian governor, who in 404 was induced by the Spartans to have him murdered.

Defeat of the Spartan fleet

Alcibiades was a colourful man, about whom many stories were remembered and many more invented. He was perhaps the most gifted Athenian of his generation; but he lacked self-discipline, and his restless ambition was disastrous to Athens. The radical leaders, Cleon and his successors, carried on a bitter feud with him, which at the critical period undermined Athenian confidence. Alcibiades could not practice his master's virtues, and his example strengthened the charge brought against Socrates in 399 of corrupting the youth of Athens.

BIBLIOGRAPHY. Plutarch's life of Alcibiades is a very readable combination of fact and fiction. J. HATZFELD, *Alcibiade* (1940), is a good critical biography (in French).

(Ru.M.)

Alcohol Consumption

People do not usually drink pure alcohol but a beverage containing alcohol, specifically ethyl alcohol. This article deals with most aspects of such drinking, from the chemical and the physiological to the medical and social.

Alcoholic beverages include wines, beers, and spirits. Wines are fermented from the sugars in fruits or berries (most commonly grapes), from various plants or their saps, from honey, and even from milk. Beers are fermented from grains after the starch in them is first converted to sugar. Spirits are distilled from wines or beers (see BREWING; DISTILLED LIQUOR; and WINE MAKING). The alcohol in all these beverages is ethyl alcohol, or ethanol (CH_3CH_2OH). Extremely small amounts of other alcohols, such as amyl, butyl, propyl, and methyl alcohol, also occur in alcoholic beverages, along with other so-called congeners that include acids, aldehydes, esters, ketones, phenols, and tannins; there are also numerous inorganic substances, including vitamins and minerals. Some of these ingredients are derived from primary plant materials; some are produced during the fermentation process and may be reduced by purification; and some are introduced during aging—for example, by contact with containers such as wooden barrels. The combinations and exact amounts of congeners vary with the type of beverage, ranging from 33 milligrams per litre in vodka, which is purified diluted alcohol, to averages of 500 milligrams per litre in some whiskies and 2,600 in specially aged whiskies or brandies. Congeners contribute special characteristics of taste, aroma, and colour to the beverages. Some have nutrient and medicinal effects. Some, in spite of their small quantity, slow the rate at which the body disposes of ethyl alcohol and may have toxic effects if very large amounts of alcoholic beverages are consumed. But the main ingredient that characterizes alcoholic beverages and the chief contributor of the effects sought by people who drink them is ethyl alcohol (hereafter, simply alcohol).

Alcohol strengths

In beers the alcohol content varies from about 2 percent in some mild Scandinavian varieties to about 8 percent in especially strong types; most U.S. beers contain between 4 and 5 percent. Natural or unfortified wines (the so-called dry wines, such as burgundy, chianti, and sauterne) usually contain between 8 and 12 percent alcohol, although most U.S. varieties contain from 12 to 14 percent. Vermouths and aperitif wines usually contain 18 percent, and dessert, sweet, and cocktail wines (such as sherry, port, and muscatel) contain 20 to 21 percent. These percentages are by volume; *i.e.*, the proportion of alcohol in the fluid volume of an average American beer is 4.5 percent. Since fermentation yields only 14 percent alcohol, the extra strength of fortified wines comes from the addition of alcohol or brandy. Spirits, including vodka, gin, and whiskies (rye, Scotch, bourbon), rum (distilled from sugarcane or molasses), brandies (distilled from fruit wines), and liqueurs (flavoured syrupy spirits) usually contain between 40 and 50 percent alcohol. Cordials, made of flavoured spirits, such as anisette, blackberry, curaçao, maraschino, and sloe gin, usually contain between 25 and 40 percent. The Asiatic kumiss (made from mare's milk) and the Russian kvass rareiy contain more than 2 percent alcohol.

The article is divided into the following sections:

I. Physiological and psychological effects of alcohol
What the body does with alcohol
Intoxication
Effects on the brain
Effects on emotional behaviour
Direct effects on organs
The acute diseases
The chronic diseases
II. Social conditions of alcohol consumption
History of the use of alcohol
In early societies
Among classical peoples
Alcohol in complex societies
Drinking patterns
Kinds and amounts
Kinds and customs
III. Alcoholism
Definitions and causal factors of alcoholism
Prevalence of alcoholism
Treatment of alcoholism
Physiological therapies
Other therapies
IV. Contemporary alcohol problems and controls
Individual and social effects
Alcohol control
The new scientific orientation

I. Physiological and psychological effects of alcohol

WHAT THE BODY DOES WITH ALCOHOL

Absorption and distribution

When an alcoholic beverage is swallowed, it is diluted by the stomach juices—it may first undergo some dilution in the mouth—and very quickly distributed throughout the body. Alcohol does not require digestion before its absorption into the bloodstream. A small portion is diffused into the bloodstream directly from the stomach wall. Most of it passes through the pyloric junction into the small intestine, where it is very rapidly absorbed and circulated. The rate of absorption will vary according to the beverage and the state of the stomach. A strong alcohol solution, when taken into an empty stomach, may cause a spasm of the pylorus that will impede its passage into the small intestine, resulting in a slower rate of absorption. The presence of food in the stomach, especially some fatty foods, will delay absorption. On the other hand, alcohol taken together with carbonated beverages will ordinarily be absorbed more rapidly. Even the emotional state of the drinker may affect the rate of absorption. Probably cocktails and highballs provide the most rapidly absorbed forms of beverage alcohol.

Alcohol is diffused in the body in proportion to the water content of the various tissues and organs, appearing in greater concentration in the blood and brain than in fat or muscle tissue. The absorbed alcohol is greatly diluted by the body fluids. Thus, the alcohol in an ounce (29.57 millilitres) of whiskey at 50 percent strength (100 U.S. proof, 87.6 British proof) will be diluted, in a man of average build, to a concentration of about 2 parts per 10,000 in the blood (0.02 percent). In a larger person the blood alcohol concentration will be proportionately less; and, in a smaller one, more.

The importance of the rate of absorption, distribution, and dilution of alcohol in the body lies in their relation to the achieved alcohol concentration in various organs, particularly the brain. After rapid intake of a large amount of alcohol that is quickly absorbed, a relatively early "high" of intoxication will occur. Figure 1 shows a set of probable average curves of blood alcohol concentrations in an average-size man after rapid absorption of various amounts of alcohol. It shows also the average rate of decline of the blood alcohol concentration over time as a result of the disposal of the alcohol by the body through metabolism (oxidation) and excretion.

The body begins to dispose of alcohol immediately after it is absorbed. An insignificantly small proportion is exhaled through the lungs, and a tiny amount is excreted in sweat. A small proportion is secreted by the kidneys and will be accumulated and retained in the bladder until eliminated in the urine. Only between 2 and 10 percent of the alcohol is eliminated by these means. The remainder, 90 percent or more of the absorbed alcohol, is disposed of by metabolic processes, mainly in the liver. Other organs are capable of metabolizing only small amounts of alcohol.

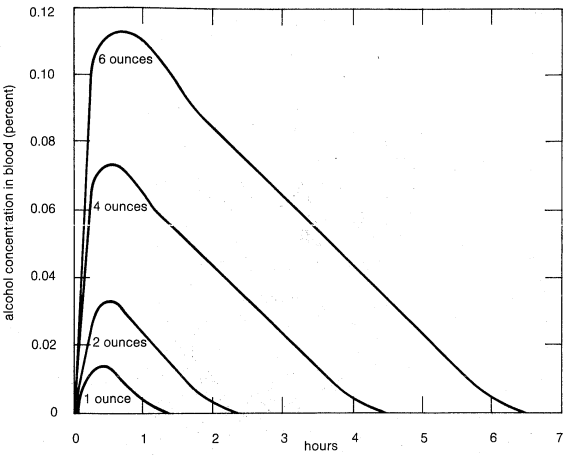


Figure 1: Percent of blood alcohol concentration in an average man at hourly intervals after drinking one, two, four, or six ounces of spirits.

How alcohol is metabolized

As absorbed alcohol is passed through the liver by the circulating blood, it is acted upon by ADH (alcohol dehydrogenase), a zinc-containing enzyme found chiefly in the liver cells. The alcohol molecule is converted by this action to acetaldehyde, itself a highly toxic substance, but this is immediately acted upon by another enzyme, aldehyde dehydrogenase, and converted to acetate, most of which enters the bloodstream and is ultimately oxidized to carbon dioxide and water. Considerable utilizable energy—7.1 calories from each gram of alcohol (about 200 calories per ounce)—is made available to the organism during these processes, and in this sense alcohol serves as a nutrient.

The two enzymatic reactions—that of ADH and aldehyde dehydrogenase—require a co-enzyme, NAD (nicotinamide-adenine dinucleotide), the acceptor of hydrogen from the alcohol molecule, for their effects. The NAD is thus changed to NADH and becomes available again for the same reaction only after its own further oxidation. While adequate ADH seems always present for the first step of alcohol metabolism, the temporary reduction of the available NAD apparently acts as a limit on the rate at which alcohol can be metabolized. That rate is about half an ounce of alcohol (equal to about an ounce of spirits, a 12-ounce bottle of beer, a four-ounce glass of dry wine, or a 2.5-ounce glass of fortified wine) per hour in an average-size man.

The limit on the rate at which the body can dispose of alcohol results in its accumulation in the organism if drinking proceeds at a faster rate than the alcohol is metabolized. The average-size man who drinks and absorbs four ounces of whiskey (at 50 percent alcohol) rapidly will have 1½ ounces of alcohol in his body at the end of an hour and a blood alcohol concentration near to 0.07 percent (Figure 1). If he drinks eight ounces within two hours, he will have a blood alcohol concentration of about 0.11 percent. The continued accumulation of alcohol in the organism faster than it can be metabolized leads to increasing degrees of intoxication. This is illustrated in Figure 2, Curve A, which shows what happens if an average man drinks two ounces of spirits each hour for four hours. At the end of the first hour, the blood alcohol concentration has passed its peak and begun to decline, but with the next drink it starts rising again; this is repeated after each drink. Only at four hours—an hour after the last drink—is the highest blood alcohol concentration reached; and, with no more drinking, it declines steadily thereafter. Curve A thus illustrates the combined effects of repeated alcohol absorption and its continuous metabolism. Curve B in Figure 2 shows what the course of blood alcohol concentration would be if the eight ounces of spirits were drunk and absorbed all at once.

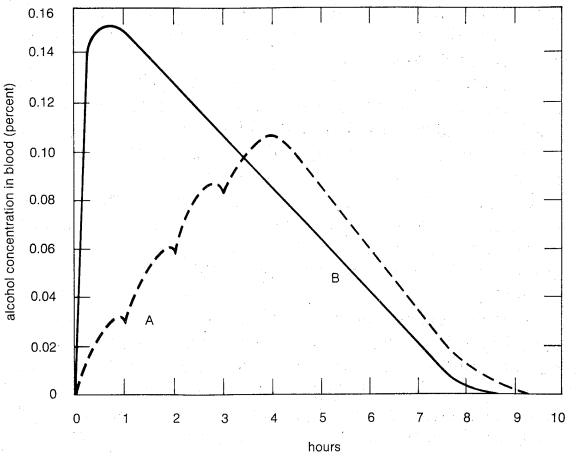


Figure 2: Percent of blood alcohol concentration in an average man at hourly intervals after drinking two ounces of spirits each hour for four hours (curve A) and eight ounces all at once (curve B).

INTOXICATION

General effects of intoxication

The drinking of small amounts of alcohol, even regularly over a long period of time, does not have any conclusively demonstrated pathological effect. A mild infrequent intoxication produces a variety of temporary biochemical disturbances: the adrenal glands may discharge hormones, sugar may be mobilized from stores in the liver, the electrolyte balance may be slightly upset, the metabolism and equilibrium of the liver may be disturbed. But these changes leave no chronic aftereffects, and the body rapidly returns to normal. A severe intoxication may produce more serious disturbances, including temporary extensive imbalance in the body chemistry, an acute hepatitis, and numerous "hangover" effects: nausea, headache, gastritis, thirst because of the shift of water from cellular to extracellular spaces, and a generalized residual malaise and physical and mental incompetence that may last as long as 24 hours after all the alcohol consumed has been metabolized. Some drinkers are satisfied to suffer the mild and even the severer aftereffects of occasional intoxication for the sake of the temporary dissociation, euphoria, or socialization that the drinking allows. But frequent intoxication, even of moderate degree, imposes a severe burden on the organism. The irritating effects of alcohol, especially in undiluted strong beverages, can result in damage to the tissues of the mouth, esophagus, and stomach, perhaps with increased susceptibility to cancer. The liver is likely to suffer if it must cope lengthily with the detoxication of large amounts of alcohol. There may also be damage to the heart muscle.

Frequent heavy drinking to severe degrees of intoxication or the prolonged steady maintenance of a high alcohol concentration in the organism is linked to many injuries of many organs. The most common linkage is with diseases caused by defects of nutrition.

Effects on the brain. Alcohol as a drug affecting the central nervous system belongs in a class with the barbiturates, minor tranquillizers, and general anesthetics and is commonly classified as a depressant. Its effect on the brain is rather biphasic: at quite low concentrations it can serve as an excitant or stimulant of some functions, but, as the concentration increases, the effect is constantly more depressant, going on to sedation, stupor, and coma. The excitement phase exhibits the well-known signs of exhilaration, loss of socially expected restraints, loquaciousness, unexpected changes of mood, and occasionally uncontrolled emotional displays. This excitement phase may result from an indirect effect of alcohol in suppressing the function of inhibitory brain centres rather than from a direct stimulation of the manifest behaviour. The physical counterparts of this state of relatively mild intoxication are slurred speech, unsteady gait, disturbed sensory perceptions, and inability to make fine discriminations. These effects are not produced by the direct action of alcohol on the misbehaving muscles and senses but by its effect on the brain centres that control their activity.

Because brain alcohol concentrations can rarely be measured in people, blood alcohol concentrations are used as the standard against which performances are compared. Most people exhibit some degree of functional depression or incapacitation at blood alcohol concentrations of 0.10 percent. Inefficiency in some tasks may begin at much lower concentrations, even at 0.03 percent. The majority of drinkers begin to show measurable impairment at just above 0.05 percent, and most people are considered intoxicated at 0.15 percent. In determining the intoxication of automobile drivers, most jurisdictions in Western countries use a blood alcohol concentration of 0.15 percent as the criterion; in some, a presumption of impairment is made at concentrations of 0.10, 0.08, or 0.05 percent.

The impairments at these concentrations may not be manifested superficially, but laboratory tests show diminished alertness, visual acuity, and capacity to distinguish between signals. Reflex responses and the time of reaction to a signal, as well as neuromuscular functions, are slowed. Complex reactions, those that require the brain to process more than one type of incoming information simultaneously, are impaired at blood alcohol concentrations too low to affect simple reflexes and reaction times.

Effects on mental processes

The most important immediate actions of alcohol are on the highest functions of the brain—those of thinking, learning, remembering, and making judgments. The biphasic action of alcohol is evident in its effect on thinking. At low blood alcohol concentrations, after the equivalent of two drinks (two ounces of spirits), well-trained and highly intelligent young men performed better at solving problems in symbolic logic than they had without alcohol. At medium concentrations, the equivalent of four drinks, their performance was about normal. But, at concentrations equivalent to six drinks, their ability to solve such problems definitely deteriorated. Similar effects were observed in tests of memorizing. More recent experiments indicate the dependence of learning on the state in which it occurs:—*i.e.*, what is learned under the influence of alcohol is better recalled under the influence of alcohol than when sober, but what is learned in the sober state is better recalled when sober. The function of judgment is difficult to define; when evaluated by such a task as willingness to drive a bus through a more or less adequate gateway, judgment progressively deteriorates with increasing blood alcohol concentrations; when evaluated by willingness to take risks, as shown in responses to hypothetical problem situations, the results are similar, although more so in younger than in older people.

As blood alcohol concentrations continue to rise above 0.15 percent, intoxication steadily increases. Well-adapted, very heavy drinkers may continue to function fairly well in some motor and mental tasks even up to concentrations of 0.3 percent, but most people will appear visibly drunk, manifesting the common symptoms of slurred speech, unsteadiness of gait, and confusion of thinking long before this level is reached. At 0.4 percent most people will be anesthetized to the extent that they will be asleep, difficult to arouse, incapable of voluntary activity—indeed, in a state in which they could undergo surgery. At higher blood alcohol levels, deep coma sets in, and, if the intoxicated person is long unattended, death may ensue by accidental stoppage of the breathing passages. Between 0.5 and 1 percent the breathing centre in the brain or the action of the heart may be anesthetized, and then death directly from alcohol intoxication will quickly follow. Ordinarily, the achievement of blood alcohol concentrations above 0.4 percent by drinking is unlikely; in a man of average build it would require the ingestion and absorption, unmetabolized, of almost a fifth of a gallon of spirits (0.757 litre) at 50 percent alcohol.

Effects on emotional behaviour. People ordinarily drink alcohol to obtain effects that they have been taught

to expect. Small amounts are drunk in the expectation of reducing feelings of tension, relieving feelings of anxiety, and, conversely, obtaining feelings of gaiety and exhilaration. A sufficient amount of alcohol will usually serve the desired purpose. It is likely, however, that the state of expectation combines with the pharmacological action of the drug to produce the desired effect. In favourable circumstances, alcohol will not merely reduce tension and anxiety but suppress them entirely, even allowing a shift of the emotional state to one of indifference or euphoria or elation. The anxiety-suppressing action of alcohol is commonly seen in the gradual removal of social inhibitions. Shy people become outgoing or bold; well-behaved people become disorderly; the sexually repressed become amorous; the fearful become brave; the quiescent or peaceful become verbally or physically aggressive. All these reactions may be the result of the stimulating effect of small amounts of alcohol or the control-anesthetizing (disinhibiting) effects of larger amounts. But they are, in part, also made possible by the social and cultural permissiveness typical of drinking situations. Alcohol is not only a psychoactive but a socioactive drug. These effects apparently occur in most normal drinkers; in alcoholics, the consumption of huge quantities apparently evokes different states of feeling and even an increase of anxiety or tension.

Alcohol and therapeutics

Alcohol is often used for medicinal and therapeutic purposes. Whiskey is popular for treating colds and snakebites, brandy for treating faintness, wine for blood building, beer for lactation, and any alcoholic beverage for treating sleeplessness or overexcitement. Many of these uses survive from folk medicine. Alcohol is administered by physicians in hospitals, usually by vein, sometimes for anesthesia before minor surgery; more often it is given for sedation after surgery and as a source of easily absorbed calories when it is desirable to bypass the digestive system. Physicians often prescribe "a drink" for a variety of purposes: to stimulate a sluggish appetite, as a sedative to induce sleep, as an anxiolytic in premenstrual tension, as a vasodilator in arteriosclerosis, to relieve the vague aches and pains that beset the elderly, and as a supplement in special diets.

Direct effects on organs. After being considerably diluted upon absorption into the bloodstream, alcohol reaches the body organs in extremely small concentrations. A pint (0.473 litre) of whiskey (at 50 percent alcohol), fully absorbed and not metabolized, will produce in a man of average build and weight a concentration of about 0.32 percent in the blood. The concentration in any organ is likely to be less. Concentrations of 1 percent are hardly sufficient to irritate the most delicate tissue. Even with very heavy continuous drinking, a post-absorptive concentration above 0.5 percent in any organ is extremely unlikely. Such small alcohol concentrations cannot be expected to inflict direct damage on the liver, the kidneys, the generative organs, the heart, or the brain, though they may cause temporary derangements of biochemistry and function. A concentration of 0.5 percent in the brain will cause coma and may be lethal. Most of the damage to these and other organs in long-time heavy drinkers consists apparently of side effects related to disturbances of metabolism or to malnutrition and poor hygiene. Some experimental work has suggested that alcohol has a directly adverse effect on the liver and on muscle tissue, and clinical observations have suggested a direct chronic pathological effect on the heart muscle. But the manner in which alcohol produces such direct damages is still unexplained.

The serious indirect effects of alcohol

The acute diseases. Excessive users of alcohol suffer from both acute and chronic diseases. The acute diseases include, first of all, intoxication, with its disturbances of neuromuscular and mental functions and of body chemistry. In addition, the intoxicated person is abnormally liable to accidents and injuries. Those who experience severe intoxication chronically, the alcoholics, are 7 times as liable to fatal accidents as persons in the general population. They are 30 times more liable to fatal poisoning, 16 times more liable to death from a fall, and 4.5 times more liable to death in a motor-vehicle accident. These liabilities obviously reflect not only the disablements of immediate intoxication but the whole lifestyle and state of alcoholics.

Another range of diseases associated with intoxication are those that occur in the post-intoxication state, the alcohol-withdrawal syndromes. The commonest and least of these is the hangover. After a prolonged bout of drunkenness, severe withdrawal phenomena often supervene, consisting of tremulousness, loss of appetite, inability to retain food, sweating, restlessness, sleep disturbances, sometimes epileptiform seizures, and abnormal changes in body chemistry, especially in the electrolyte balance.

In extreme cases, delirium tremens will develop, a gross trembling of the whole body, sometimes with seizures, mental clouding, disorientation, and hallucinations both visual and auditory. Depending on the amount and quality of care and treatment, as well as the possible occurrence of intercurrent diseases, delirium tremens lasts from 3 to 10 days, with a reported fatality rate ranging from 1 to 20 percent. Acute alcoholic hallucinosis may develop with or without preceding delirium tremens and may last days to weeks. Other brain disorders that may follow a prolonged bout include Wernicke's disease, resulting from an acute, complete deficiency of thiamine (vitamin B_1) and marked by clouding of consciousness and paralysis of the optic nerves; the acute form of Korsakoff's syndrome, marked by loss of recent memory, with a tendency to make up for the defect by confabulation; and Jolliffe's encephalopathy, resulting from an acute complete deficiency of nicotinic acid and marked by delirium, together with rigidities of the extremities and sucking and grasping reflexes.

Early signs of polyneuropathy, a degenerative disease of the nervous system, include tenderness of calf muscles, diminished tendon reflexes, and loss of vibratory sensation. Acute hepatitis (inflammation of the liver) is common. Less often seen are severe pancreatitis and Zieve's syndrome, the latter being a combination of jaundice with hemolytic anemia, hyperlipemia, and hepatic tenderness. Numerous other disturbances, for example, severe muscle cramps and defects in nerve conduction, are occasionally observed.

The disorders of the brain and nervous system and possibly many of the symptoms in the other acute consequences of alcoholism are caused largely by dietary deficiencies. Alcohol provides large numbers of calories, but they are empty calories—such as those of refined sugar—devoid of vitamins and other essential microelements of nutrition, including minerals and amino acids. The small amounts of vitamins and minerals present in beers and wines—rarely in spirits—are insufficient for dietary needs. In heavy bouts of drinking, alcoholics neglect normal eating or, because of digestive difficulties, cannot absorb enough of the essential food elements. These nutritional defects are the cause of many of the chronic diseases associated with alcoholism.

Another acute disorder associated with alcohol is sometimes misnamed pathological intoxication in which, after downing a small amount of drink, a person suddenly breaks into a maniacal fury, smashing furniture and attacking people. The seizure ends as suddenly as it began, and the victim usually falls into a stuporous sleep, from which he awakens after some hours—usually without memory of the events. The cause is uncertain; personality factors combined with a temporary deficiency of blood sugar have been suspected.

Degenerative diseases

The chronic diseases. In long-lasting alcoholism, one or more of the chronic nutritional deficiency diseases may develop. The severer effects of vitamin B_1 deficiency are probably commonest—degeneration of the nervous system, with permanent nerve damage in extreme cases, and beriberi heart disease; another common disease is pellagra, due chiefly to deficiency of nicotinic acid. Other diseases include scurvy, resulting from ascorbic acid deficiency; hypochromic microcytic anemia, caused by iron deficiency; and anemia resulting from vitamin B_{12} deficiency. Severe open sores on the skin of alcoholic derelicts whose usual drink is the cheapest form of alcohol—

low-quality fortified wines—are sometimes miscalled wine sores, but they result from a combination of multiple nutritional deficiencies and poor hygiene.

The classical disease associated with alcoholism is cirrhosis of the liver (Laënnec's type), frequently preceded by a fatty enlargement of the organ. The exact mechanism by which this cirrhosis develops is still unclear; it appears to involve the consequences both of the metabolism of excessive amounts of alcohol and of defective nutrition. In its severest form, Laënnec's cirrhosis is often fatal; its successful treatment or the retardation of its progress is impossible in an alcoholic who cannot be stopped from drinking.

Besides the mental symptoms that may accompany pellagra, other mental disorders that are more specifically related to the consumption of alcohol include chronic hallucinations and Korsakoff's psychosis, which may last months to years. A relatively uncommon chronic brain disorder is Marchiafava-Bignami's disease; because the symptoms are not specific, this degeneration of the corpus callosum in the brain is usually diagnosed only at autopsy. Other brain damage occasionally reported in alcoholism includes cortical laminar sclerosis, cerebellar degeneration, central pontine myelinolysis, and enlargement of the ventricles as a result of atrophy of brain substance. Alcoholics, especially older ones, frequently show signs of neurological defect and impaired brain function, and these are sometimes blamed on direct damage to the central nervous system by alcohol; it is possible, however, that they are the result of brain damage caused by accidents and blows. Many of those who survive long years of alcoholism show a generalized deterioration involving the brain, muscles, endocrine system, and vital organs, giving an impression of premature old age. Alcoholics die more frequently in all types of accidents and from numerous diseases; their average life span is reduced by about 10 or 12 years.

II. Social conditions of alcohol consumption

HISTORY OF THE USE OF ALCOHOL

In early societies. The origin of alcoholic beverages is lost in the mists of prehistory. Fermentation can occur in any sugar-containing mishmash, such as one of grapes, fruits, berries, or honey, if left exposed in a warm atmosphere. Airborne yeasts will act on the sugar, converting it to alcohol and carbon dioxide. This is the process of fermentation. Alcoholic beverages were thus probably discovered accidentally in the pre-agricultural gathering stage. Early man presumably liked the effects, if not the taste, and proceeded to purposeful production; from merely gathering the wild-growing raw materials, he went on to regular cultivation of the vine and other suitable crops. Few preliterate people did not learn to convert some of the fruit of the earth into alcohol. In the case of starchy vegetation, quite primitive agriculturists learned how to convert the starch to fermentable sugar by providing the necessary zymase from their saliva through such a simple process as preliminary mastication.

Universality of alcohol production

The making of wines and beers has been reported from several hundred preliterate societies. Their importance is evident in the multiplicity of customs and regulations that developed around their production and uses. They often became central in the most valued personal and social ceremonials, especially in the rites of passage, and are ubiquitous in such activities as births, initiations, marriages, compacts, feasts, conclaves, crownings, magic, medicine, worship, hospitality, war making, peace making, and funerals. Alcohol is thus the oldest and still probably the most widely used drug.

The manufacture and sale of alcoholic beverages was already common and commercialized and regulated by government in the earliest civilizations. The oldest known code of laws, that of Hammurabi of Babylonia (*c.* 1770 BC), regulated drinking houses. Sumerian physician-pharmacists prescribed beer (*c.* 2100 BC) in relatively sophisticated pharmacopoeias found on clay tablets. The later Egyptian doctors, in their medical papyri (*c.* 1500 BC), included beer or wine in about 15 percent of their prescriptions. The recently discovered Semitic cuneiform literature of the northern Canaanites, in pre-Biblical Ugarit, contains abundant references to the ubiquitous religious and household uses of the intoxicating fluid.

Early religious uses of alcohol

Water, a precious commodity in the earliest agriculturally dependent civilizations, was probably the original fluid used as offering in worship rites. In time, other fluids—milk, honey, and later wine (in some religions, beer)—were substituted. That alcoholic beverages should have displaced other fluids in early religions, both as offering and drink, is not surprising: its capacity to help the shaman or priest and other participants reach a desired state of ecstasy or frenzy could not long have escaped observation, and its appreciation was naturally attributed to supernatural spirits and gods. The red wine in religious uses was eventually perceived as symbolizing the blood of life and, in this spiritual sense, ultimately passed into the Christian Eucharist. The records of the Egyptian as well as of the Mesopotamian civilizations attest that drinking and drunkenness had passed from the state of religious rite to common practice, often troublesome to government and accompanied by acute and chronic illnesses. There are ample indications that some people so loved drink and were so abandoned to drunkenness that they must be presumed to have been alcoholics.

Among classical peoples. The significance of the classical history of drinking arises from the fact that after about 300 BC the Greek, Hebrew, and Roman cultures became mingled in a mix that was to influence powerfully the development of European culture. The surviving records of ancient Greek and Roman culture, in classic pictorial and plastic art as well as in the literature transmitting prehistoric memories enshrined in myths, reveal the common and copious use of wine by the gods, as well as by people of all classes. The worship of Dionysus, or Bacchus, the wine god, was the most popular; his festival, the Bacchanalia, has given English one of its literary names for a drunken orgy. His female devotees, the Maenads, worshipped him in drunken frenzies. The Greco-Roman classics abound with descriptions of fulsome drinking and often drunkenness. The wine of the ancient Greeks, like that of the Hebrews of the same time, was usually drunk diluted with an equal part or two parts of water, and, thus, the alcohol strength of the beverage was presumably between 4 and 7 percent. But, as a standard drink, diluted wine was apparently more common than plain water, and there were topers who preferred their wine straight.

The literature of the Greeks does not lack warnings of the evil effects of excess in drinking, but in this they are excelled by the classics of the Hebrew Canaanites. The earliest references in the Bible show that abundant wine was regarded as a blessing, on a par with ample milk and honey, grain and fruit. The eyes of the Judaeans were to be bright from wine, which also was known to gladden the heart of man and to bring relief to those who were bitter of heart or ready to perish. In the national religious culture that developed into the Judaism that has survived to the present, drinking was intertwined with all important ceremonial occasions—from the celebration of the eight-day-old boy's circumcision to the toasting of the soul of the departed and, in between, the wedding, the arrival and departure of every sabbath and festival, and, indeed, any sort of celebration. The purpose of drinking thus became integrated with an attitude of reverence for the sanctity or the importance of the occasion, to the extent that over-drinking, becoming tipsy, would manifestly be inappropriate and disapproved. Drunkenness then became a culturally negative, an alien and rejected behaviour and generally vanished from the Jewish communities. In contemporary terms, drinking was under effective social control, and the result has been the seeming paradox, fascinating to modern students of sociocultural phenomena, that a people with the highest proportion of drinkers exhibit the lowest rates of alcoholism and other alcohol-related problems.

Religious sanctions against drunkenness or drinking

Quite a different kind of religious control was adopted later (in the 7th century) in Islām: the Qur'ān simply

condemned wine, and the result was an effective prohibition wherever the devout followers of Muḥammad in Arabia and other lands prevailed. (The same process occurred some 1,000 years later in Europe, after the Reformation, and later still in the United States, when a number of ascetic Christian sects, resting their ideology on the Bible, made abstinence a fundamental tenet.)

Like the early agriculturists of the Near East, the people of the Far East discovered the technology of manufacturing alcoholic beverages in prehistoric times. Barley and rice were the chief crops and the raw materials for producing the drink that, here too, was incorporated into religious ceremonial, both as drink and libation, with festivals featuring divine states of drunkenness. Here too, in time, the sacred drink became secularized, even while its religious uses survived, and evoked public as well as private disorders. The history of China includes several abortive efforts at control or prohibition. But, in the Far East, prohibition was effective only when religiously motivated. The Hindu Ayurveda (*c.* 1000 BC) skillfully describes both the beneficent uses of alcoholic beverages and the consequences of intoxication and the diseases of alcoholism. The devout adherents of Buddhism, however, which arose in India in the 5th and 6th centuries BC and spread over southern and eastern Asia, abstain to this day, as do the members of the Hindu Brahmin caste. But most of the peoples in India, as well as in Ceylon and the Philippines, in China and Japan, have continued, throughout, to ferment a portion of their crops and nourish as well as pleasure themselves with the alcoholic product.

In Africa, maize, millet, bananas, honey, the saps of the palm and the bamboo, and many fruits (including that of the sausage tree) have been used to ferment nutrient beers and wines, the best known being Kaffir beer and palm wines. Most of the peoples of Oceania, on the other hand, seem to have missed the discovery of fermentation. Many of the pre-Columbian Indians of North America are also exceptional in lacking alcoholic beverages until introduced to firewater by Europeans, with explosive and disastrous consequences. But the Papago Indians of the southwestern United States made a cactus wine, and the Tarahumara of northern Mexico made beers from corn and species of agave, while throughout Central and South America the aborigines made chicha and other alcoholic beverages from maize, tubers, fruits, flowers, and saps. For the most part, their drinking appears to have been regulated so as to inhibit individual alcoholism and limit drunkenness to communal fiestas.

Variety of uses for alcohol

Alcohol in complex societies. *Contrasts and similarities between primitive and modern uses.* In primitive societies the uses of alcoholic beverages were multiply motivated. These beverages had important nutritional value. They were the best medicine available for some illnesses and especially in relieving pain (in any case, a patient given a prescription to be taken in beer or wine, with the instruction to drink it liberally, was likely to feel better, regardless of whether the various ingredients affected his disease). They facilitated religious ecstasy and communion with the mystical super-mundane powers thought to control tribal and individual fate. They enabled periodic social festivity and the personal jollification of the participants, thus also serving as the mediator of popular recreation. They helped reduce anxiety, tension, and fears connected with concerns over subsistence, safety, or with the need to engage in aggressive or dangerous activities, such as warfare. They could calm anger or tranquillize hostility so as to allow pacification and reduce suspicion, making possible peaceful associations and commercial or ceremonial relations. And, in individuals with extraordinary responsibilities, such as chiefs, shamans, medicine men, they helped to assuage the personal anxieties and tensions connected with those exceptional roles. In some cases the formalized public binge could serve as the permissive loosener of interpersonal aggressions, allow an interlude of verbal or even physical hostility within the family or clan group, which otherwise the mores of the cohesive small society necessarily forbade. The insults and wounds of the discordant interlude could easily be forgiven by blaming them on the alcohol-induced irresponsibility. Under these circumstances drunkenness could be approved or even mandatory and still serve an integrative social function. In short, the most general effect of alcohol, suggested by its very equivocal uses, appears to be as a facilitator of mood change in any desired direction.

The conditions of primitive or early societies foreshadow the conditions of complex societies, including the contemporary highly industrialized ones. As food, hardly anything but the caloric value of alcohol remains, and this is essentially of no importance in modern dietaries, unless such large amounts are consumed that the value becomes negative, since the carbohydrate calories in modern beverages are deficient in other necessary nutrients, such as vitamins and minerals, provided by other foods. As medicament, alcohol has a limited surviving value in the practice of contemporary physicians equipped with more efficient anesthetics and taught to prescribe more variegated analgesics and tranquillizers. In religion, wine has been relegated to a highly specific, essentially symbolic role, where not completely eliminated. What remains is the satisfaction of personal and group needs that have particular urgency arising out of the very complexity of modern life.

The most distinctive features of the complex technological society, in contrast to the simple primitive type, consist of specialization of functions; membership in limited interest groups (though often with conflicting roles); inequalities of earnings and prestige; social and economic stratification, in addition to religious differentiation; and an irreconcilable disharmony of individualism and interdependence, marked by the fact that individuals command skills and privileges indispensable for the welfare of others (only a baker can produce bread for the supermarket's customers; only a licensed physician can write a prescription), yet the same individuals are totally dependent on others for their own vital needs. All these factors contribute to competitiveness, distrust, insecurity, anxiety, alienation. Those who have abundant goods or positions of power and prestige are not thereby freed of these troubles. The result is, if anything, a more intense need for relief from tension, for relaxation from anxiety, for tranquillization from the effects of competitive effort.

It is not that the ancient uses have been forgotten: a drink is still the symbolic announcer of friendship, peace, agreement, in personal as well as in business or political relations. But in the complex society many people discover that many drinks often help them to suppress the overwhelming inhibitions, shynesses, anxieties, tensions that interfere with urgent needs to function effectively, either socially or economically. In the complex society, characterized by multifarious frustrations of gratifying interpersonal relationships, the capacity of alcohol to serve as socio-activating medicine comes into high value.

Government regulation of alcohol use

Conflicts over drinking. Modern complex societies are troubled by a lack of consensus around many issues of right and wrong or proper and improper behaviour; and drinking, since the latter part of the 18th century, has been a focus of disagreement, sometimes amounting to political warfare among subgroups making up larger national societies. In the United States, the late 18th-century temperance movement became, by the middle of the 19th century, an anti-alcohol movement that culminated in national Prohibition (enacted by constitutional amendment in 1919 and repealed in 1933). Echo movements in other countries had somewhat similar histories. The lack of consensus regarding who may drink how much of what and where and when and with whom is illustrated by the crazy quilt of local regulations extant in the United States. In some localities there is total prohibition or prohibition only of distilled spirits and strong wines; in some, only those over 18 or over 21 years of age may buy drinks; in some, married women under age may buy but not married under-age men; in some, until recently, Indians might not buy; in some, liquor may be sold only by the bottle, not by the drink; in some, drinks may be served only together with food, in others only without

food; in some, drinking in public places is permitted only if the drinkers are curtained or only if they are uncurtained or only when they are seated; or men may stand to drink, but women must be seated. Dissonant attitudes toward a custom as common as drinking is believed by many sociologists to account for the inability of the society to establish firm rules inhibiting immoderate behaviour, with a resulting high incidence of damaging use, drunkenness, and many other problems related to alcohol. The Chinese and the Italians, as well as the Jews, are cited as examples of groups having well-developed cultural consensus against drinking to drunkenness, with resulting low rates of alcohol problems. In parallel, France is cited as a country with a consensus favouring steady copious drinking, with a resulting high rate of alcoholism.

The modern conflict over drinking reflects the complex interactions of individual man and small group and larger society. Small groups, formed by common interests in business or occupation or recreation or neighbourhood or politics or ethnicity or religion, use drinking together to facilitate mixing, to engender solidarity, to reduce normal inhibitions against trust in and collaboration with "strangers," to symbolize and ratify accord, to ensure that gatherings for celebration will succeed as festive occasions. Individuals use alcoholic beverages as an agreeable effector of desired mood alteration, such as moderate relief from anxiety or depression or masking of unease and pain, and to enable participation in the various small groups with which they are required to associate. Given favourable contexts, moderate amounts of alcohol apparently function well in fulfilling these purposes. Given consensual practices, moderate amounts of drink function integratively within families and in common-interest groups. This is thought to account for the survival of the drinking custom from early times, in spite of the problems it has engendered and the opposition it has provoked. But individuals and, sometimes, groups, whether formally or informally organized, also indulge in immoderate, self-injurious, and socially damaging drinking. These dysfunctional behaviours account, in part, for the organized societal opposition to any drinking, although the antagonism is sometimes based in more fundamental perspectives of the anti-alcohol sentiment: alcohol is, from olden times, the facilitator of pleasurable, morally lax, hedonistic behaviour; as such, it evokes the displeasure and condemnation of those favouring moral strictness and an ascetic way of life.

DRINKING PATTERNS

Places of drinking

The universal and general functions of drinking are displayed in a great variety of ways and customs in different parts of the world and among various subgroups and subcultures within larger societies. The places of drinking vary infinitely: home may be the only place to drink or the one place where it is forbidden. Drinking may be a ceremonial or informal family affair, with or before meals. It may be a lone practice at home, in commercial drinking places, or in private hideaways or a group practice in membership clubs, neighbourhood taverns, beer gardens, sidewalk cafés, or skid-row alleys. The purposes and occasions are infinitely disparate. It may be the benign drunkenness of the men in the quasi-religious fiestas of Central and South American natives, the fiestal abandonment in song and dance—after relatively little drinking—of the members of the Jewish sect of Ḥasidim, the socializing celebrations of fraternal collegians, or the exhibitionistic libationism of business meetings and professional conventions. Or it may be the repeated "killing" of bottles by men or women fixed mindlessly before the television set, or the quick potations of cocktail parties, or the gulping of forbidden liquors by defiant, rebellious youth groups needing to assert their grownupness. Not only an individual's taste, predilection, or psychological need but also his sex, age, residential neighbourhood, education, associations, church and other memberships, and socio-economic status may determine whether, when, what, how much, or with whom he shall drink. In the United States, where nearly one-third of adults are abstainers, the better educated and the economically advantaged are more likely to be drinkers than the poor, though, among the poor who do drink, the proportions of heavy drinkers are higher. By contrast, in France, where abstainers are in a very small minority, they are more likely to be found among the better educated and upwardly mobile. The attitudes toward drinking and abstaining, among or within different countries, are as varied as the practices.

National comparisons of drinking patterns

Kinds and amounts. The drinking patterns of different countries can be compared in many possible ways; one is by the types of beverages consumed and the actual amounts of alcohol from each. The Table shows the consumption in 25 countries in the latest year for which both alcohol and population statistics were available. The consumption is shown in terms of litres of the total of absolute alcohol contained in all the beverages, as well as in litres of each main class of beverage and of alcohol derived from each class. The importance of converting each beverage class to the absolute alcohol it provides arises from the enormous difference in alcohol content among beverages, ranging from 2 to 5 percent in beers to 8 to 17 percent in wines and to over 45 percent in some spirits. The consumption is further broken down, in the Table, as an average per person in the population of each country aged 15 years and over. The choice of 15 as the "drinking age" is based on observations that in the countries dealt with, for the most part, the population aged under 15 consumes only an insignificant portion of the alcoholic beverages; the resulting division gives a more realistic picture of the amounts people drink than would the per capita of total population. Of course, a more realistic comparison would be based on a division of the alcohol among actual drinkers, omitting abstainers. But the proportions of drinkers and abstainers are known reliably only in the United States, Canada, and Norway. In these three countries, the corrected division would raise the amount of alcohol available to each drinker by 47, 39, and 43 percent respectively.

In terms of total alcohol consumption, the Table shows France heading the list, being highest in wine, very high in distilled spirits, and with a medium ranking in beer and cider. Italy, second highest in wine and with a medium spirits and low beer ranking, is second. The United States, with the second highest spirits ranking and medium in beer and low in wine, is thirteenth, exactly in the middle of this group of 25 countries in rank as well as in quantity of absolute alcohol. Lowest is Israel, in spite of a higher consumption of spirits than beer, and more beer than wine—in fact, ranking higher in spirits than seven other countries.

Neither the amounts nor the classes of beverages in themselves are sufficient to describe the drinking patterns of countries nor the resulting effects in terms of problems or pathologies.

Drinking in France and Italy

Kinds and customs. In both France and Italy, wine consumption is high, but attitudes as well as patterns and amounts differ in the two countries in many ways. French parents tend to exhibit strong attitudes, either favourable or negative, to their children's drinking; Italian parents typically introduce their children to wine drinking without any emotional overtones. Italian standards of safe limits for drinking are lower than those of the French, and they regard getting drunk with disdain, while the French look on it with good humour or even, in men, as a mark of virility; also, as already noted, the French are heavily given to the use of spirits. Although these patterns are not consistent among the regional populations and socio-economic groupings of either country, they are thought to be significant in accounting for the much higher mortality and morbidity from alcoholism in France.

Among the Scandinavian countries (including Finland), the pattern is not one of drinking daily or with meals but rather of telescoped drinking on weekends or special occasions; this is believed to account for the relatively high rate of alcohol-connected problems, such as intoxication, even in a country such as Finland, where the total alcohol consumption is relatively very low. The Scandinavian

Apparent Consumption of Distilled Spirits, Wine, and Beer*
(in litres per capita†)

	year	spirits		wine		beer		total
		beverage	absolute alcohol	beverage	absolute alcohol	beverage	absolute alcohol	absolute alcohol
France	1968	7.96	3.97	146.72	15.40	78.26‡	3.25	22.62
Italy	1969	4.23§	2.12	152.51	12.20	14.95	0.67	14.99
Austria	1970	7.14	2.86	54.27	5.70	130.52	5.87	14.43
West Germany	1970	10.21§	3.88	16.00	1.60	182.00	8.01	13.49
Portugal	1970	1.52	0.68	95.78	11.97	17.59	0.79	13.44
Switzerland	1966–69	6.18	2.47	51.90	5.45	111.25‡	5.00	12.98
Spain	1971	7.20§‖	3.24‖	75.15	9.39	4.93	0.22	12.85
Australia	1969	2.55	1.45	11.50	1.96	168.79	8.44	11.85
Hungary	1969	5.95§	2.98	48.73	5.36	68.61	2.75	11.09
Belgium	1967	2.52	1.26	11.89	1.66	159.23‡	7.96	10.88
New Zealand	1967	2.91§	1.66	5.46	0.93	163.25	8.16	10.75
Czechoslovakia	1969	7.29	2.92	17.51	2.20	179.17	5.38	10.50
U.S.	1971	9.92	4.47	7.87	1.25	98.03	4.43	10.15
Denmark	1969	3.40	1.70	6.66	1.00	132.66	5.97	8.67
Canada	1969	7.14§	2.86	4.82	0.77	95.47	4.77	8.40
U.K.	1970	2.14	1.22	4.91	0.83	114.70	5.16	7.21
Sweden	1970	8.40	3.30	8.10	1.00	65.80	2.90	7.20
Ireland	1970	3.71	2.11	4.81	0.51	98.71	4.44	7.06
The Netherlands	1969	5.19	2.59	6.79	1.16	70.90	3.19	6.94
Poland	1970	8.69§	4.35	7.61	1.06	42.66	1.19	6.60
Japan	1968	4.17§	1.32	19.32§¶	3.08	31.13	1.40	5.80
Finland	1969	5.38	2.42	4.97	0.70	60.09	2.64	5.76
Norway	1970	4.83	2.07	3.10	0.44	48.70	2.22	4.73
Iceland	1971	8.52	3.41	2.86	0.41	16.00♀	0.56♀	4.38
Israel	1970	3.44	1.72	5.23	0.63	15.80	0.80	3.15

*The alcohol content of beverages varies among countries. Calculations for this table are based on the best available information from each country. Only countries from which adequate recent statistics on both alcohol and population are available have been included. †Population aged 15 years and over. ‡Includes cider. §These values converted from original reports in different terms. ‖Based on 1970 data. ¶Includes sake. ♀ Estimated from 1966 and 1968 data.
Source: V. Efron, M. Keller, and C. Gurioli, *Selected Statistics on Consumption of Alcohol and on Alcoholism* (1972).

countries have also had strong anti-alcohol movements, often supported by government funds, and have had abstainer societies with large memberships. It is possible, therefore, that alcohol is consumed by a smaller number of drinkers than is represented by the drinking-age population; in Norway, for example, more than 20 percent of men and nearly 40 percent of women are abstainers.

In England and Ireland, the pub has maintained its popularity as a main locus of drinking. In both countries, beer is by far the most popular beverage—in the United Kingdom accounting for more than three-fourths of the total alcohol consumed and in Ireland for two-thirds. The marked preference for beer is seen in other countries overwhelmingly settled and influenced by British populations—Australia, New Zealand, and much of Canada. In these countries, too, the pub tends to dominate the drinking style, apparently allowing an agreeable pattern of prolonged but not necessarily intimate socializing, while preserving the degree of independence that individual drinkers may prefer. Drinking to a moderate grade of intoxication appears to be acceptable and in some of these countries is thought to constitute the permissive societal attitude that partly underlies the development of alcoholism.

The drinking patterns of very few total populations in European countries have been subject to formal examination by social scientists. Studies have been focussed, for the most part, on segments of the population regarded as problem giving, such as alcoholics, traffic offenders, criminals, inmates of mental hospitals, or youths, especially students. Information from eastern European countries tends to be of the same order. There are indications that, in Poland, the shifting of a young population from rural-farm to new urban-industrial centres, with higher cash incomes and few outlets, has resulted in what the authorities regard as an increase in heavy drinking or alcoholism; vodka is the common drink. Somewhat similar reports emanate from Yugoslavia, where old village customs, while permissive of much drinking, nevertheless used to curb gross excesses. In the Soviet Union the growing program of establishing sobering-up stations and treatment clinics in many cities, often with research-oriented staffs, indicates a recognition that alcoholism is a serious problem. Vodka is the national drink, though native wines are increasingly being promoted. But the situation is evidently quite varied in different parts of the country. In the Soviet republic of Georgia, a viticultural region, wine is the favoured drink, and the drinking patterns are much more like the Italian than those of western Russia, with, apparently, a decidedly low rate of problems.

Drinking in eastern Europe

Much of the description of drinking patterns in South and Central America, as in Africa and Oceania, has consisted of anthropologists' observations of small, preliterate groups, with occasional notations on the effects of acculturation. Studies of national patterns do not exist, although, as in Europe, there are occasional reports on special, either local or problem-giving, populations. In Chile, a middle class population is described as similar to some European populations in drinking patterns—moderate consumption, drinking at home with meals, and frowning on drunkenness being typical. The much larger working class customarily drinks outside the home, in male company, on weekends or paydays, seeking intoxication that is valued as signalling both friendship and virility. A third population, identified as Indian, has a pattern such as that of the working class. The favoured drinks are pisco, a strong native brandy, and wine. Drinking accompanies all secular and religious fiestas, as well as the celebrations of births, baptisms, marriages, funerals; women, however, are expected to drink very moderately. Similar patterns have been reported from various areas in Bolivia and Peru.

In Japan, traditional drinking, still practiced in large circles, allowed heavy drinking and drunkenness in well-delimited social situations and was socially integrative. The traditional beverage was sake, often called rice wine but, more properly, a beer, brewed by highly skilled methods up to a strength of at least 14 percent alcohol and possibly up to 16 or 17 percent. A great many drinking customs and rituals involving sake were connected with religious and social occasions. Next to sake the common beverage was a distillate from sake mash, *shōchū*, containing about 25 percent alcohol. There are evidences of heavy drinking and alcoholism and various attempts to impose prohibition, as well as the practice of abstinence by the followers of Buddhism and of some revered

Japanese philosophers. Modernization was accompanied in the last quarter of the 19th century by a temperance movement stimulated, in part, by the Woman's Christian Temperance Union and the Salvation Army. Since World War II, the widespread Americanization has included a growing popularity of beer and an increased use of imported beverages, especially whiskey, for which younger people express a preference. Among the young, the development of an ambivalence toward drinking, also characteristic of U.S. attitudes, has been reported.

Drinking in the United States

Drinking patterns and attitudes in the United States have been studied more systematically and completely than those of any other country. The results indicate that there is no pattern or set of attitudes typical of the nation as a whole; instead, there is a variety of patterns, customs, and attitudes. U.S. drinking patterns are a conglomerate of customs brought over by repeated waves of immigrants from different places and of diverse ethnic stocks, modified over time by intermixture, economic circumstances, political developments, and the emergence of some indigenous ways. Nevertheless, certain generalizations are possible. In the post-Prohibition and post-World War II era, several changes in U.S. drinking practices and attitudes have been observed and confirmed by formal systematic studies. The proportion of abstainers has declined, especially among women. In recent years, approximately 77 percent of adult men and 60 percent of adult women are drinkers. Men, however, as is almost the universal custom, drink substantially larger amounts, on average, than women. Among adolescents, about 57 percent of boys and 43 percent of girls are drinkers; these percentages rise with age, and the age class 21–29 contains the highest proportion of drinkers; drinking appears to be associated with separation from home and with courtship behaviour. As they continue to age, apparently many Americans become abstainers, perhaps settling down to living as they were brought up to do by their parents or as they would like to exemplify to their children. Rural populations, those with fewer years of education, lower income, more frequent religious attendance, and membership in fundamentalist Protestant denominations, contain larger proportions of abstainers. Among drinkers, beer tends to be the preferred drink of men and, to a lesser extent, of unskilled and blue-collar workers. Spirits are preferred by middle and upper class drinkers and by women—especially in the form of mixed drinks, such as cocktails. Repeal of Prohibition was advocated with the promise of no return of the saloon. Its successors, the bar, tavern, or cocktail lounge, are a ubiquitous U.S. institution. A good deal of drinking, however, takes place at home—though not, most commonly, with meals—and at parties, especially at the second indigenous U.S. drinking way, the cocktail party, now sometimes a substitute for the vanished frontier drinking style. The popularity of the cocktail party and, more specifically, the cocktail on coming home from work and before the evening meal has been attributed to its function as a separator between the working day and the relaxing evening—a sort of rite of passage, easing the way from one into the other. Whether or not this is the actual reason, the cocktail hour has become increasingly adopted in many countries, like other U.S. customs that are paradoxically sneered at while being assiduously imitated.

Although, in general, styles and customs of drinking are influenced by geographic and ethnic backgrounds, Americans tend to be members of multiple small societies, and, to some extent, they drink differently within each of these societies. People from diverse origins may drink alike when joined in some special association—as fellow collegians, members of a business convention, comrades in one of the armed services, or guests at a special kind of social function. Even then, the expected way and amount of drinking is likely to be at least modified by an individual's background. Even within a situation where drinking—indeed, drunkenness—is institutionalized (that is, on the skid rows), patterned differences are discernible to the systematic observer, giving rise to named classes, such as lushes, bums, winos, and rubby-dubs (habitual drinkers of non-beverage alcohols, such as rubbing alcohol and canned heat), and distinct consequences. The fact that most Americans drink, that drinking rather than abstinence is the norm, does not prevent the paradoxical existence of ambiguous attitudes about the behaviour among the drinkers themselves, many of whom express such views as that which maintains that alcohol, or drinking, is more harmful than beneficent, more wrong than right. These ambivalences account for the massive array of regulations on the sale and distribution of alcohol, most of them intended to interfere with the availability of beverages at certain times, in certain places, or to certain classes of persons.

III. Alcoholism

The conception of inveterate drunkenness as a disease appears to be rooted in antiquity. The Roman philosopher Seneca classified it as a form of insanity. The term alcoholism, however, appears first in the classical essay "Alcoholismus chronicus" (1849) by the Swedish physician Magnus Huss. The phrase chronic alcoholism rapidly became a medical term for the condition of habitual inebriety conceived as a disease; and the bearer of the disease was called an alcoholic or alcoholist (*e.g.*, Italian *alcoolisto*, French *alcoolique*, German *Alkoholiker*, Spanish *alcohólico*, Swedish *alkoholist*).

DEFINITIONS AND CAUSAL FACTORS OF ALCOHOLISM

Alcoholism consists of a repetitive intake of alcoholic beverages to an extent that causes repeated or continued harm to the drinker. The harm may be physical or mental; it may also be social or economic. Implicit in the conception of alcoholism as a disease is the idea that the person experiencing repeated or long-lasting injury from his drinking would alter his behaviour if he could. His failure to do so shows that he cannot help himself, that he has "lost control over drinking." This conception incorporates the idea of addiction or dependence.

Formal definitions of alcoholism vary according to the point of view of the definer. A simplistic old-fashioned medical definition calls alcoholism a disease caused by chronic excessive drinking. A purely pharmacological–physiological definition of alcoholism classifies it as a drug addiction recognizable by the need for increasing doses to produce desired effects and by the occurrence of a withdrawal syndrome when drinking is stopped. This definition is inadequate, since alcoholism does not resemble other addictions in the need for increased doses. Opium addicts become adapted to and require as much as hundreds of times the normal lethal dose, but the increased amounts to which alcoholics become adapted are well below the normal single lethal dose. Moreover, the withdrawal syndromes in alcoholism occur inconsistently, sometimes failing to appear in the same persons who experienced them previously and apparently never occurring in some persons who cannot be distinguished from confirmed alcoholics.

Behavioral rather than pharmacological–physiological signs are much more consistent and reliable in defining and diagnosing alcoholism. A sophisticated definition representing modern conceptions of comprehensive medicine classifies alcoholism as a disease of unknown cause, without recognizable anatomical signs, manifested by addiction to or dependence on alcohol. A more comprehensive definition incorporating the perspectives of both psychological and physical medicine recognizes that alcoholism may be either a symptom of another underlying, possibly psychological, disorder or a disease itself: alcoholism, in this view, is a chronic and usually progressive disease or a symptom of an underlying psychological or physical disorder, characterized by dependence on alcohol (manifested by loss of control over drinking) for relief from psychological or physical distress or for gratification from alcohol intoxication itself, and characterized also by a consumption of alcoholic beverages sufficiently great and consistent to cause physical or mental or social or economic disability. Here, the conception of disease undoubtedly rests on the evidence of disablement.

A comprehensive definition of alcoholism

The various definitions that rely on the symptom of loss of control over drinking often consider the loss of control to consist of an inability to stop drinking once it is started, implying that the alcoholic can choose not to take the first drink. But the more comprehensive definition sees the alcoholic as starting a drinking episode because he cannot refrain. Nor does the loss of control over drinking hold true all the time. As with symptoms in many diseases, the loss of control is active in most alcoholics only inconsistently. This means that an alcoholic is not always under internal pressure to drink and can sometimes resist drinking, or, if he drinks, he can sometimes drink in a controlled way. The inconsistency of the loss of control is, however, consistent with a definition of alcoholism based on learning psychology: alcoholism, in this definition, is a learned (or conditioned) dependence on (or addiction to) alcohol that irresistibly activates resort to drinking whenever a critical internal or environmental stimulus (or cue) presents itself. This definition leaves room for the conception that alcoholism may start as a symptom of an underlying disorder, which induces the learning of the alcoholismic pattern, and that once the pattern is fixed or conditioned it may become a disease in its own right (that is, an addiction), capable of surviving even the disappearance of the original underlying cause. Some theorists who regard alcoholism as primarily a symptom do not necessarily subscribe to the idea that it is learned, although they recognize that it may progress to the state of a primary disease.

Alcoholism is a multifarious phenomenon requiring more than one definition. Epidemiologists need a definition that will enable them to identify alcoholics within a population not available for individual examination. Such a definition may rely on a quantity-and-frequency measurement of drinking and also on behavioral features, including injurious effects measurable by instrumental indexes, such as a formula resting on the relation of alcoholism to diseases of known frequency among alcoholics or a drinking-history questionnaire or a preoccupation-with-alcohol scale. Sociological–behavioral definitions emphasize deviance from a norm, especially drinking that exceeds customary dietary use or diverges from the social customs of the drinker's community; such a definition may use as a criterion the way a drinker is regarded by those who know him; his arrests, hospitalizations, and clinical diagnoses; or his membership in a self-defining group, such as Alcoholics Anonymous. Legal definitions tend to rest on habitual intemperance that endangers others, injures the public welfare, or threatens the health, welfare, or competence of the person himself.

Causal factors

Many theories of the cause of alcoholism rest on the limited perspectives of specialists in particular disciplines or professions. Thus, alcoholism has been thought to be caused by defects of heredity, nutrition, disorders of endocrine function, latent homosexuality, economic misery or affluence, bad social influences, or sinful gluttony. More discerning definitions and descriptions take into account the complexity of alcoholism, acknowledging that its causes are not yet knowable with certainty. The most comprehensive conceptions recognize that alcoholism may have a genetic or constitutional underlying factor—not a fateful heredity but a predisposition that renders some people more vulnerable to alcoholism than others. Some think the genetic vulnerability is specific not to alcoholism but rather more generally to a neurosis or an affective disorder that may manifest itself as alcoholism; the alcoholism may possibly represent a "choice of symptom" and be for some individuals a useful "sickness." Others think the genetic factor may impose not vulnerability but, on the contrary, immunity to alcoholism, meaning that some people are unable to adapt to drinking on a level sufficient to gain the peculiar rewards that dispose a person to the development of an alcoholismic life pattern.

The comprehensive etiological view suspects that factors in infancy or early childhood, such as lack of parental care and love, overindulgence, or inconsistency in rearing practices, may lay the foundation of a vulnerable personality. On such a foundation, a dependent personality type or one marked by dependence–independence conflict may emerge; in adolescence this may manifest itself in an insecure self-sex image and a need to overcompensate—for example, by defiant exhibitionistic deviance. Such a problem-ridden personality may find exceptionally effective assuagement and reward in alcohol and learn to rely on intoxication as a mechanism for coping with problems. If this learning process is not interrupted and especially if the social surroundings respond encouragingly or permissively or ambivalently to heavy drinking and intoxication, then the vulnerable personality will become conditioned to react to difficulties by resort to intoxication. If the process lasts long enough, the outcome will be addiction to alcohol or a confirmed alcohol dependence. This comprehensive conception takes into account not only the possible genetic, pharmacological, psychological, and social factors but also the sociocultural context. It recognizes that the society defines and labels the phenomenon of alcoholism, that the culture contributes to its development or inhibition, and that behaviour that in one culture matches an adequate rational definition of alcoholism may not constitute alcoholism in another. Thus, periodic intoxication causing sickness for several days and necessitating absence from work may define alcoholism in a modern industrial community, but, in a rural Andean society, periodic drunkenness at appointed communal fiestas, resulting in sickness and suspension of work for several days, is normal behaviour. An essential aspect of the difference is that the drunkenness at fiestas is not individually deviant behaviour.

PREVALENCE OF ALCOHOLISM

Estimates of the prevalence of alcoholism vary greatly, depending on how it is defined as well as on the methods of estimation. In the United States in 1970, according to one sophisticated estimate, there were approximately 5,400,000 alcoholics—about 4,500,000 men and 900,000 women. In percentage terms, 7.3 percent of men and 1.3 percent of women were alcoholics or 4.2 percent of adults aged 20 and over. There were large variations among regions and states, the rates being higher in urban and industrialized areas. There was no objective evidence that the rates of alcoholism had risen since World War II, although the absolute numbers had increased substantially with the growth of the adult population. A widespread impression that alcoholism was increasing among women apparently reflected the greater visibility of female alcoholics caused by changing public and professional attitudes; formerly, there had been more masking of alcoholism in women than in men.

Alcoholics and pre-alcoholics

A constant rate of alcoholism, without any increase in numbers except in proportion to the growth of population, requires an annual incidence of several hundred thousand new cases. The process of becoming an alcoholic usually takes several years. Since in many cases the process is not carried to completion, there must be a population of several million pre-alcoholics (for example, "heavy drinkers" or "heavy-escape" drinkers) from whom the several hundred thousand new cases of alcoholism emerge each year. On the basis of national surveys of U.S. drinking patterns in the mid-1960s, it is estimated that the size of the "pre-alcoholic" population is about 4,000,000; these, together with the 5,400,000 alcoholics, may be considered the total of problem drinkers. There are indications, however, that among the pre-alcoholics the sex differential is rather smaller than among the alcoholics, probably about four men to one woman. This implies that among women pre-alcoholics a smaller proportion cross the line to become full-fledged alcoholics.

The existence of over 5,000,000 alcoholics in the U.S., plus possibly 4,000,000 other problem drinkers, of whom perhaps between 5 and 10 percent become alcoholics each year, places alcoholism in the front rank of public-health problems. Its gravity is underlined by the higher rates of mortality (2.5 times normal) among alcoholics. Suicide rates are 2.5 times higher; accidental death rates are seven times higher; and there is an enormously higher

rate of general morbidity among alcoholics. One study found that among patients in general hospitals, those identifiable as alcoholics range from 13 to 29 percent. Alcoholism-related psychoses accounted for about 15 percent of the male and 3 to 4 percent of the female admissions to public and private psychiatric hospitals in the U.S. in the 1960s. Admissions of alcoholics without psychosis—usually to participate in alcoholism-treatment programs—accounted for another 40 percent of the men and 13 percent of the women admitted to public mental hospitals and for 15 percent of the men and 4 percent of the women admitted to private mental hospitals. These statistics do not include pre-alcoholics and problem drinkers, although, from the viewpoint of preventive public health, they are those most in need of study and help.

Estimates for different countries

Variations in the definition of alcoholism make it difficult to compare U.S. rates with those of other countries. The most comparable statistics are those of Canada, where the rate of alcoholism is much lower than in the U.S., about 2.4 percent, and the ratio of incidence among the sexes is about five men to one woman, as in the U.S. A rate of 3.5 percent has been reported from Sweden and 1.1 percent from Finland, each with a ratio of five men to one woman, and 0.8 percent from Northern Ireland, with a ratio of three men to one woman; other rates include 5.4 percent in Chile and 0.41 percent in Italy, with no indications of sex ratios. In England and Wales different estimators have suggested rates varying from as low as 1.1 percent to as high as 8.8 and 11 percent; and in Switzerland the suggested rates have varied from 2.2 to 13 percent. The rate in France has been estimated at as high as 15 percent of the adult population, but more conservative estimates suggest 9.4 percent.

Although the rate in France is probably higher than in any of the other countries mentioned, the degree of validity that may be attached to these estimates is so uncertain that all comparisons must be considered as unreliable. There is a strong subjective element in statistics of alcoholism. From time to time, professional opinion becomes aroused, a cry of alarm is raised, and the assumption is made that alcoholism is increasing. High estimates are then likely to emerge, based on local and insufficiently refined data. Often, increased admissions to hospitals for alcoholic mental disorders and sometimes increased consumption of alcohol are cited in evidence. But these data invariably fail to take account of changes in availability or use of facilities, changes in admission or diagnostic policies, or changes in the source of beverages —for example, from unrecorded to recorded supplies. In the U.S.S.R. a change in the internal political situation with the death of Stalin resulted in a shift from official denial that any significant problem of alcoholism existed to an outcry that its prevalence was widespread and serious, though no statistics were provided.

TREATMENT OF ALCOHOLISM

The various treatments of alcoholism may be classified as physiological, psychological, and social. Many physiological treatments are given as adjuncts to psychological methods, but sometimes they are applied in "pure" form, without conscious psychotherapeutic intent or even with an effort to avoid it.

Drugs that prevent drinking

Physiological therapies. *Chemical fences.* One of the popular modern drug treatments of alcoholism, initiated in 1948 by Eric Jacobsen of Denmark, uses disulfiram (tetraethylthiuram disulfide). The usual technique is to administer half a gram in tablet form daily for a few days; then, under carefully controlled conditions and with medical supervision, the patient is given a small test drink of an alcoholic beverage. The presence of disulfiram in the drinker's body causes a reaction of hot flushing, nausea, vomiting, a sudden sharp drop of blood pressure, pounding of the heart, and even a feeling of impending death. These symptoms result from an accumulation of the highly toxic first product of alcohol metabolism—acetaldehyde. Normally, as alcohol is converted to acetaldehyde, the latter is rapidly converted, in turn, to other harmless metabolites, but in the presence of disulfiram—itself harmless—the metabolism of acetaldehyde is blocked, with the resulting toxic symptoms. The patient is thus dramatically shown the danger of attempting to drink while under disulfiram medication. A smaller daily dose of disulfiram is then prescribed, and the dread of the consequences of drinking acts as a "chemical fence" to prevent the patient from drinking as long as he continues taking the drug. Most therapists use the period of enforced abstinence to apply psychological and rehabilitative measures that should enable the patient ultimately to refrain from drinking without the chemical crutch. Variations of the technique include group-reaction tests and the substitution of motion pictures or verbal descriptions for the reaction test.

Citrated calcium cyanamide is another drug used with similar effect, preferred by some therapists because the reaction with alcohol is milder, though its protective potency is briefer. In Japan some therapists have reported giving very small doses of the cyanamide compound, thereby allowing the patient to drink very moderately without suffering a severe reaction but provoking the reaction if he attempts to drink immoderately. Other substances that can produce disagreeable reactions with alcohol include animal charcoal, the mushroom *Coprinus atramentarius*, numerous antidiabetic drugs, and the ground pine *Lycopodium selago*, but, except for the latter, which has had some trial in the Soviet Union, they have attracted little clinical interest.

Aversion. The U.S. psychiatrist W.L. Voegtlin developed a method of creating a conditioned reflex of aversion to alcohol by repeatedly giving the patient a precisely timed injection of an emetic drug just before a drink of his favorite beverage, resulting in nausea and vomiting before the alcohol could be absorbed. The consequent association of vomiting with drinking, causing aversion to the taste, smell, and sometimes even sight of alcoholic beverages, does not last indefinitely but may be reinforced periodically. Similar techniques have been tried in several European countries. Other methods of conditioning applied by behaviour therapists and learning psychologists include associating drinking with mild to painful electrical shocks or with temporary interruption of breathing by injection of a paralyzing drug.

Nutrition, hormones, drugs. A genetotrophic theory of disease holds that alcoholism is caused by a genetically determined need for extraordinary amounts of one or more vitamins. Accordingly, alcoholics have been treated with massive doses of multivitamins. Another theory holds that alcoholism is caused by some defect of the endocrine system, the adrenal-hypophyseal axis being most commonly implicated, and, accordingly, alcoholics have been treated by injections of adrenal steroids and adrenocorticotropic hormones. Other physical and drug therapies that have been tried in alcoholics include intravenous injections of alcohol, apomorphine, injections of autoserum and alcoholized serum, brain surgery, carbon-dioxide inhalation, oxygen by injection, nicotinic acid, nicotinamide-adenine dinucleotide, lysergide (LSD, lysergic acid diethylamide), strychnine, antihistaminic agents, and many tranquillizing and energizing drugs. None of these treatments has been shown in controlled studies to be more effective than others. With some, controlled studies are extremely difficult to carry out. In many cases, moreover, the treatments are accompanied by simultaneous measures having potentially psychotherapeutic and socially rehabilitative effects, especially membership in Alcoholics Anonymous. It is possible that the treatment that works best is the one that is most suitable for the particular patient. But it is also possible that the most effective therapy is the one the therapist believes in, and this may account for the inferior results achieved when any of them is tested in controlled experiments. In the use of psychoactive drugs such as LSD, the aim often is not directly to affect the alcoholism but to produce changes in the patient's emotional state that will help him respond to other psychosocial measures.

Other therapies. *Psychological therapies.* Psychotherapy in alcoholism encompasses the entire range of modalities applied in treating the psychoneuroses and char-

Treatment of neurotic factors

acter disorders, including individual and group techniques. The aim varies from eliminating some underlying cause to effecting just enough shift in the patient's emotional state so that he can function at least temporarily without drinking. Psychoanalysis is rarely tried, having shown little success in alcoholism; analytically oriented therapies are more usual, chiefly with supportive aims. The only psychological technique developed specifically for alcoholism consists of gaining the patient's recognition and acceptance of his actual condition, which alcoholics often resist. Such acceptance may then be followed by a therapeutic–rehabilitative regimen. Group therapies are regarded as more effective than individual modalities with alcoholics. These range from instructional lectures and superficial discussions to deep analytic explorations, psychodrama, hypnosis, psychodynamic confrontation, and marathon sessions. Mechanical aids include didactic motion pictures, movies of the patients while intoxicated, and taped records of previous sessions. Some therapists have experimented, as yet without definitive results, with milieus that reward and reinforce socializing behaviour, hoping thereby to extinguish the desocializing drinking behaviour. Many institutional programs rely on "total push," subjecting the patient to a bombardment of methods, including drugs, hypnosis, physiotherapies, group sessions, lectures, Alcoholics Anonymous meetings, and individual psychological and religious counselling, with the hope that each patient will be affected favourably by whatever is most suitable for him. Other institutional programs rely on mere removal from the stressful outside environment, with a period of enforced abstinence. The therapists themselves may be psychoanalysts, psychiatrists, clinical psychologists, pastoral counsellors, social workers, nurses, police or parole officers, or lay counsellors—the latter often former alcoholics with special training.

The places of treatment are as varied as the modalities, ranging from general hospitals to mental hospitals to mental-health outpatient clinics to specialized inpatient sanitariums and specialized alcoholism clinics to jails and penitentiaries to medical and psychiatric private offices, with patients often moving, randomly or systematically, from one milieu to another.

Awareness of the social and environmental elements in alcoholism has led to the development of treatment for spouses and occasionally for whole families, either separately or jointly, in the recognition that "the patient" is not just the alcoholic but the family unit.

A new trend in the United States, partly stimulated by court decisions prohibiting the jailing of alcoholics for public intoxication, is the establishment of detoxication centres that provide first aid along with guidance toward more fundamental treatment. But even if adequate programs and facilities for treating alcoholism were available, it is unlikely that they would solve the problem, given the large number of new cases each year. Only preventive public-health programs can eliminate alcoholism and thus far no likely methods of prevention have been devised.

Alcoholics Anonymous. The patient-centred self-help fellowship of men and women called Alcoholics Anonymous enables its members to share their common experience and thus to help each other. AA was founded in 1934 by two alcoholics, one of whom had been strongly influenced by the Oxford Movement, and has grown to a worldwide community of hundreds of thousands. The members strive to follow Twelve Steps, a nonsectarian spiritual program the central points of which are reliance on God or a higher power as each individual understands that concept and the value of help to other alcoholics. The fellowship is organized in local groups of indeterminate size, has no dues, and accepts contributions for its expenses only from those attending meetings—where members narrate the stories of their alcoholic careers and their recovery in AA. Affiliation of the society or its groups with churches, politics, organizations, or institutions is barred by the AA Twelve Traditions. AA apparently meets deep-seated needs among its members by enabling them to associate with kindred sufferers who understand them, to accept the disease concept of alcoholism, to admit their powerlessness over alcohol and their need for help, to depend without shame or stigma on others, and to involve themselves in activities within the group and in helping other alcoholics. These seem to provide adequate substitutes for the alcoholismic way of life. AA is thought by many to be the single most successful method yet devised for coping with alcoholism. It has spawned some allied but independent organizations: Al-Anon, for spouses and other close relatives and friends of alcoholics, and Alateen, for their adolescent children, the aim of which is to help the members understand alcoholism, understand themselves better, and learn how to help an alcoholic or, at any event, how to live with one. Professionals in the field tend to think of AA as an inexpensive form of group therapy and a useful ally but recognize, as do the more sophisticated members, that it is not suitable for all types of alcoholics. Most experienced therapists agree that any form of treatment is likely to show a higher rate of success if the patient can be persuaded simultaneously to join Alcoholics Anonymous.

The effectiveness of Alcoholics Anonymous

AA groups around the world resemble each other and generally use the ideological literature (including translations) published by the central office in New York, although there are some variations in style and conduct. In some countries the AA groups are sponsored by or affiliated with national temperance societies or accept financial support from government health agencies. There are also clubs for former alcoholics, usually sponsored by a particular institution for its former patients. One Scandinavian group seeks to achieve a stable degree of moderate drinking, rather than total abstinence.

Results of treatment. The success of treatment in behavioral or personality disorders is always difficult to appraise, and this is the case in alcoholism. The effects of new treatments tend to be reported enthusiastically, but critical examination of the results tends to reduce or cast doubt on the rate of apparent success. Controlled studies, when carried out, usually undercut the claims. Follow-up studies of persons treated have usually been too brief to determine whether permanent results had been achieved, and in most cases the investigators failed to locate a substantial proportion of the former patients. Moreover, the measures of "success" are inconsistent. Some investigators regard only total abstinence as a successful outcome; others are satisfied if drinking bouts are curtailed and the patient's life adjustment is improved. Perhaps between 25 and 50 percent of alcoholics who receive some form of treatment either become abstinent or achieve some abatement of the severity of their illness. Alcoholism treatment programs connected with businesses and industries, in which the alcoholic must participate if he wants to keep his job, have reported even higher success rates. Forms of frankly compulsory treatment, even if grudgingly endured by alcoholics, seem to have a high rate of effectiveness. Some investigators have suggested that the older the patient and the longer the duration of his alcoholism, the more frequent is the occurrence of "spontaneous" recovery.

IV. Contemporary alcohol problems and controls

Useful versus detrimental uses of alcohol

Effects of drugs are commonly discussed in terms with negative inference—as if, with drugs, including alcohol, effect and detriment were necessarily synonymous. The personally functional and socially integrative uses of alcohol tend to be overlooked. Alcohol does indeed serve in prophylactic, nutritional, and medicinal roles. The vast majority of drinkers in most of the world are light, occasional, and moderate drinkers—normal drinkers who experience no harm from their own use of alcoholic beverages. Thus, relatively small minorities fall into the class of heavy or excessive or problem drinkers, including alcoholics. Nevertheless, the problem drinkers invoke sufficient troubles for themselves, their families, their employers, their occupational or social associates, and their communities and society generally so that "alcohol problems" are major causes of disorder and suffering, as well as costly focusses of study and responses.

INDIVIDUAL AND SOCIAL EFFECTS

First, in the realm of health, the most serious and detrimental effect is alcoholism. Although drinking itself is hardly ever regarded as the sufficient cause of alcoholism, this disease could not arise without the use of alcohol. Next come the alcoholic diseases—physical and mental disorders that are caused directly or indirectly by alcoholism or heavy drinking. Moreover, alcoholics and heavy drinkers are especially susceptible to the development of some other diseases, not specifically alcoholic, and are then less able to withstand the vicissitudes of ill health. Finally, alcoholics and problem drinkers undoubtedly contribute to the deterioration of the mental health of other members of their families and often to the breakdown of family life. This is seen in the enormously increased rates of divorce and separation in families in which one of the spouses is an alcoholic (4 to 12 times higher than in comparable populations) and in the extraordinarily higher rate of alcoholism in the offspring of alcoholics (up to 10 times higher than in comparable samples). This last is now considered to be a social consequence rather than a result of genetic defect.

The social and economic costs of alcoholism and heavy drinking are essentially incalculable. An estimate of $2,000,000,000 as the cost of health and welfare services provided to alcoholics and their families in the United States alone suggests the measure of effects worldwide. Furthermore, the millions of problem drinkers who have jobs and businesses are more frequently absent and often less efficient than their occupational associates. In one study in a very large industry, identified alcoholics accumulated two and a half times as many days absent from work and collected nearly three times as much in sickness payments as matched samples of fellow employees. Crude projections of the annual costs of alcoholism to the national economy of the United States range from $7,000,000,000 to $10,000,000,000.

Crime and motor violations

The role of alcohol in crime has often been gauged by the presence of alcoholics in prisons and jails. It is necessary to discount the self-exculpatory blaming of "drink" by convicts as the cause of their crimes. Moreover, the high proportion of alcoholics among those convicted of serious crimes—estimates range from 22 to 43 percent—may misrepresent the role of alcoholism. It is possible that alcoholics, ineffective in other aspects of life, are also less efficient than nonalcoholic criminals in escaping detection and conviction; this may account for their high frequency in penal institutions. There is no question, however, that a great portion of the work of police departments and costs of local courts and jails is attributable to arrests, prosecutions, and brief incarcerations for public intoxication.

A special offense related to drinking is alcohol-impaired driving of motor vehicles and the resulting high rate of accidents, with fatalities, personal injuries, and property damage. Alcohol, for example, is involved in about one-third of the more than 50,000 annual road traffic fatalities in the United States, in possibly 500,000 injuries to persons, and in more than $1,000,000,000 worth of property damage. Although people with very low alcohol concentrations in their blood—the equivalent of one or two drinks—do not figure in accidents more often than those with no alcohol, the chances of being involved in a traffic accident rise precipitously with increasing blood alcohol concentrations beyond these minimal levels. Laws making specified blood alcohol concentrations prima facie evidence of being drunk, impaired, under the influence of alcohol, or unfit to drive have been passed in most jurisdictions. In some Scandinavian countries the limit is 0.05 percent; in most countries, 0.08 percent; in West Germany, 0.13 percent; in U.S. jurisdictions, from 0.08 to 0.15 percent. Attempts to curb driving while affected by alcohol have taken the form of severe punishments—heavy fines, mandatory jail sentences, and loss of driving licenses for periods ranging from six months to ten years. Newer ways of coping with this problem are being sought with the consideration that, since drivers and pedestrians involved in accidents after drinking usually have very high blood alcohol concentrations, they are probably alcoholics and should be dealt with under some therapeutic rather than exclusively penal regimen.

ALCOHOL CONTROL

Governmental efforts to control alcoholic beverages go back as far as recorded history. The Code of Hammurabi included regulations of prices, taverns, and sellers. That the laws often failed to produce the desired effects—temperance and good public order and perhaps revenue exceeding the social costs of excess—is inferred from the frequent legislative attempts at total prohibition in numerous lands throughout history, all apparently without lasting success. The most resounding failure was that in the United States from 1920 to 1933. Current prohibitions in parts of India appear to be equally ineffective.

Types of governmental control

Less totalitarian efforts to control the use of alcohol include licensing systems that limit the number and locations of places of sale; restriction of days and hours of sale; prohibitions of sale to the young, with ages varying from 16 in Yugoslavia to 21 in parts of the United States; regulation of the strength of beverages, the size of containers, advertising, prices, or profits. Some governments—for instance, those of Finland and several states of the U.S.—have sought to eliminate the private-profit motive from the sale of alcoholic beverages by reserving a monopoly in the trade to themselves. All Communist countries have government monopolies. But indications that this has made any difference in the kinds, degree, or severity of problems are lacking; apparently problem drinkers do not need to be persuaded by profit seekers. Some governments—for instance, those of Sweden, Finland, and the U.S. state of Ohio—have attempted to control individual drinking by a system of personal ration books for purchases. In Sweden this system was abandoned after 38 years of trial; evidently, those who needed to drink a lot could find supplies—even when their ration books were withdrawn. The most universal regulation of alcoholic beverages takes the form of taxation (or, in government monopolies, an added profit), which is often quite heavy. Usually, however, though the taxing policy may have the ostensible purpose of reducing consumption or controlling licensees, the real object is revenue. In any event, none of the common forms of government control have proved themselves able to promote temperance in those whose need to drink heavily is uncontrollable. The persistence of massive restrictions and regulations, with costly enforcing bureaucracies, reflects the tendency of legislatures to give some satisfaction to the substantial minority of convinced opponents of alcohol, and the tendency of segments of the drinking population—and even of some people in the liquor trade—to accept the naïve notions of the anti-alcoholists that these regulations do some good.

Control by temperance movements and religions

Though not significantly influenced by governmental efforts, the rate and severity of alcohol problems have indeed been influenced by nongovernmental movements and agencies. The most obvious example is the success of religious movements, such as Buddhism, Islām, and numerous Christian denominations and sects, in confirming their followers as total abstainers. The Methodist and Baptist denominations, the Quakers, the Mormons, the Christian Scientists, the Seventh-day Adventists, and the Jehovah's Witnesses are examples of Christian churches that have made abstinence a condition of loyal membership, though the Methodist Church has modified its stand in recent times. Other Christian denominations, such as the Congregationalists, have advocated abstinence without making it a requirement. Though not formally allied with these churches, the Woman's Christian Temperance Union (WCTU) and other temperance societies, particularly in the United States, once drew much support from them. In several European countries the abstinence movement also drew some support from the Socialist-influenced labour movement and found some organizational expression in the form of fraternal orders, particularly the Order of Good Templars. The importance of the religious orientation is indicated by the relatively larger proportion of abstainers in the

United States than in countries where the ideal was more largely politically motivated. The decline in the numbers of abstainers in recent times may reflect the changing character of religious adherence.

THE NEW SCIENTIFIC ORIENTATION

In the past generation the character and influence of citizen movements have changed markedly. Whereas in former times the personnel, teaching aids, and ideologies of the temperance movement had generally dominated the research and education regarding alcohol, the tendency now is toward deriving objective information from academic and scientific sources. Among the major efforts to bring a scientific orientation to bear on the consideration of alcohol problems has been the founding of a systematic documentation and publication of the biological and social-science knowledge of the entire world; this is now a function of the Center of Alcohol Studies at Rutgers University in New Brunswick, New Jersey.

International concern

The new trend had its repercussions also on international cooperation. The International Bureau Against Alcoholism, founded in 1907, became, in 1964, the International Council on Alcohol and Alcoholism—more recently renamed the International Council on Alcohol and Addictions. The change of name represented a change in aims and policies, from total opposition to any drinking to advocacy of an objective consideration of alcohol problems. This change was manifested also in the character of the international congresses convened by anti-alcohol organizations since 1885 and by the International Bureau since 1925. Formerly devoted essentially to descriptions of the horrible effects and denunciations of the evils of alcohol, beginning with the 26th Congress in 1960 the program has been infiltrated by presentations from the scientific-academic world, and the 28th Congress in Washington, D.C., in 1968 was marked by the total absence of representation from the remnants of the old temperance movement; the papers and lectures offered by representatives of religious organizations and societies were on an equal level of scholarship and objectivity with those from the scientific and academic community.

These developments, in turn, have had repercussions in government activities. In the United States and Canada they have led to the establishment of some 55 state and provincial agencies, some independent, most attached to departments of health or mental health, with missions chiefly to provide treatment for alcoholics but often also to participate in education and occasionally to engage in research. A few county and city agencies have also been created. On the federal level in the United States, a National Institute on Alcohol Abuse and Alcoholism has been established, and the Department of Transportation is engaged in a program aimed to reduce the alcohol-and-traffic problem. Other governments have shown recognition of the potential of newer, science-oriented approaches and have supported research and education as well as therapeutic activities, sometimes through special institutions such as the Canadian Addiction Research Foundation, supported by the province of Ontario; the Finnish Foundation for Alcohol Studies; the Norwegian National Institute for Alcohol Research; and the Northern Committee for Alcohol Research, with membership from all the Scandinavian countries. The new ferment discernible in the 1970s in the field of alcohol problems is thus far stimulated mainly by concern over the human and economic costs of existing problems and the aim to alleviate them. The idea that only the prevention of alcohol problems or that only the reduction of the numbers of persons who become newly involved in alcohol problems can effect permanent gains is universally given lip service but rarely is the object of direct effort. This is owing to two facts: those who have been chiefly involved in bringing about the new ferment have been inspired mainly by the humanitarian and economic motive to bring relief to those who are already suffering from the consequences of alcohol problems, and practical new methods of prevention, going beyond scattershot education, have yet to be invented.

BIBLIOGRAPHY. The prime sources for scientific, scholarly, and professional information and documentation on all aspects of the human use of beverage alcohol and its effects are the Rutgers University Center of Alcohol Studies, *Quarterly Journal of Studies on Alcohol:* Part A, *Originals*, Part B, *Documentation* (1940–), and its *Classified Abstract Archive of the Alcohol Literature* (CAAAL), a cumulative indexed collection of world scientific literature, on edge-notched finding cards (1939–). A key to the meanings of both technical and popular terms relating to alcohol is provided by MARK KILLER and MAIRI MCCORMICK, *A Dictionary of Words About Alcohol* (1968). The most complete critical reviews of the acute and chronic pharmacological, physiological, and psychological effects of alcohol intake are in HENRIK WALLGREN and HERBERT BARRY III (eds.), *Actions of Alcohol*, vol. 1, *Biochemical, Physiological and Psychological Aspects*, vol. 2, *Chronic and Clinical Aspects* (1970); H. KALANT, "The Pharmacology of Alcohol Intoxication," *Q. Jl. Stud. Alcohol*, suppl. no. 1, pp. 1–23 (1961); and H. CAPPELL and C.P. HERMAN, "Alcohol and Tension Reduction: A Review," *ibid.*, 33:33–64 (1972). BERTON ROUECHE, *The Neutral Spirit: A Portrait of Alcohol* (1960), is a well-researched, popular account of the history and the beneficent as well as harmful effects of alcoholic beverages. R.G. MCCARTHY (ed.), *Drinking and Intoxication* (1959), is a rich collection of informative writings from the sociological as well as biological fields on every aspect of its title; while D.J. PITTMAN and C.R. SNYDER (eds.), *Society, Culture, and Drinking Patterns* (1962), brings together the best essays on social and cultural factors associated with drinking patterns in many kinds of communities and groups. Anthropological studies of the drinking practices of early societies, as well as their survival and transmission to and significance for modern man, are described in E.M. LOEB, "Primitive Intoxicants," *Q. Jl. Stud. Alcohol*, 4:387–398 (1943); CHANDLER WASHBURNE, *Primitive Drinking: A Study of the Uses and Functions of Alcohol in Preliterate Societies* (1961); and M.K. BACON *et al.*, "A Cross-Cultural Study of Drinking," *Q. Jl. Stud. Alcohol*, suppl. no. 3 (1965). The prevalence of drinking and abstaining and the distribution and demographic interrelations of alcohol problems and alcoholism in the United States and other countries are dealt with in DON CAHALAN, IRA H. CISIN, and HELEN M. CROSSLEY, *American Drinking Practices: A National Study of Drinking Behavior and Attitudes* (1969); and K. BRUUN *et al.*, "Surveys of Drinking and Abstaining: Urban, Suburban and National Studies," *Q. Jl. Stud. Alcohol*, suppl. no. 6 (1972); while VERA EFRON, MARK KELLER, and CAROL GURIOLI, *Statistics on Consumption of Alcohol and on Alcoholism* (1972), provides comparisons of alcohol consumption in many countries and presents estimates of alcoholism and statistics on alcoholic psychoses and mortality in the United States. The nature and causes of alcoholism and descriptions of its symptomatology are dealt with in MARK KELLER, *Some Views on the Nature of Addiction* (1969); DAVID LESTER, "Self-Selection of Alcohol by Animals, Human Variation, and the Etiology of Alcoholism," *Q. Jl. Stud. Alcohol*, 27:395–438 (1966); E.M. JELLINEK, *The Disease Concept of Alcoholism* (1960), and "Phases of Alcohol Addiction," *Q. Jl. Stud. Alcohol*, 13:673–684 (1952). Treatments of alcoholism are described in M.J. HILL and H.T. BLANE, "Evaluation of Psychotherapy with Alcoholics: A Critical Review," *Q. Jl. Stud. Alcohol*, 28:76–104 (1967); and MORRIS E. CHAFETZ, HOWARD T. BLANE, and MARJORIE J. HILL (eds.), *Frontiers of Alcoholism* (1970), while the ideology and role of the leading self-help fellowship are detailed in *Alcoholics Anonymous: The Story of How Many Thousands of Men and Women Have Recovered from Alcoholism* (1955); and BARRY LEACH *et al.*, "Dimensions of Alcoholics Anonymous," *Int. J. Addict.*, 4:507–541 (1969). Analyses of classic as well as contemporary attempts to cope with alcohol problems are outlined in the NEW YORK STATE MORELAND COMMISSION ON THE ALCOHOLIC BEVERAGE CONTROL LAW, *Study Paper No. 1: The Relationship of the Alcoholic Beverage Control Law and the Problems of Alcohol* (1963); and S.D. BACON, "The Classic Temperance Movement in the U.S.A.: Impact Today on Attitudes, Action and Research," *Br. J. Addict. Alcohol*, 62:5–18 (1967). Broad data and perspectives on alcohol-related health and social problems are presented in THOMAS F.A. PLAUT, *Alcohol Problems: A Report to the Nation, by the Cooperative Commission on the Study of Alcoholism* (1967); and the NATIONAL INSTITUTE ON ALCOHOL ABUSE AND ALCOHOLISM, *Alcohol and Health, 1st Special Report . . . from the Secretary of Health, Education, and Welfare* (1972).

(M.Ke.)

Alcohols, Phenols, and Ethers

Alcohols, phenols, and ethers are organic chemical compounds—that is, they belong to the general family of

compounds whose molecules are based on chains of carbon atoms. These substances differ from all other organic compounds in that their molecules contain at least one oxygen atom joined to a carbon atom solely by a single bond (as the connections between atoms are called). In alcohols and phenols this oxygen atom is part of a hydroxyl, or $-OH$, group; that is, the oxygen is joined to a hydrogen atom as well as to a carbon atom, $\equiv C-OH$. In ethers the oxygen atom is bonded to two carbon atoms, $\equiv C-O-C\equiv$.

The hydroxyl group

Generally, the oxygen-bearing carbon atoms in alcohols, phenols, and ethers are part of hydrocarbon chains, series of carbon atoms to which hydrogen atoms are attached. Phenols differ from alcohols in that the hydrocarbons to which their oxygen atoms are joined are aromatic. Used in this sense, the word aromatic refers to certain rings of carbon atoms containing alternating single and double bonds, a system with unusual properties.

It is instructive to compare the basic chemical structures of these compounds to the structure of the water molecule (see Table 1). Examination of these structures

Table 1: Comparison of the Structures of Water, Alcohols, Phenols, and Ethers

classification	structure*
Water	H–O–H
Alcohols	R–O–H
Phenols	Ar–O–H
Ethers	R–O–R′

*Ar = aromatic ring; R and R′ = hydrocarbon chains.

suggests that alcohols and phenols can be considered derivatives of water, in which a hydrocarbon chain has replaced a hydrogen atom. Experimentally, it is found that alcohols, in particular, resemble water closely, especially when the hydrocarbon group is small. In ethers, both hydrogen atoms have been replaced, and these substances do not resemble water in their behaviour at all.

Another way of characterizing alcohols is to say that a hydroxyl ($-OH$) group has replaced a hydrogen atom in a hydrocarbon molecule ($RH \rightarrow ROH$; the symbol R represents a group of atoms that remains unchanged in a reaction). The most typical alcohols are those in which a hydroxyl group is linked to a carbon atom attached to other atoms solely by single bonds. When the hydroxyl function is attached directly to an aromatic ring, as in phenols, the differences in physical and chemical properties are of such a magnitude that those compounds generally are not considered to be alcohols at all.

ALCOHOLS, PHENOLS, AND ETHERS AS A GROUP

Alcohols are quite reactive substances chemically, and either the hydrogen of the hydroxyl group or the entire hydroxyl group can combine with or be replaced by other groups. Generally, alcohols are neutral compounds showing neither acid nor alkaline properties. Alcohols combine with inorganic compounds to form crystalline products, such as $MgCl_2 \cdot 6CH_3OH$, which are similar to common hydrate salts (inorganic compounds combined with water molecules). Primary alcohols are easily oxidized to aldehydes and to acids ($RCH_2OH \rightarrow RCHO \rightarrow RCO_2H$); secondary alcohols are oxidized to ketones ($R_2CHOH \rightarrow R_2C=O$); and tertiary alcohols on oxidation yield various decomposition products in which the carbon chain is broken ($R_3COH \rightarrow$ decomposition products). Alcohols generally react with carboxylic acids to produce neutral compounds called esters ($RCO_2H + R'OH \rightarrow RCO_2R' + H_2O$). Alcohols also can be converted to ethers under suitable dehydrating conditions ($2ROH \rightarrow ROR + H_2O$).

Every alcohol contains a grouping of atoms $\equiv C-OH$ in which the three bonds on the carbon atom may be joined to three hydrogen atoms, three alkyl (hydrocarbon) groups, or any combination of these. Monohydric alcohols, those with single hydroxyl groups, can exist in various structural modifications. A primary alcohol contains at least two hydrogens on the carbon atom attached to the hydroxyl group; a secondary alcohol has one hydrogen on that carbon; there are no hydrogens on the hydroxyl carbon of a tertiary alcohol (see Table 2).

Table 2: Structure of Primary, Secondary, and Tertiary Alcohols

classification	examples	
	name	formula
Primary ($-CH_2OH$)	methyl alcohol	CH_3OH
Primary ($-CH_2OH$)	propyl alcohol	$CH_3CH_2CH_2OH$
Secondary ($>CHOH$)	isopropyl alcohol	$(CH_3)_2CHOH$
Tertiary ($\geq COH$)	trimethylcarbinol	$(CH_3)_3COH$

Typical phenols have an acidic nature (as contrasted to alcohols and ethers, both of which are neutral, neither acidic nor basic). As acids, phenols exhibit a tendency to form salts with strong bases. Resonance structures (representations of the electronic structure of the molecule) of phenol itself and of the phenoxide ion, which is formed from phenol by loss of a proton, H^+, help explain the acidity of phenols. The resonance forms of phenol involving distribution of negative charge over the benzene ring can be produced only by a simultaneous development of positive charge on the oxygen atom. This means that separation of charges, an energy-consuming process, operates in producing these structures, and they make relatively little contribution to the overall state of the phenol molecule. With the phenoxide ion, on the other hand, no charge separation is involved; negative charge appears on all resonance structures, and those in which the charge is on the benzene ring serve to distribute charge already present. Thus, these forms make significant contributions to the structure of the ion, and it is favoured energetically over the phenol molecule. Since loss of the proton is what converts phenol to the phenoxide ion, the phenol molecule readily loses a proton or behaves in the fashion characterized as acidic.

The phenolic aromatic ring undergoes substitution reactions, such as nitration, chlorination, and sulfonation, more readily than benzene does, due to the presence of the negative charge on the ring, which serves to attract the substituting group. Because the negative charge on the ring appears only at certain locations (called the *ortho* and *para* positions: *ortho* for the ring carbon adjoining the one with the oxygen bond; *para* for the ring carbon diametrically opposite the one with the oxygen bond), it is primarily at these sites that substitution occurs. Under proper conditions, alkyl halides and olefins can react with phenols to form ethers. Esters of phenols are formed by reaction with acid chlorides or anhydrides.

Ethers are far less reactive than alcohols and phenols; they resemble hydrocarbons in that they are relatively inert at ordinary temperatures to strong bases, alkalies, and alkali metals. With strong acids, however, ethers form oxonium salts [$ROR' + HX \rightarrow (RR'OH)^+ + X^-$]. ($X^-$ stands for a negatively charged ion, such as chloride, Cl^-, or bromide, Br^-. These react further with strong acids to produce alkyl halides and alcohols or water.

$$(RR'OH)^+X^- \xrightarrow{HX} RX + R'OH$$

$$(RR'OH)^+X^- + HX \xrightarrow{HX} RX + R'X + H_2O$$

ALCOHOLS

Nomenclature. The word alcohol is derived from the Arabic *kuḥl* (also *kohl* or *kohol*). It originally was used to mean a "very fine powder" but gradually came to connote "essence." Later the term was applied to wine spirits, which were referred to as *alcool vini*, and eventually simply alcohol.

A heterogeneous system of nomenclature has evolved for mono-, di-, and polyhydric alcohols, and the more important members have two or more common names.

Table 3: Examples of Alcohol Nomenclature

formula	parent hydrocarbon	common name	carbinol name	IUPAC name*
CH_3OH	methane	methyl alcohol	carbinol	methanol
CH_3CH_2OH	ethane	ethyl alcohol	methylcarbinol	ethanol
$CH_3CHOHCH_3$	propane	isopropyl alcohol	dimethylcarbinol	2-propanol
$CH_3CH=CHCH_2OH$	2-butene	crotyl alcohol	propenylcarbinol	2-buten-1-ol
$(CH_3)_3COH$	2-methylpropane	*tert*-butyl alcohol	trimethylcarbinol	2-methyl-2-propanol

*Name approved by the International Union of Pure and Applied Chemistry.

Multiplicities of names

Ethanol, for example, is known as ethyl alcohol, grain alcohol, and spirits of wine. The first system of nomenclature for alcohols, which was based on the use of the term "carbinol" for methanol, was originated by a 19th-century German chemist, Hermann Kolbe. In this system, all alcohols are named as derivatives of carbinol, the various substituents being named first. Many secondary and tertiary alcohols are readily named by this system, as are aromatic and heterocyclic derivatives. Carbinol nomenclature is gradually fading out of usage, however.

The presently accepted systematic nomenclature for alcohols was adopted at a meeting of the International Union of Pure and Applied Chemistry (IUPAC), in Paris in 1957. According to the IUPAC system, the name of the alcohol is derived from the name of the parent hydrocarbon (which corresponds to the longest straight chain in the molecule incorporating the alcohol function). The final "-e" of the hydrocarbon name is replaced by "-ol" (or is followed by "-diol," "-triol," etc., for increasing numbers of alcohol functions). An arabic number before the name gives the position in the chain of the hydroxyl group, with the numbering beginning at the end of the chain that gives the hydroxyl the smallest number; *e.g.*, 2-butanol for $CH_3CH_2CHOHCH_3$. For unsaturated alcohols, the position of the hydroxyl group is given by a number before "-ol" and the number indicating the position of the double bond appears first, as 2-buten-1-ol for $CH_3CH=CHCH_2OH$. Certain other functional groups in the molecule, such as the carboxyl group, take precedence over the hydroxyl group in naming molecules and are given the terminal position; the prefix "hydroxy-" is used to denote the hydroxyl group. For example, 2-hydroxypropanoic acid is the systematic name for lactic acid, $CH_3CHOHCOOH$.

A frequent practice in forming common names is to indicate straight chains with a prefix "*n*-" (for normal) and certain branched chains with the prefix "iso-." The latter prefix is usually associated with those compounds containing a terminal $-CH(CH_3)_2$ group and no other branching. This prefix specifically applies to isopropyl, isobutyl, and isoamyl alcohol, but it is occasionally used in a more general sense to indicate any branched-chain structure.

Confusingly enough, the various alcohol nomenclature systems are in more or less simultaneous use. Generally, however, conversions from one form to another are made without difficulty (see Table 3).

Compounds containing two or more hydroxyl groups on separate carbons are named systematically according to the IUPAC rules as polyols, with separate numbers indicating the position of each hydroxyl. These compounds are also given common names, the diols being called glycols and the triols glycerols. Among the higher polyols are pentitols and hexitols.

General physical properties. Typical alcohols are colourless liquids or solids. Normal primary alcohols containing fewer than 12 carbon atoms are liquid at room temperature, whereas those with 12 or more carbons are solid at room temperature. Polyhydric alcohols usually are syrups; alcohols with complex arrangements of carbon atoms, such as the sterols (like cholesterol, which has 27 carbon atoms in its molecule), generally are solids.

Solubilities

Lower alcohols, with up to three carbons in their molecules, are soluble in water, and all monohydroxy alcohols are soluble in organic solvents. The water solubility of alcohols decreases as the molecular weight rises, because the structure becomes relatively more hydrocarbon in nature. Conversely, the boiling points, vapour pressures, densities, and viscosities of alcohols increase with higher molecular weight. Adding hydroxyl groups to an alcohol without increasing the number of carbon atoms increases the solubility in water and decreases the solubility in ether and ethanol; at the same time the melting point, boiling point, density, and viscosity are increased. Hydrogen bonding—that is, secondary attractive forces between hydroxyl groups in adjacent molecules—is the main reason the alcohols have higher melting and boiling points than the corresponding hydrocarbons.

Additional hydroxyl groups in an alcohol enhance its sweetness. For example, ethanol, C_2H_5OH, is not sweet; propylene glycol, $C_3H_6(OH)_2$, is slightly sweet; glycerol, $C_3H_5(OH)_3$, is quite sweet; and mannitol, $C_6H_8(OH)_6$, is so sweet that it is known as a sugar alcohol. Lower alcohols have characteristic odours; alcohols with 8 to 12 carbons have roselike or lilylike odours, which find use in perfumes. The higher alcohols (with more than 12 carbons) are practically odourless. Physical properties of some common alcohols are shown in Table 4.

Occurrence and preparation. Alcohols in free form are not a common occurrence in nature; they are found mainly in the essential or volatile oils obtained from the flowers, leaves, and stems of plants. Chiefly, these are a group of monohydric primary alcohols with carbon chains of 7 to 12 atoms. Their characteristic odours make them valuable ingredients for perfumes. More complex alcohols with branched structures and unsaturated linkages, such as citronellol and geraniol, both 10-carbon alcohols, are found in rose and geranium oils and are

Occurrence in essential oils

Table 4: Physical Properties of Alcohols

IUPAC name*	common name	no. of carbon atoms	formula	melting point (°C)	boiling point (°C)	density (d_4^{20})
Methanol	methyl alcohol	1	CH_3OH	−93.9	64.96	0.7914
Ethanol	ethyl alcohol	2	CH_3CH_2OH	−117.3	78.5	0.7893
1-Propanol	*n*-propyl alcohol	3	$CH_3CH_2CH_2OH$	−126.5	97.4	0.8035
2-Propanol	isopropyl alcohol	3	$(CH_3)_2CHOH$	−89.5	82.4	0.7855
2-Propen-1-ol	allyl alcohol	3	$CH_2=CHCH_2OH$	−129	97	0.8540
1-Butanol	*n*-butyl alcohol	4	$CH_3(CH_2)_2CH_2OH$	−89.53	117.25	0.8098
1-Pentanol	*n*-amyl alcohol	5	$CH_3(CH_2)_3CH_2OH$	−79	138.0	0.8144
1-Dodecanol	lauryl alcohol	12	$CH_3(CH_2)_{10}CH_2OH$	26	259	0.8309^{24}
1-Octadecanol	stearyl alcohol	18	$CH_3(CH_2)_{16}CH_2OH$	59.4	332	0.8124^{59}
9-Octadecen-1-ol	oleyl alcohol	18	$CH_3(CH_2)_7CH=CH(CH_2)_7CH_2OH$	6–7	334	0.8489
1,2-Ethanediol	glycol	2	CH_2OHCH_2OH	−11.5	198	1.1088
1,2,3-Propanetriol	glycerol	3	$CH_2OHCHOHCH_2OH$	20	290	1.2613
Phenylmethanol	benzyl alcohol	7	$C_6H_5CH_2OH$	−15.3	205.35	1.0419

*Name approved by the International Union of Pure and Applied Chemistry.

isolated from citronella oil. The fatty alcohols (primary alcohols with long, straight carbon chains), cinnamic alcohol ($C_6H_5CH=CHCH_2OH$), phenylpropyl alcohol ($C_6H_5CHOHC_2H_5$), and menthol and terpineol (cyclic 10-carbon alcohols) are other alcohols that occur naturally. Several sugar alcohols, containing, respectively, four and five hydroxyl groups on as many carbon atoms, are found in mosses, lichens, and yeasts. Sorbitol, a six-carbon, six-hydroxyl alcohol, occurs widely in fruits, berries, algae, and red seaweed. Mannitol, an isomer of sorbitol, is a major component of the exudate of plane, olive, and manna ash trees, and is found in pumpkin, grasses, mushrooms, marine algae, and brown seaweed. Sterols, such as cholesterol, are present in both the animal and vegetable kingdoms. They are found, for example, in bile, gallstones, brains, blood, egg yolk, and many plant products.

In nature, alcohols are many times more prevalent in combined form than as the free compounds. Fats, for instance, are esters of the trihydric alcohol, glycerol; and certain waxes and oils are esters of fatty alcohols. Many sterols occur in nature in esterified form.

Destructive distillation

Destructive distillation of hardwood was a common process in the early 1900s for producing wood alcohol as well as charcoal, acetic acid, and acetone. The major constituent of wood alcohol is methanol. It is quite difficult to purify this distillate to obtain the pure alcohol, however, so synthetic production has largely supplanted destructive distillation as the chief method used to make high-purity methanol in large quantities for industry.

Carbohydrates (sugar derivatives, such as starch) are a major group of natural compounds that can be made to yield ethyl alcohol by the biological process of fermentation. Before fermentation is begun, grain is cooked to form a starch mass; this mixture is cooled, and malt is then added to supply the enzymes that hydrolyze starch to fermentable sugar. Molasses also can be used as a starting material for fermentation. It contains enough sugar that it can be fermented without preliminary hydrolysis. In either case, yeast is added to bring about the fermentation; enzymes present in the yeast convert the sugars in the natural material to ethyl alcohol and carbon dioxide.

Fermentation

The fermentation process also may be used to produce other alcohols; *n*-butyl alcohol, for example, has been obtained from sterilized molasses by fermentation with a pure culture of bacteria (a species of *Clostridium*). Treatment of glycerol with certain other bacteria yields 1,3-propanediol, as well as *n*-butyl alcohol. Fusel oil, a mixture of impurities that occurs in fermentation ethanol, contains *n*-propyl, isobutyl, isoamyl, and active amyl alcohol (the *d*-isomer of this secondary alcohol). In Europe, glycerol has been produced as a by-product of the alcohol fermentation of sugar. The substance 2,3-butanediol is an example of a glycol obtained from sugar, molasses, or wheat mash by fermentation. Because of their occurrence in nature, many alcohols can be obtained from natural products by purely physical means such as distillation or, as indicated above, they can be liberated from complex chemical structures by means of enzymes, heat, or hydrolysis.

Synthetic production. Specific alcohols, such as ethanol, methanol, and glycerol, are consumed in enormous amounts for literally thousands of applications. Generally, the alcohols produced from natural sources are expensive, difficult to purify, and somewhat unpredictable in their availability. Therefore, in order to meet the demands of industry for alcohols of many descriptions in pure form and at low cost, producers have had to develop special processes that form alcohols from readily available raw materials under carefully controlled conditions.

Aldol condensation. Aldehydes or ketones with a hydrogen atom on the carbon adjacent to the carbonyl group can be condensed to β-hydroxy aldehydes or ketones, respectively. Dehydration followed by catalytic hydrogenation then produces the alcohol; for example, acetaldehyde undergoes self-condensation to give, after dehydration, crotonaldehyde. This product then can be reduced to the alcohol 1-butanol.

$$2CH_3CHO \longrightarrow CH_3CHOHCH_2CHO$$

acetaldehyde — condensation product

$$CH_3CHOHCH_2CHO \xrightarrow[-H_2O]{\text{acid}} CH_3CH=CHCHO$$

condensation product — crotonaldehyde

$$\xrightarrow[2H_2]{\text{reduction}} CH_3(CH_2)_2CH_2OH$$

1-butanol

The Grignard reaction. The Grignard reaction involves addition of an alkyl (or aryl) magnesium halide to an aldehyde, ketone, ester, or epoxide. Hydrolysis of the product produces an alcohol. In generalized terms, reaction of an alkylmagnesium halide with a carbonyl compound is as follows:

$$>C=O + RMgX \longrightarrow >C(R)OMgX \xrightarrow{+H_2O} >C(R)OH + MgOHX.$$

Hydration of olefins (alkenes). A hydration process (addition of water to a double bond) is used industrially to produce ethanol directly from ethylene. The reaction is a general one and may be written as follows:

$$>C=CH_2 + H_2O \rightleftharpoons >C(OH)-CH_3.$$

Indirect hydration may be carried out by adding the olefin to sulfuric acid, with subsequent hydrolysis of the intermediate alkyl sulfate.

$$>C=CH_2 + HSO_4H \longrightarrow >C(OSO_3H)-CH_3 \xrightarrow{H_2O} >C(OH)-CH_3$$

Secondary and tertiary alcohols may be produced in this manner.

Hydroformylation of olefins. Lower-molecular-weight olefins react with carbon monoxide and hydrogen in the presence of a catalyst in a reaction called hydroformylation or the oxo reaction.

$$>C=C< + CO + H_2$$

olefin — carbon monoxide — hydrogen

$$\xrightarrow{\text{cobalt catalyst}} >CH-C(-C(=O)H)$$

aldehyde

The resulting aldehyde is subsequently hydrogenated to form an alcohol.

$$>CH-C(-C(=O)H) + H_2 \longrightarrow >CH-C-CH_2OH$$

Hydrolysis of halides and sulfates. Alkyl halides and sulfates readily undergo alkaline hydrolysis to yield alcohols.

$$RX + HOH \xrightarrow{\text{alkali}} ROH + HX$$

This general procedure produces primary and secondary alcohols. Glycerol can be synthesized from propylene by a series of reactions including hydrolysis of a halide as one step in the process. Unsaturated alcohols can be prepared by high-temperature chlorination followed by hydrolysis of the intermediate halide of an alkene. An example is the production of allyl alcohol from propylene.

$$CH_2=CHCH_3 + Cl_2 \longrightarrow CH_2=CHCH_2Cl + HCl$$

propylene — chlorine — allyl chloride — hydrogen chloride

$$CH_2{=}CHCH_2Cl + NaOH \longrightarrow CH_2{=}CHCH_2OH + NaCl$$

allyl chloride — sodium hydroxide — allyl alcohol — sodium chloride

Hydrolysis of organic esters. The hydrolysis of esters, which is catalyzed by acids and bases, at one time was of considerable importance in producing alcohols from natural esters. While not practiced commercially in the United States, it is employed in Europe and Japan.

$$R'COOR + HOH \longrightarrow R'COOH + ROH$$

Oxidation of hydrocarbons. Saturated hydrocarbons can be oxidized to alcohols in air by the use of high temperatures (260°–370° C), moderate pressures (15–50 atmospheres), and short periods of contact of the hydrocarbon with a catalyst.

Pyrolysis of fatty acids. In one particular instance the sodium salt of a hydroxyl-substituted, unsaturated fatty acid (ricinoleic acid) on pyrolysis gives an alcohol (2-octanol) among the products.

$$CH_3(CH_2)_5CHOHCH_2CH{=}CH(CH_2)_7COONa \xrightarrow[\Delta]{NaOH} CH_3(CH_2)_5CHOHCH_3 + NaOOC(CH_2)_8COONa$$

2-octanol — sodium sebacate

Reduction of carbonyl compounds. At high temperatures and pressures carbonyl compounds can be reduced in the presence of complex catalysts.

$$\diagup C{=}O + H_2 \longrightarrow \diagup CHOH$$

The largest application of this reaction is the production of methanol from carbon monoxide:

$$CO + 2H_2 \longrightarrow CH_3OH.$$

As carbonyl compounds, esters are reduced by alkali metals in alcohol or, in a more modern technique, by high-pressure hydrogenation over a copper oxide–chromium oxide catalyst. The process is used industrially to produce fatty alcohols from the esters of fatty acids.

$$RCOOR' + 2H_2 \longrightarrow RCH_2OH + R'OH$$

Catalytic polymerization of olefins. Since about 1955 straight-chain primary alcohols have been prepared in the petroleum industry from ethylene with the use of triethyl aluminum (Ziegler) catalysts. The course of the reaction is as follows:

$$Al(C_2H_5)_3 + (x{+}y{+}z)CH_2{=}CH_2 \longrightarrow Al\begin{matrix}(CH_2CH_2)_{x+1}H\\(CH_2CH_2)_{y+1}H\\(CH_2CH_2)_{z+1}H\end{matrix}$$

$$Al\begin{matrix}(CH_2CH_2)_{x+1}H\\(CH_2CH_2)_{y+1}H\\(CH_2CH_2)_{z+1}H\end{matrix} + \tfrac{3}{2}O_2 \xrightarrow{H_2O} H(CH_2CH_2)_{x+1}OH + H(CH_2CH_2)_{y+1}OH + H(CH_2CH_2)_{z+1}OH + Al(OH)_3$$

Principal chemical reactions. In general, the chemical reactions undergone by alcohols are essentially related to the hydroxyl groups and their position in the molecule. The nature of the hydrocarbon radical and the presence of other functional groups, however, also exert considerable influence.

Neutrality of alcohols

Alcohols are neutral compounds that do not ionize in water. Nevertheless, when they react with alkali metals, alcohols do act as acids, and when they react with strong acids they are functioning as bases. Tertiary alcohols are the most basic of the alcohols because the alkyl groups donate electrons to the carbon to which the oxygen is attached; this process increases the electron density at the oxygen atom and makes it more attractive toward a proton (*i.e.*, more basic).

Oxidation and the formation of esters and ethers (see above) are probably the most common reactions of alcohols. A number of other reactions also are encountered frequently in work with alcohols.

Ammonolysis and aminolysis. Alcohols generally react with ammonia to give amines and with amines to give further alkylation products.

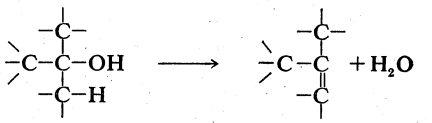
$$ROH + NH_3 \longrightarrow RNH_2 + H_2O$$

alkylamine

$$ROH + R'NH_2 \longrightarrow \begin{matrix}R\\ \diagdown\\ NH\\ \diagup\\ R'\end{matrix} + H_2O$$

dialkylamine

Dehydration. Dehydrating agents remove water from alcohols with the production of unsaturated compounds or, if two molecules of alcohols are involved in the reaction, with formation of ethers. Tertiary alcohols tend to be dehydrated to alkenes the most easily; primary alcohols the least. For this reason, primary alcohols are converted to ethers more readily (without competition from alkene formation).



tertiary alcohol — alkene

$$2ROH \longrightarrow ROR + H_2O$$

ether

Dehydrogenation. Like oxidation, dehydrogenation of primary alcohols gives aldehydes and dehydrogenation of secondary alcohols gives ketones. The high-temperature dehydrogenation of ethyl alcohol to acetaldehyde is a good example.

$$CH_3CH_2OH \xrightarrow{300^\circ C} CH_3CHO + H_2$$

Ethoxylation. The reaction of alcohols with ethylene oxide gives polymeric products in which many units of the ethoxy group ($-CH_2CH_2O-$) are incorporated.

$$ROH + nCH_2\text{–}CH_2 \longrightarrow R(OCH_2CH_2)_nOH$$

polyethoxylated alcohol

Halogenation. Conversion of alcohols to alkyl halides can be accomplished by treatment with hydrogen halides or phosphorus trihalides.

$$ROH + HCl \longrightarrow RCl + H_2O$$

alkyl chloride — water

$$3ROH + PCl_3 \longrightarrow 3RCl + P(OH)_3$$

Reaction with alkali metals. Alkali metals convert alcohols to their alkali metal alkoxides. Reaction of an alcohol with metallic sodium, for instance, produces the corresponding sodium alkoxide.

$$2ROH + 2Na \longrightarrow 2RONa + H_2$$

sodium alkoxide

Sulfation. Treatment of alcohols with fuming sulfuric acid gives alkyl sulfates, a reaction analogous to ester formation with carboxylic acids.

$$ROH + H_2SO_4 + SO_3 \longrightarrow ROSO_3H + H_2SO_4$$

Biological and medical properties. With the exception of methanol, most monohydric saturated alcohols in industrial use are not hazardous as far as cumulative injury

Toxicity

to internal organs is concerned, but some have a cumulative effect on the nervous system. The toxic effect of aliphatic alcohols on the central nervous system is a narcosis, generally increasing with the molecular weight of the alcohol. Apparently the increased effect with increasing molecular weight results from the larger hydrocarbon portion of those molecules, which favours solubility in lipids (fatty materials) of the nervous system. Physiologically, all alcohols are weaker narcotics than their parent hydrocarbons, but metabolic differences influence the degree and nature of toxicity. Ethyl alcohol, for example, is a noncumulative poison because the human body rapidly converts it to carbon dioxide and water. Methyl alcohol, by contrast, is labelled a poison because it has a slower rate of elimination and decomposes to formaldehyde and formic acid, both of which are poisonous to nervous tissue and are able to damage ocular tissues.

As the number of carbon atoms in an alcohol exceeds eight, the resultant low volatility and poor water solubility of the molecule reduces the likelihood of its ingestion or inhalation. There is a decrease of toxicity in a series of alcohols of the same number of carbon atoms in going from primary to secondary to tertiary. Unsaturation brings about a pronounced increase in toxicity, with allyl alcohol being about 30 times more toxic than *n*-propyl alcohol.

Industrial applications. As a group, alcohols are among the most important organic chemicals, judged by the large amounts produced. For example, in the U.S. in the early 1970s, 5,960,000,000 pounds of synthetic methanol, 1,870,000,000 pounds of synthetic ethanol, and 3,300,000,000 pounds of synthetic ethylene glycol were produced annually. The most important industrial use of alcohols is as chemical intermediates, chiefly because alcohols can readily be converted into a large variety of other compounds by simple chemical reactions. For instance, reaction with carboxylic acids leads to fats and waxes; dehydration produces olefins and ethers; sulfonation and ethoxylation form surface active agents for use in detergents; and other syntheses yield compounds used as plasticizers, emulsifiers, lubricants, emollients, and foaming agents.

Alcohols as chemical intermediates

Ethyl alcohol is second in importance only to water as a solvent for drugs, lacquers, perfumes, cosmetics, detergents, and plastics. Recent applications are in aerosol and mouth-wash products, a motor fuel, and a rocket fuel.

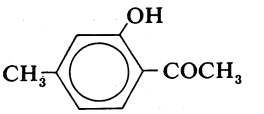
Identification and analysis. Physical methods, such as spectroscopy and chromatography, are frequently used to identify alcohols. Chemically this is done by forming a derivative that characterizes the functional hydroxyl group. An important first step is to convert the crude alcohols or more complex mixtures containing other substances to organic or inorganic esters, which can then be isolated in pure form. Additionally, if the esters are crystalline in nature with well-defined melting points, the melting points may serve for purposes of identification. Esters frequently used for the characterization of alcohols are the phthalates, benzoates, and borates. Carbamates, commonly known as urethans, are popular derivatives of primary and secondary alcohols, particularly *α*-naphthylurethan and *p*-nitrophenylurethan. Formation of complex crystalline salts with calcium chloride is another technique for isolation and purification of alcohols.

Detection of impurities

Often in analysis of an alcohol the substances of interest are the impurities rather than the alcohol itself. Hence, many analytical methods involve separation of minor components by laboratory manipulations, whereas the amount of alcohol is determined by a simple reading on an instrument. The most significant and useful determination involves the hydroxyl function; it is analyzed most frequently by acetylation, phthalation, formylation, and oxidation reactions or by the measurement of active hydrogen. Colour, odour, taste, refractive index, density, and melting and boiling points are properties that aid in identifying alcohols and their states of purity.

PHENOLS

Nomenclature. Phenol is both the specific name for monohydroxybenzene (C_6H_5OH) and the generic name for compounds containing one or more hydroxyl functions attached to an aromatic ring. As with alcohols, nomenclature has been built around trivial names, consequently such names as phenol, cresol (for methylphenols), and naphthol (for hydroxynaphthalenes) are widely used and have been adopted as part of the IUPAC rules. In systematic nomenclature for phenols with more elaborate ring systems, the name for the phenol is obtained by dropping the final "e" from the hydrocarbon name and adding "ol"; thus, the phenols formed from chrysene are termed chrysenols. Frequently, in naming phenols with many substituents, the hydroxyl group is treated as merely another substituent; in this case it is denoted by the prefix "hydroxy-" and a number to locate it on the ring as 4-methyl-2-hydroxyacetophenone for the compound.



General physical properties. Phenols are colourless liquids or white solids at room temperature. Monohydric phenols have a characteristic burning taste. The lowest member of the family, phenol, has a bland and sweetish smell, whereas higher homologues are sharp and spicy smelling. Dihydric phenols, such as resorcinol and hydroquinone, have only slight odours; and trihydric and other polyhydric phenols have essentially no odour at all. Phenols have considerably higher melting points, boiling points, and densities than the parent hydrocarbons from which they are derived (see Table 5).

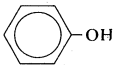
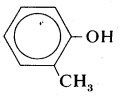
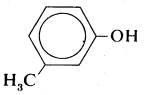
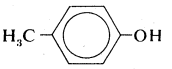
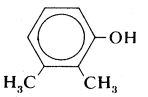
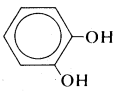
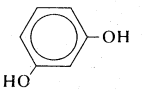
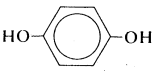
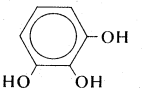
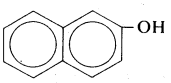
Since the phenolic hydroxyl group is more strongly polarized than that of alcohols, phenols tend to form stronger hydrogen bonds than do alcohols. As a result, phenols show increased solubilities in water and higher boiling points than alcohols, and their solvency for polar organic molecules is greater. Monohydric phenols are partially soluble in water, and polyhydric phenols are much more soluble. Most phenols are readily soluble in oxygen-containing organic solvents but less so in nonpolar solvents such as the aliphatic hydrocarbons. Alkyl substitution on the benzene ring of phenols, however, enhances their solubility in nonpolar solvents.

Phenols have antiseptic and germicidal properties, which increase to a maximum as the length of an alkyl side-chain substituent reaches about six carbon atoms. The lowest member, phenol, is highly toxic and caustic, but these properties progressively diminish with the higher members of the series. Polyhydric phenols are still markedly toxic but less caustic than the monohydric compounds.

Occurrence and preparation. So-called "natural" phenol is obtained by the destructive distillation of coal or from wood tar although it also can be obtained from petroleum and from organic matter of vegetable origin. Other phenols of truly natural origin (that is, not dependent on prior chemical transformation) are found in essential oils, particularly thymol and carvacrol. Phenol itself was first obtained from coal tar in 1834 by the German chemist Friedlieb Ferdinand Runge and called carbolic acid. The French chemist Auguste Laurent in 1841 produced the pure compound from coal tar and named it phenic acid. Although coal tar represented the major source of phenol before World War I, the great use of phenol in plastics spurred the use of synthetic methods of production, with the result that by the mid-1960s over 95 percent of the phenol produced was of synthetic origin. Three general procedures are used for synthesis, each of which utilizes benzene as a starting material; the methods are: sulfonation, chlorination, and alkylation with propylene (which leads to cumene, or isopropylbenzene, as the first intermediate).

In addition to phenol itself, the cresols, xylenols, and higher phenol homologues are obtained from coal tar and other products of the destructive distillation of organic matter (usually of vegetable origin). Coal tar has been a prime source of phenols because of its extensive production in gasworks and coke-oven plants, which heat coal to produce heating gases for fuel and coke for steel

Table 5: Physical Properties of Phenols

	formula	melting point (°C)	boiling point (°C)	density (d^{20}_4)
Phenol		43	181.75	1.0576
o-Cresol		30.94	190.95	1.0273
m-Cresol		11.5	202.2	1.0336
p-Cresol		34.8	201.9	1.0178
2,3-Dimethyl-phenol		75	218	1.16
Pyrocatechol		105	245*	1.1493 (21°)
Resorcinol		111	280	1.2717
Hydroquinone		173–174	285–287	1.328 (15°)
Pyrogallol		133	309	1.453 (4°)
2-Naphthol		123	295	1.28

*At 750 mm Hg.

manufacture. Depending on the source of the coal and the conditions used for carbonization, various complex mixtures of acidic tars are obtained along with tar bases and naphthalene. When this coal tar is distilled, a "middle oil" fraction is obtained that is rich in phenols. Further treatment with aqueous sodium hydroxide and acidification with carbon dioxide separate phenol from neutral and basic impurities and subsequent distillation yields phenol, cresols, and cresylic acids.

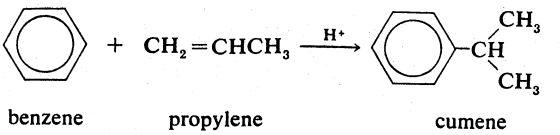
Phenols in free form occur in the so-called essential oils, which are derived from seeds or leaves of plants. Oils of thyme and origanum may contain from 60 percent to 80 percent of thymol or carvacrol. Thymol is also obtained in substantial quantities from Indian ajowan seed oil. Modern demand for these natural phenols has spurred synthetic processes, which are used to supplement material from natural sources. A nonvegetable source of phenols is petroleum distillates; several processes exist for isolating phenols from petroleum tars.

Synthetic production. The following synthetic methods are widely used to make phenol and, on a far smaller scale, certain other monohydroxybenzene derivatives.

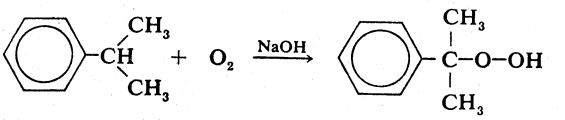
The cumene process. In the first step in this procedure, benzene is alkylated with propylene to give cumene.

Oxidation of cumene

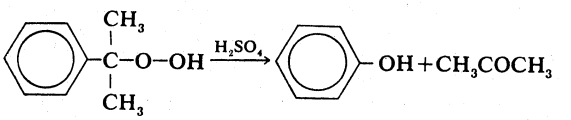
Cumene is then oxidized to cumene hydroperoxide by air in the presence of an alkaline catalyst, and the hydroperoxide is cleaved with an acid catalyst to phenol and acetone. This process may also be used to make cresols from toluene.



benzene propylene cumene

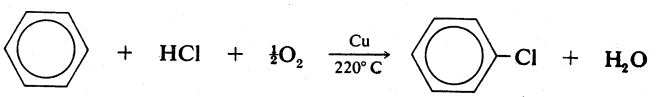


cumene oxygen cumene hydroperoxide

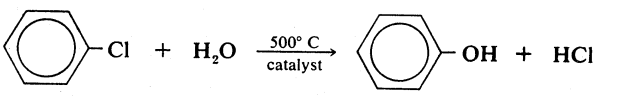


cumene hydroperoxide phenol acetone

Chlorination. In a modification of the old German Raschig process, chlorination of benzene is accomplished with hydrogen chloride and oxygen. Chlorobenzene is hydrolyzed in the vapour phase over a catalyst to give phenol and regenerate hydrogen chloride.



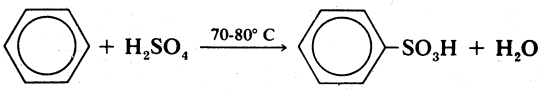
benzene hydrogen chloride oxygen chlorobenzene water



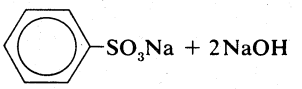
chlorobenzene water phenol hydrogen chloride

In an alternate process the intermediate chlorobenzene is purified by distillation and hydrolyzed with aqueous sodium hydroxide at high temperature and pressure. Formation of a by-product, diphenyl ether, is repressed by recycling this compound in the feed to maintain an equilibrium concentration. Sodium phenoxide is separated by solvent extraction and acidified with hydrochloric acid to give phenol, which is purified by distillation.

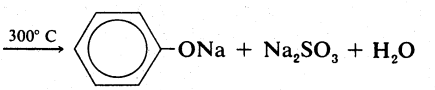
Sulfonation. Benzene is sulfonated in the liquid state with sulfuric acid, and then neutralized with sodium sulfite. The resulting sodium benzenesulfonate is fused with sodium hydroxide, and the phenoxide neutralized with sulfur dioxide to phenol.



benzene sulfuric acid benzenesulfonic acid water

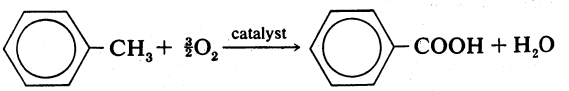


sodium benzenesulfonate sodium hydroxide

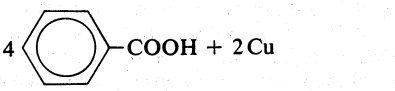


sodium phenoxide sodium sulfite water

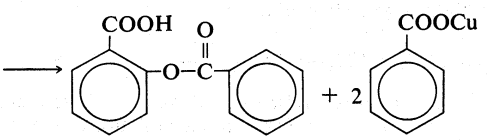
Toluene oxidation. Toluene is oxidized to benzoic acid by the use of air at 130–140° C with a cobalt salt catalyst. The second step employs air, steam, and a copper–magnesium salt mixture to form an intermediate cupric benzoate, which decomposes to cuprous benzoate and benzoylsalicylic acid. Steam hydrolyzes the acid ester to benzoic and salicylic acids, and the latter compound decomposes to phenol and carbon dioxide:



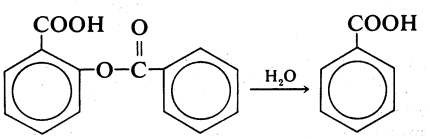
toluene oxygen benzoic acid water



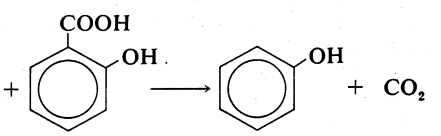
benzoic acid copper



benzoylsalicylic acid cuprous benzoate



benzoylsalicylic acid benzoic acid



salicylic acid phenol carbon dioxide

Cyclohexane oxidation. Cyclohexane is oxidized with air at 130° C in the liquid phase with a cobalt naphthenate catalyst to cyclohexanol and cyclohexanone. The crude product is purified by distillation and dehydrogenated at 380° C over a platinum catalyst. The reaction mixture is fractionated to yield phenol.

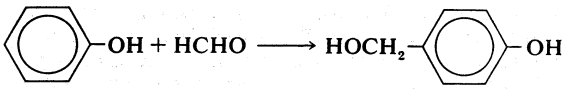
Principal chemical reactions. The chemical properties of phenols differ from those of alcohols chiefly because of the acidic character of the phenols. Because of their acidity, the phenols react with strong bases to form alkali-metal salts known as phenoxides, phenolates, or phenates. These salts are readily decomposed by acids or carbon dioxide.

Phenols alkylate more readily than their parent hydrocarbons, primarily in the *ortho* and *para* positions. Alkylation with dienes forms alkenyl phenols or cyclic compounds or both. Alkyl halides or dialkyl sulfates form phenolic ethers readily by reaction with alkali metal phenoxides. Phenol esters result from reactions of phenols with acid chlorides or anhydrides.

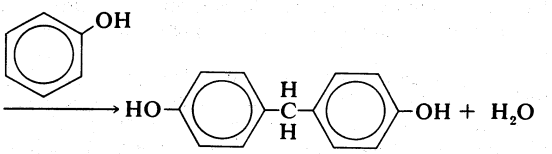
As a simple derivative of benzene, phenol can be readily chlorinated, nitrated, and sulfonated. It can be oxidized to give di- and trihydric phenols, derivatives of diphenols, and various decomposition products. Catalytic hydrogenation produces cyclohexanol and cyclohexane.

Phenol–formaldehyde resins

The most important reaction of phenol is its condensation with formaldehyde to form useful synthetic polymers, the phenol–formaldehyde resins, which are widely used to make articles of plastic. Under alkaline conditions, phenol reacts with formaldehyde to form methylolphenols and methylenediphenol.



phenol formaldehyde methylolphenol



methylenediphenol

Antiseptic properties

Biological and medical properties. The antiseptic activity of phenol is a property that accounts for much of its industrial utilization. In dilute solutions (of 1 to 2 percent) it also finds application as an agent for combatting itching of the skin. Undiluted phenol is highly corrosive to mucous membranes and skin, and it is considered a nerve poison. It may enter the human system by many methods, including absorption by skin, oral ingestion, and vapour inhalation. Phenol and the three cresol isomers are of about the same order of toxicity and produce identical symptoms in poisoned animals.

The lethal dose of phenol for adults has been estimated at 10 to 15 grams orally and less when absorbed from body cavities or mucous membranes. Systemic toxicity of ingested phenol depends on the absolute amount consumed. Continuous exposure even to low levels may cause a chronic poisoning. A safe allowed concentration of phenol vapour in air for an eight-hour exposure is 5 parts per 1,000,000. The metabolic effects of phenol are not well known.

Industrial applications. In the 19th century, phenol was used extensively as an antiseptic, and the term "phenol coefficient" is widely used as a measure of the relative strength of disinfectants. Due to its irritating effect on humans, the use of phenol has been largely discontinued for general antisepsis. Higher-molecular-weight homologues and derivatives of phenol have supplanted phenol itself in the areas of germicidal cleaning.

World War I created a large demand for phenol as a starting material for production of a trinitro derivative, picric acid, which was used extensively as an explosive. In the 1930s the expanding plastics industry began to require large amounts of phenol for the Bakelite-type resins, and this usage expanded to the point in the 1950s where it consumed about 70 percent of total phenol production. In the 1960s this use fell to about 50 percent.

The chemical and pharmaceutical industries are large users of phenol for conversion to many different products. The salicylates (carboxyphenols) are starting materials for the preparation of aspirin, dyes, and flavours. Chlorophenols and their derivatives find application as fungicides, bactericides, and selective weed killers. Alkyl phenols produce an important group of synthetic detergents and are also used as oxidation inhibitors. Phenolsulfonic acid can be converted to synthetic tanning agents, and triphenyl phosphate is a plasticizer for cellulose acetate film. Other large areas of application are as starting materials for production of cyclohexanol, which is converted eventually to nylon, and bisphenol-A, used as a solvent and in the production of epoxy resins.

Identification and analysis. Natural phenol may contain small amounts of moisture, oils, tars, and naphthalene and substantial amounts of *o*-cresol; it normally assays in the range of 82 to 99 percent. Synthetic phenol, on the other hand, runs above 98 percent, with better grades assaying over 99.5 percent. The moisture content of phenol samples can be determined with a special reagent known as the Karl Fischer reagent. When a known impurity is present in phenol, depression of the freezing point provides a method of determining the amount.

Purification by means of acidity

Since phenols are weak acids, treatment with aqueous sodium hydroxide forms water-soluble phenolates, which may then be readily separated from water-insoluble compounds. Acidification regenerates phenol from the phenolate. This is a useful purification method.

Characteristic solid products result from the reaction of phenols with picric acid. The well-defined melting points of these products permit identification of the specific phenols that produce them. In the absence of other phenol compounds, a substitution reaction that yields the 2,4,6-tribromophenol can be used to determine the phenol on the basis of the amount of bromine consumed in the reaction.

Qualitative determination of phenols is readily achieved by the well-known reaction with ferric chloride, which produces highly coloured compounds with the different phenols; phenol itself and resorcinol give violet colours, cresol forms a blue colour, and catechol yields a green-coloured complex. Instrumental methods, especially gas and liquid chromatography and infrared spectrophotometry, are widely used in the detection, identification, and analysis of phenols.

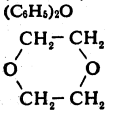
ETHERS

Nomenclature. The name ether apparently was applied to diethyl ether for the first time in 1730 by F.G. Frobenius, who referred to the compound as *spiritus aethereus*, or *vini vitriolatus*. The word ether is derived from the Greek *aithēr* meaning the clear sky or air; it represents the airy or volatile nature of typical compounds of this class. As with the term phenol, ether is used both for a specific compound (diethyl ether) and as a general term; in this case, however, use of the general term in the specific sense has not been officially sanctioned and is discouraged. In common nomenclature of ethers, simple symmetrical compounds with both R groups the same are named for the parent hydrocarbon; thus, $CH_3CH_2OCH_2CH_3$ is called diethyl ether or, simply, ethyl ether. With complex ethers containing different R groups both are named, as methyl ethyl ether for $CH_3OCH_2CH_3$. Alternatively, ethers may be designated by the IUPAC system as hydrocarbon derivations; $CH_3OCH_2CH_2CH_2CH_3$, for example, is called in this system 1-methoxybutane. Trivial or unsystematic names also are used for some complex ethers, such as anisole for methyl phenyl ether ($CH_3OC_6H_5$) and phenetole for ethyl phenyl ether ($CH_3CH_2OC_6H_5$).

General physical properties. At room temperature, ethers are colourless, neutral liquids with pleasant odours. Saturated and unsaturated aliphatic ethers have densities that range from 0.7 to 0.8, whereas the densities of aromatic ethers range from 1.0 to 1.1. Compared to the corresponding alcohols, ethers are generally lower in density, less water soluble, and have lower boiling points. The boiling points of ethers tend to approximate those for hydrocarbons of the same molecular weight, from which it can be concluded that association in liquid form is not significant. Ethers are easily soluble in organic liquids. Dioxane, however, is unusual among ethers in that it is completely soluble in water as well. A large number of organic compounds form azeotropes (constant-distilling mixtures) with ethers. Because of these properties, ethers are extensively used for the extraction and separation of organic chemicals. Physical properties of some common ethers are shown in Table 6.

Solubilities

Table 6: Physical Properties of Ethers

	formula	melting point (°C)	boiling point (°C)	density (d_4^{20})
Methyl ether	$(CH_3)_2O$	−138.5	−23.0	—
Ethyl ether	$(C_2H_5)_2O$	−116.2	34.51	0.7138
Isopropyl ether	$[(CH_3)_2CH]_2O$	−85.89	68.0	0.7241
n-Butyl ether	$(n\text{-}C_4H_9)_2O$	−95.3	142.0	0.7689
Vinyl ether	$(CH_2{=}CH)_2O$	−101	28	0.773
Allyl ether	$(CH_2{=}CHCH_2)_2O$	—	94.0	0.8260
Cyclopropyl methyl ether	$CH_2CH_2CHOCH_3$	−119	44.7	0.8100
Benzyl ether	$(C_6H_5CH_2)_2O$	3.60	298	1.0428
Diphenyl ether	$(C_6H_5)_2O$	26.84	257.93	1.0748
1,4-Dioxane		11.8	101.1	1.0353

Occurrence and synthetic production. Simple ethers are not found commonly in nature, but the ether linkage is present in such natural products as sugars, starches, and cellulose.

Dehydration. Catalytic dehydration of alcohols is the oldest and most important method of producing ethers. Sulfuric acid is the dehydration catalyst most commonly used; it remains behind when water and ether are removed by distillation.

$$ROH + H_2SO_4 \xrightarrow{\text{cold}} RHSO_4 + H_2O$$

$$RHSO_4 + ROH \xrightarrow{\text{heat}} ROR + H_2SO_4$$

Williamson ether synthesis. The reaction of alkyl halides with sodium alcoholates is generally known as the Williamson ether synthesis. Both simple and mixed ethers are made in this manner. Metallic oxides may be used in place of the alcoholate to produce symmetrical ethers.

Reaction of olefins with alcohols or phenols. In the presence of sulfuric acid, alcohols or phenols combine with olefins to produce mixed ethers.

$$(CH_3)_2C{=}CH_2 + ROH \xrightarrow{H_2SO_4} (CH_3)_3COR$$

Reaction of boron trifluoride alcoholates. Heating of boron trifluoride alcoholates under pressure yields simple ethers.

$$n\text{-}C_4H_9OH \cdot BF_3 \xrightarrow[\text{autoclave}]{\text{benzene}} (n\text{-}C_4H_9)_2O$$

dibutyl ether

Use of diazomethane. Methyl ethers can be made from alcohols by reaction with the methylating agent, diazomethane.

Principal chemical reactions. Ethers are relatively inert with regard to chemical reaction, and in this regard they resemble the corresponding alkanes (which carry no functional groups). Since neither alkali metals nor alkalies attack ethers at low temperatures, they may be dried over metallic sodium.

Autoxidation. In the presence of oxygen, ethers oxidize to unstable peroxides, and this reaction may create the danger of explosion in stored ethers.

Salt formation. Strong inorganic acids and Lewis acids (compounds capable of accepting electron pairs) form oxonium compounds with ethers.

Cleavage reactions. Heat decomposes ethers, especially in the presence of alumina catalysts, to form olefins and water as the principal products.

Strong acids, such as hydriodic and nitric acid, cleave one or both of the carbon–oxygen linkages to form alkyl derivatives, and, in those cases in which only one bond is cleaved, alcohols.

Complex formation. Ether is used extensively as a solvent for Grignard reagents because it forms soluble complexes with most alkylmagnesium halides. It also forms complexes with inorganic halides, boron trifluoride, lithium aluminum hydride, and lithium borohydride.

Biological and medical properties. Ether has a widespread use as a surgical anesthetic; it acts on the central nervous system to render the patient insensible. Administration is by inhalation, rapid absorption taking place from the lungs. Ether is not a severe skin irritant, but it may cause slight irritation when swallowed or applied rectally. Ether is rapidly excreted from the lungs; small amounts also are excreted in the urine. The compound undergoes no chemical change in the human body. Concentrations of ethyl ether of more than 400 parts per 1,000,000 in air are considered dangerous when applied by inhalation. Ingestion of more than a fluid ounce may cause death.

Use as an anesthetic

Isopropyl and vinyl ethers are less effective anesthetics than ethyl ether, and the isopropyl compound is 1.5 to 2.0 times more toxic than ethyl ether. *n*-Butyl ether is more toxic than isopropyl ether by inhalation but less so by oral administration. Irritation of mucous membranes by the butyl derivative is greater than by isopropyl ether.

Industrial applications. A major application of ethers is as solvents for fats, oils, waxes, perfumes, resins, dyes, gums, and hydrocarbons. Since ethers are insoluble in water, they function well as extracting agents for organic compounds in such uses as separating acetic acid from dilute solutions. The solvent properties of mixtures of ethyl ether and alcohol with respect to cellulose nitrate are enhanced, with the result that this solvent mixture is used in the manufacture of collodion films and pyroxylin plastics, both of which are based on cellulose nitrate.

Use as solvents

In the chemical laboratory and to a certain extent in the chemical industry, ethers are used as anhydrous, inert media for the preparation of organic compounds by such special reactions as the Grignard and Wurtz–Fittig reactions. By changing the alkyl group of the ether, solvents with a wide range of boiling points and densities are obtained, thereby providing materials with specific physical properties for specific uses. Special precautions to prevent fires and explosions are always necessary.

Vapours of specific ethers, such as bis(2-chloroethyl)-ether, are toxic to insects, and these substances are therefore employed as insecticides, miticides, and soil fumigants. Ethyl ether is used to denature several alcohol formulations, making the product unsuitable for human consumption. Ethyl ether also is employed as a starter for diesel engines and as an entrainment agent (a substance that carries water with it on distillation) in the dehydration of ethyl and isopropyl alcohol.

Benzyl ether is a plasticizer for cellulose nitrate and acetate, improving their properties for use as plastics. Isopropyl ether is a good solvent for antiknock compounds, and it serves as a high-octane blending agent for gasoline.

Identification and analysis. The solubility of ethers in cold concentrated sulfuric acid is used to distinguish these compounds from the paraffinic hydrocarbons and the alkyl halides. The so-called Zeisel method, which cleaves ethers with hydriodic acid, measures the methoxy and ethoxy groups when the other hydrocarbon chain of the ether is higher than ethyl and is not tertiary.

Ethyl ether is the most important member of the ether family, and many of the methods used for its analysis are applicable to other ethers.

Ethyl ether is commercially available in various specified grades: anesthetic, absolute (for use as a chemical reagent), industrial, solvent, and synthetic. Analytical methods have been developed to measure small amounts of impurities, especially peroxides.

BIBLIOGRAPHY

Alcohols: UNION CARBIDE CORP., *Alcohols* (1969); R.L. BENT, "Alcohols," in *Encyclopedia of Chemical Technology*, 2nd ed., vol. 1, pp. 531–541 (1963); J.A. MONICK, *Alcohols* (1968); E.H. RODD (ed.), *Chemistry of Carbon Compounds*, 2nd ed., vol. 1A–2D (1965–70).

Phenols: L.F. and M. FIESER, *Advanced Organic Chemistry*, ch. 22 (1961); F.M. KONRAD, "Phenols," in *Encyclopedia of Polymer Science and Technology*, vol. 10, pp. 73–92 (1969); S.J. LEDERMAN and N. POFFENBERGER, "Phenol," in *Encyclopedia of Chemical Technology*, 2nd ed., vol. 15, pp. 147–159 (1968); E.H. RODD (ed.), *Chemistry of Carbon Compounds*, vol. 3, pt. A, ch. 8 (1954).

Ethers: UNION CARBIDE CORP., *Ethers* (1960); L.F. and M. FIESER, *Advanced Organic Chemistry*, pp. 303–314 (1961); A.P. LURIE, "Ethers," in *Encyclopedia of Chemical Technology*, 2nd ed., vol. 8, pp. 470–498 (1965).

(J.A.Mo.)

Aldehydes and Ketones

Aldehydes and ketones are families of similar organic chemical compounds, the molecules of which contain a special grouping of atoms called the carbonyl group. Largely because of the chemical properties of this group, aldehydes and ketones have many important industrial uses. Acetone, for example, is a solvent commonly employed in vast quantities in the production of paints, lacquers, and other substances. Acetaldehyde, like many other aldehydes and ketones, is a chemical intermediate, a substance converted by chemical reactions into other compounds widely employed in the production of plastics, dyes, and pharmaceuticals. In biological systems, many important compounds, such as the sugars and certain of the steroid hormones, contain aldehyde or ketone groups.

Industrial uses

GENERAL CONSIDERATIONS AND PROPERTIES

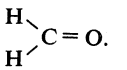
The carbonyl group, by which aldehydes and ketones are identified, consists of a carbon atom joined by two of its four available bonds to the two available bonds of an oxygen atom, and by the remaining two bonds to other carbon or hydrogen atoms. The molecular structure of the carbonyl group is written in chemical symbols as



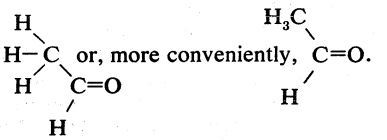
in which the letters C and O are the symbols for atoms of carbon and oxygen, respectively, and the lines represent chemical bonds between the atoms. As this partial formula indicates, the oxygen atom forms two chemical bonds (or in chemical terminology has a valence of two), and the carbon atom forms four bonds (valence, four), two of the valence bonds of carbon being utilized in forming the carbonyl bond with oxygen and two being available for joining to other atoms in the molecules of aldehydes and ketones.

Aldehydes are distinguished from ketones in that they have at least one hydrogen atom (symbol H, valence one) joined directly to the carbon of the carbonyl group. The simplest aldehyde is formaldehyde, which is used in tanning, preserving, and embalming, and is a constituent of certain plastics. In its molecules two atoms of hydrogen are attached to the carbonyl carbon; its structural formula, therefore, is

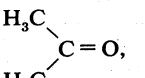
Molecular structures



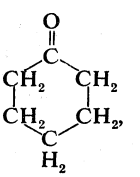
The next most complex aldehyde is acetaldehyde, in the molecule of which one carbon atom is joined to the carbonyl carbon, and the four valences of the second carbon are completed by joining three hydrogen atoms to it. The result is a molecule the structure of which can be designated by the formula



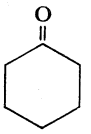
The CH_3 unit or group is called a methyl group. When two methyl groups are joined to the carbonyl carbon the result is acetone,



but acetone is a ketone, not an aldehyde, because there is no hydrogen atom attached directly to the carbonyl carbon. Some ketones with higher molecular weight have structures in which the carbonyl group is present in a carbon chain, the ends of which are joined to form a closed ring, each carbon having enough hydrogens joined to it to complete carbon's four valence bonds. Compounds in which the chain of carbon atoms forms a ring are called ring or cyclic compounds (which does not mean recurrent but merely that the molecule has a ring structure). The compound called cyclohexanone, for example, is a cyclic ketone, its molecules consisting of a ring of six carbon atoms, one of which carries the doubly bonded oxygen of the carbonyl group. The structural formula of cyclohexanone is



or, in a notation customarily used by chemists,



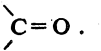
In this abbreviated form each angle of the polygon is understood to represent a carbon atom together with enough hydrogen atoms to complete the valence of the carbon atom.

Aldehydes and ketones are so similar in their chemical and physical properties that they are generally classified together and often they are termed, collectively, carbonyl compounds, since it is the chemistry of the carbonyl group that identifies them. Each individual aldehyde or ketone has properties of its own, and all the members of both families readily undergo a wide variety of chemical reactions (a property which makes them highly suitable for use in chemical synthesis, such as the production of medicinal agents). Aldehydes generally are more chemically reactive than ketones. The simpler ones, such as formaldehyde and acetaldehyde, readily combine with themselves in repeated linkages to produce enormously long chains, containing tens of thousands of the simple molecule. Such macromolecules, known as polymers, can also be produced from the combination of aldehydes (especially formaldehyde) with other types of molecules. Many of these polymers are the familiar plastics of commerce; *e.g.*, Bakelite. Although some aldehydes and ketones occur in nature, including—in small amounts—acetone and acetaldehyde, most carbonyl compounds of commercial importance are prepared synthetically.

Function of the carbonyl group

The properties and behaviour of aldehydes and ketones are explained in large part by the nature of the carbonyl group. The addition of oxygen atoms to any element or compound is called oxidation; the removal of oxygen is called reduction. In a general definition of these reactions, oxidation means the removal of electrons from an atom or molecule, while reduction is the addition of electrons to an atom or molecule. Thus, an oxidizing agent removes electrons and, in the process of adding them to itself, becomes reduced. Similarly, a reducing agent becomes oxidized, since it donates electrons to the substance which it reduces. Actually, in many oxidation and reduction reactions, electrons are not physically removed or added but a compound is formed between the reactants. For example, if B oxidizes A some of A's electrons will also be held by B in the resulting molecule AB. An element or molecule may have more than one electron removed, either in a single oxidation process or in several steps. In the latter case, the substance will have more than one oxidation state, or number. The most completely oxidized state of the carbon atom is found in carbon dioxide, CO_2, which means that the carbon atom has all four of its valence electrons attached to the two oxygen atoms, bonding all three atoms together. The fully reduced state of carbon occurs in methane (CH_4), in which the four electrons of the hydrogen atoms are held by the carbon. The carbonyl group thus represents a partially oxidized state, and aldehydes and ketones are subject to either further oxidation or further reduction—a situation that accounts in part for their great chemical reactivity and for the wide variety of chemical reactions in which they take part.

Electronic structure of the carbonyl group. The chemical bonds that hold the nuclei of atoms together in molecular structures are the result of the motion of electrons around and between the constituent atoms. Thus, any group of atoms can be considered in terms of the electronic configuration that is associated with the bonded nuclei. Because it requires energy to disrupt any chemical bond, and because different kinds of bonds have different strengths, chemical reactions often can be predicted on the basis of known bond strengths. The electronic structure of the carbonyl group will clearly influence the physical and chemical characteristics of aldehydes and ketones. The carbonyl group may be pictured in terms of molecular orbitals that represent the "standing wave forms" of the electrons that are associated with particular bonds, in this case the carbon–oxygen double bond. The electronic structure of the carbonyl group can also be described in an alternate but chemically equivalent fashion as a resonance hybrid; *i.e.*, a combination of two structures, neither of which alone adequately depicts the molecule but which, taken together, represent its true structure. One of the resonance forms of the carbonyl group is the doubly bonded form already described:

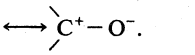


The other is one in which a positive charge appears on the carbon atom and a negative charge on the oxygen atom:

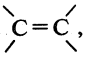


Hybrid state

The hybrid state of the actual carbonyl group is indicated in symbols by showing both forms connected by a double-headed arrow, as



The presence of plus and minus charges in one of the resonance structures suggests that the carbonyl group is highly polar (that is, that it shows an uneven distribution of electrons that makes the oxygen electrically negative and the carbon electrically positive). In contrast to a carbon–carbon double bond



which is essentially nonpolar and weaker than should be expected from the combination of two carbon–carbon single bonds, the carbon–oxygen double bond is strongly polar, possessing more than twice the bond energy of a carbon–oxygen single bond.

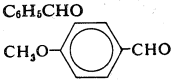
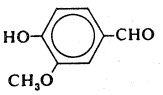
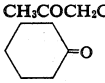
Physical properties. The polarity of the carbon–oxygen double bond is responsible for many of the physical properties of carbonyl compounds, including their behaviour in electrical fields, called their dipole moment. The dipole moment of a substance is a measure of the uneven distribution of electrons that produces an asymmetry of electrical charges in the molecule. The highly polarized character of the carbonyl group introduces a certain degree of attraction between the molecules of carbonyl compounds; *i.e.*, since the oxygen of each carbonyl molecule carries more negative charge than the carbon atom, it will be attracted to the carbon end of other carbonyl molecules. This attraction in turn affects the melting points and boiling points of the compounds, both of which properties depend on the ease of separability of the molecules. Thus, the melting points of the simpler aldehydes and ketones are from 50° to 80° C higher than those of hydrocarbons, compounds of comparable size, shape, and molecular weight but lacking these polar structural units. Although formaldehyde is a gas at ordinary temperatures, other low-molecular-weight carbonyl compounds are volatile liquids. The water solubility of simple carbonyl compounds is considerably enhanced because of the polar nature of the carbon–oxygen double bond (polar compounds generally being more water soluble than nonpolar compounds; see the Table for boiling points, melting points, and solubilities of common aldehydes and ketones).

Chemical properties. The chemical properties, too, of aldehydes and ketones are largely determined by the particular electronic character of the carbonyl group. Thus, substances attracted to positive charges (called nucleophilic reagents because they contain electron-rich atoms) generally attack the carbonyl carbon, which bears a partial positive charge resulting from the contribution of the charged resonance form

Nucleophilic sources

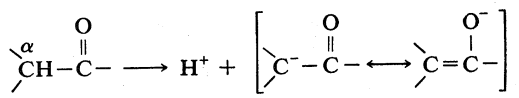


Aldehydes and Ketones

carbonyl compound	common name	systematic name	C=O infrared absorption frequency (cm^{-1})	melting point (°C)	boiling point (°C)	water solubility	industrial uses
Aliphatic aldehydes							
HCHO	formaldehyde	methanal	1743*	−92	−21	very soluble	monomer for polymers (*e.g.*, Delrin), explosives (*e.g.*, Cyclonite)
CH_3CHO	acetaldehyde	ethanal	1752*	−121	21	soluble in all proportions	synthesis of acetic acid, acetic anhydride, ethyl acetate, acrylonitrile, crotonaldehyde, 1-butanol
Cl_3CCHO	chloral	trichloroethanal	1762†	−58	98	very soluble	synthesis of DDT
Aromatic aldehydes							
C_6H_5CHO	benzaldehyde	benzaldehyde	1704†	−26	178	slightly soluble	chemical intermediate, flavouring substance
	anisaldehyde	4-methoxybenzaldehyde	1700†	0	250	slightly soluble	perfume ingredient
	vanillin	4-hydroxy-3-methoxy-benzaldehyde	1700†	77–79	285	slightly soluble	flavouring substance
Aliphatic ketones							
CH_3COCH_3	acetone	2-propanone	1742*, 1715†	−95	56	soluble in all proportions	solvent, synthesis of mesityl oxide and methyl methacrylate
$CH_3COCH_2CH(CH_3)_2$	methyl isobutyl ketone	4-methyl-2-pentanone	1720†	−85	117	slightly soluble	solvent
	cyclohexanone	cyclohexanone	1710†	−45	155	slightly soluble	synthesis of monomer for polymers (*e.g.*, Nylon-6)
Aromatic ketones							
$C_6H_5COCH_3$	acetophenone	acetophenone	1686†	21	202	insoluble	chemical intermediate, perfume ingredient
$C_6H_5COC_6H_5$	benzophenone	benzophenone	1662†	48	306	insoluble	chemical intermediate
Unsaturated aldehydes and ketones							
$CH_3CH=CHCHO$	crotonaldehyde	2-butenal	1685†	−74	104–5	very soluble	synthesis of 1-butanol
$(CH_3)_2C=CHOCH_3$	mesityl oxide	4-methyl-3-penten-2-one	1690†	−53	130	soluble	synthesis of methyl isobutyl ketone

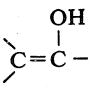
*In vapour phase. †In solution.

Hydrogen atoms on the carbon atom immediately adjacent to the carbonyl, the so-called α-carbon, are easily removed from the molecule because the resulting negatively charged group (called an anion) is a resonance hybrid (see above) of two forms, a situation that provides extra stability to the structure.



carbonyl compound with α-hydrogen atom — proton — resonance hybrid

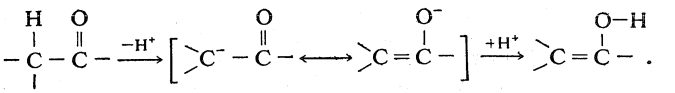
The effect of the resonance stabilization of the anion is to cause the α-hydrogen (hydrogen attached to the α-carbon) to be easily lost as a proton, that is, a transferable hydrogen nucleus represented by the symbol H^+, also called a hydrogen ion; in other words, the resonance stabilization makes the α-hydrogen acidic. A further consequence of the ease of loss of α-hydrogens is that carbonyl compounds bearing such hydrogens are easily converted to their so-called enol



forms (the name being derived from a combination of -ene, the ending for double-bond compounds,



and -ol, the ending for hydroxyl compounds, —OH). The interconversion usually occurs through the resonance-stabilized anion (see above). The reaction is represented in equation form in the following manner:



Generally, aldehydes and ketones enter into a wide variety of chemical reactions involving the carbonyl group itself or the α-carbons (see below for specific examples). Aldehydes typically are more reactive than ketones, in part because the carbonyl carbon in aldehydes is more open to attacking groups because the hydrogen atom attached to it is less bulky than the carbon-containing groups attached to it in ketones, and this can be seen readily by comparing the formulas:



aldehyde — ketone

Nomenclature. The naming of organic compounds is complicated by the fact that any chain of carbon atoms may have branches or rings attached to it. When a dominant group, such as the carbonyl, is present, the basic design of the nomenclature is to chart complicated structures by reference to the dominant group. Many aldehydes and ketones are referred to by common names, and in such usage aldehydes are named by analogy with the carboxylic acids to which they are related; thus, formaldehyde from formic acid, acetaldehyde from acetic acid. Ketones, however, are commonly designated by the groups (called substituent groups) attached to the carbonyl carbon, along with the appellation ketone; *e.g.*, methyl ethyl ketone for the compound $CH_3COCH_2CH_3$, in which a methyl (CH_3) and an ethyl (C_2H_5) are the substituent groups, both being joined to the carbonyl carbon. In formal, or systematic, nomenclature of aldehydes and ketones, the longest chain carrying the carbonyl group is

Common names and systematic names

specified as the parent structure, and the name of the compound is derived from the name of the parent structure by use of the suffix -al if the compound is an aldehyde (propanal, CH_3CH_2CHO—common name, propionaldehyde from propionic acid) or -one if it is a ketone (propanone, CH_3COCH_3—common name, acetone). To indicate the location on the parent chain of any carbon atom, in order to specify where a substituent group may be attached, the carbon atoms are numbered, beginning with the aldehyde carbon. With the ketones, because the carbonyl atom may be anywhere in the chain, the numbering begins at that end of the chain that gives the ketone carbon the smallest possible number:

$$\overset{4}{C}H_3\overset{3}{C}H_2\overset{2}{C}O\overset{1}{C}H_3, \text{ not } \overset{1}{C}H_3\overset{2}{C}H_2\overset{3}{C}O\overset{4}{C}H_3.$$

Designation of substituents

In common terminology, the position of substituents is indicated by Greek letters, designating the carbon immediately adjacent to the carbonyl group with the Greek letter α, the second with β, and so on. Examples of these usages are: $ClCH_2CH_2CHO$, named 3-chloropropanal, or, commonly, β-chloropropionaldehyde; and

$$CH_3CHCOCH_3,$$
$$\quad\;\;|$$
$$\quad CH_3$$

3-methyl-2-butanone or methyl isopropyl ketone. (A list of representative aldehydes and ketones of various types, with their common and systematic names, properties, and uses, is included in the Table.)

Principal methods of synthesis. Aldehydes and ketones can be synthesized by a wide variety of methods, some being generally useful for both classes of compounds and others being more specific in their application. These methods can be generally categorized as involving: (1) oxidation with cleavage of carbon–carbon bonds, (2) oxidation without such cleavage, (3) reduction (without cleavage), and (4) formation of new carbon–carbon bonds.

PRINCIPAL CHEMICAL REACTIONS

Further discussion of aldehydes and ketones requires a knowledge of electronic structure, chemical reactions, and vocabulary that cannot be provided in this article. The necessary background, however, can be found in detail in the articles ATOMIC STRUCTURE; CHEMICAL ELEMENTS; PERIODIC LAW; CHEMICAL COMPOUNDS, ORGANIC; and CHEMICAL REACTIONS. The polarity of the carbonyl group, which reduces the electron density both at the carbonyl carbon,

$$\text{>C=O} \longleftrightarrow \text{>C}^+-\text{O}^-,$$

and at the α-hydrogen,

$$-\overset{H}{\underset{|}{C}}-C=O \longleftrightarrow \overset{H^+}{\text{>C}}=C-O^-,$$

promotes easy reaction with nucleophilic agents (groups that attack positively charged centres) and enhances the acidity of the α-proton. Two of the characteristic reactions of aldehydes and ketones, therefore, are addition processes in which a nucleophile adds to the carbonyl carbon—

Addition and displacement reactions

$$R_2C=O + Nu^- \longrightarrow R_2\overset{O^-}{\underset{Nu}{C}},$$

in which R is any hydrocarbon chain and Nu any nucleophilic agent—and displacement processes wherein the carbonyl compound loses an α-hydrogen and itself becomes a nucleophile—$RCH_2COR \rightleftharpoons R^-CHCOR + H^+$.

Major categories of reactions. *Addition of noncarbon nucleophiles to the carbonyl carbon.* The reactivity of nucleophiles based on atoms other than carbon (heteroatom nucleophiles) toward $R_2C=O$ increases according to the following sequence: halogen < oxygen < nitrogen < sulfur. Thus, the very weakly nucleophilic halide anions have virtually no tendency to become joined to aldehydes and ketones, and water molecules are added only to a few especially reactive ones. Acetone, for example, is hydrated only to a limited extent in aqueous solution, but hexafluoroacetone, strongly activated by the electron-withdrawing fluorine atoms, is almost completely converted to its hydrate under comparable conditions. Alcohols (oxygen compounds), amines (nitrogen compounds), and mercaptans (sulfur compounds), on the other hand, are added more effectively than water. The initial addition of such compounds is often followed by a dehydration step that may be either intramolecular (within the molecule) or intermolecular (between molecules). Intermolecular dehydration is observed following the acid-catalyzed addition of alcohols to ketones; the initially formed product, termed a hemiketal, reacts, via dehydration, with a second molecule of alcohol to give a second product called a ketal.

Addition of carbon nucleophiles to the carbonyl carbon. Carbon nucleophiles are especially reactive toward aldehydes and ketones, and this type of reaction forms the basis for many important synthetic methods. Among the carbon nucleophiles most frequently used for this purpose are hydrogen cyanide (HCN), Grignard reagents (RMgX, in which R is a hydrocarbon, Mg a magnesium atom, and X a halogen, usually chlorine or bromine), and other organometallic compounds, active methylene nucleophiles, including the nitroalkanes (RNO_2), and such highly polarizable substances as diazomethane (CH_2N_2) and certain ylides.

Comparison with carboxylic acid derivatives

A feature that generally differentiates aldehydes and ketones from carboxylic acid derivatives is the tendency of the latter to undergo nucleophilic displacement rather than addition; *i.e.*,

$$\overset{R}{\underset{Y}{\text{>C=O}}} + NuH \longrightarrow \overset{R}{\underset{Nu}{\text{>C=O}}} + YH; \text{ not } R-\overset{OH}{\underset{Y}{C}}-Nu.$$

Alpha-trihalo ketones ($RCOCX_3$) and β-diketones ($RCOCH_2COR$), however, resemble carboxylic acid derivatives in this respect, the driving force for the displacement in these cases resulting from the special stability of the groups displaced, the $-CX_3$ and $-CH_2COR$ anions.

Displacement at the α-carbon. The capacity of the carbonyl group to attract and stabilize a negative charge facilitates the displacement of hydrogens attached to the α-carbon. The resulting resonance-stabilized (see above) anion

$$\text{>}\overset{O}{\overset{\|}{C}}-\overset{-}{C}- \longleftrightarrow \text{>}\overset{O^-}{\overset{|}{C}}=C-$$

acts as a carbon nucleophile in a variety of reactions. Among others, it can react with: (1) molecular bromine, to give α-bromo compounds; (2) alkyl halides, to give α-alkylated compounds; and (3) other carbonyl compounds to give β-hydroxy compounds that may lose water to form unsaturated substances.

Conjugated systems

Conjugate addition. Electronic effects can be transmitted within organic molecules by way of conjugated systems, which are series of alternate single and double bonds between adjacent carbon atoms or between carbon atoms and atoms of other elements, such as oxygen. (Conjugated systems contain two or more directly-attached pi bonds—bonds in which the molecular orbitals are unsymmetrical around the bond axis.) Because aldehydes and ketones with a double bond between the α-carbon and the β-carbon (α–β-unsaturated carbonyl compounds) are conjugated systems, they exhibit this transmission of electronic effect. As a result, they exist as resonance hybrids (see above) in which one of the forms has a positive charge on the β-position rather than on the carbonyl carbon

$$\text{>C=C-C=O} \longleftrightarrow \text{>C=C-C}^+-O^- \longleftrightarrow \text{>C}^+-C=C-O^-.$$

As a result, nucleophiles may attack the β-carbon rather than the carbonyl carbon, leading to addition across the

conjugated system, a process generally referred to as conjugate addition.

Oxidation–reduction reactions. Aldehydes and ketones are capable of gaining or losing electrons (*i.e.*, of undergoing, respectively, reduction or oxidation). Nucleophilic additions to aldehydes and ketones actually are examples of reduction, as is most explicitly shown when the hydride ion is the nucleophile (lithium aluminum hydride, sodium borohydride, and other metal hydrides are sources of the hydride ion) reducing aldehydes to primary alcohols ($RCHO \rightarrow RCH_2OH$) and ketones to secondary alcohols ($R_2C = O \rightarrow R_2CHOH$). When certain metals, such as magnesium, are used as reducing agents, the conversion of carbonyl compounds to dimeric (incorporating two molecules of the starting material) products may take place, a process called bimolecular reduction. Clemmensen reduction, one of the several methods available, employs amalgamated zinc (zinc–mercury alloys) in the presence of hydrochloric acid as the reducing medium. Another method consists of the conversion of the starting compound to a thioacetal or thioketal (sulfur-containing derivatives) followed by hydrogenolysis (cleavage with hydrogen) in the presence of a catalyst: $R_2C = O \rightarrow R_2C(SC_2H_5)_2 \rightarrow R_2CH_2 + 2C_2H_5SH$.

Clemmensen reduction

Aldehydes are considerably more susceptible than ketones to oxidation, since they are readily oxidized to carboxylic acids under conditions that have little effect on ketones. Under more strenuous conditions (such as heating with strong oxidizing agents), however, ketones can also be oxidized, by a process involving cleavage of carbon–carbon bonds associated with the carbonyl group. Thus, the action of hot nitric acid on cyclohexanone (see above and the Table) cleaves the cyclohexane ring and gives the dicarboxylic acid, adipic acid, a substance used in the preparation of nylon. In the presence of strong base, an aldehyde lacking any α-hydrogen atoms (*e.g.*, HCHO, C_6H_5CHO, and R_3CCHO) undergoes a self-oxidation-reduction (the Cannizzaro reaction); treatment of benzaldehyde with strong aqueous sodium hydroxide yields a mixture of sodium benzoate, an oxidation product, and benzyl alcohol, a reduction product.

Polyfunctional aldehydes and ketones. The presence of additional functional groups within the molecule greatly influences the properties and behaviour of aldehydes and ketones. Such compounds, for example, frequently show altered infrared absorption characteristics (light absorption properties indicative of the electronic state of the carbonyl group), as shown in the Table. Conjugate addition (see above) is a direct consequence of the interaction of a carbonyl function with an α,β-unsaturated group within the molecule. Furthermore, carbonyl groups frequently are found to undergo direct chemical reaction with other functional groups within the same molecule. In many sugars, intramolecular addition of a hydroxyl group to the aldehyde or ketone function is known to occur. Glucose, for example, exists almost entirely in a cyclic, or ring, form.

Similarly, in polycarbonyl compounds the effect of one carbonyl function upon another may frequently be observed. The enhanced reactivity of β-diketones toward substitution is a good example of such mutual influence. In some dicarbonyl compounds, an internal chemical reaction between the various portions of the molecule occurs. For example, intramolecular condensation is observed during the so-called Robinson procedure for preparing six-membered rings of carbon atoms, a method frequently employed in synthetic processes.

Effect of one carbonyl on another

Synthesis with aldehydes and ketones. Because of (1) their ease of preparation by many different methods; (2) their relative stability and convenience; (3) their ability to react with protective groups, which insure their unreactivity during other chemical manipulations; and (4) their capacity for undergoing chemical reactions of almost unlimited variety, the aldehydes and ketones are nearly ideal chemical intermediates. There are, in fact, few multistep synthetic procedures that do not involve aldehydes or ketones at some stage in the process. To cite one example among many, a laboratory synthesis of longifolene—a hydrocarbon obtained from certain turpentines, involves carbonyl chemistry at nearly every stage of the process.

IDENTIFICATION AND USES

Analysis of aldehydes and ketones. Aldehydes and ketones as a group are readily characterized by their ability to form derivatives such as phenylhydrazones, oximes, and semicarbazones (see above). It was the use of phenylhydrazine to form characteristic crystalline derivatives of the sugars that permitted chemists to unravel many of the complexities of carbohydrate chemistry in the latter part of the 19th century. Characteristic, strong absorption bands in their infrared spectra (patterns of infrared light absorption) near 1720 cm^{-1} (reciprocal centimetres, a standard unit for light absorption measurements) also are useful in identifying aldehydes and ketones. Various substituents in the neighbourhood of the carbonyl group cause shifts in the absorption band that may be used to help identify the substituents. For example, as shown in the Table, aromatic and olefinic groups adjacent to the carbonyl function cause the absorption due to the carbonyl bond to move to the neighbourhood of 1680 cm^{-1}. Aldehydes are specifically characterized by their easy oxidation to carboxylic acids by such mild oxidants as silver ion (Tollens reagent) and cupric ion (Fehling reagent), neither of which has an appreciable effect on ketones. Nuclear-magnetic-resonance spectroscopy, a technique that measures the intensity of magnetic fields of atomic nuclei, also is useful in identifying aldehydes; in the spectra obtained with this technique, a characteristic band is found with aldehydes.

Characterization of aldehydes

Aldehydes and ketones of commercial, physiological, and medicinal importance. A number of aldehydes and ketones are important industrial compounds and are manufactured on a large scale, in millions of tons. These and other aldehydes and ketones (see Table) are useful industrially as solvents, polymer components, flavouring agents, perfume ingredients, and chemical intermediates. Many compounds with important physiological properties contain carbonyl functions. Among those possessing an aldehyde group are retinene and pyridoxal phosphate. Retinene, obtained by oxidation of vitamin A_1, combines with a protein to form a pigment called rhodopsin, or visual purple, which plays a key role in the phenomenon of human vision. Pyridoxal phosphate is a coenzyme that, in association with a variety of enzymes, catalyzes certain essential life processes. The sugars are another important class of compounds in which either an aldehyde or ketone function is present. Among the common sugars are glucose (an aldohexose) and fructose (a ketohexose). Many compounds of medicinal value contain aldehyde or ketone functions. Many natural and synthetic steroid hormones bear aldehyde or ketone groupings. Molecules of the anti-inflammatory agent cortisone have three ketone groups, and aldosterone, a salt-retaining hormone, has two ketone groups and an aldehyde group.

Physiologically active carbonyl compounds

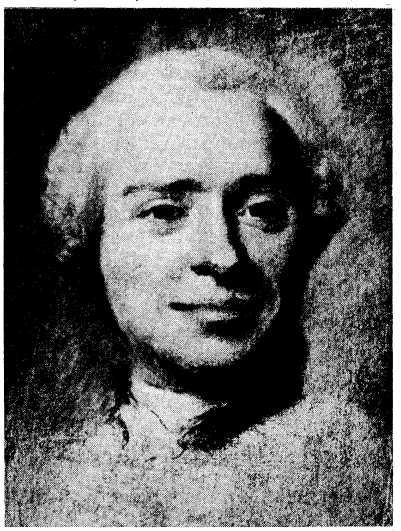
BIBLIOGRAPHY. C.D. GUTSCHE, *The Chemistry of Carbonyl Compounds* (1967), a short treatise, written for the student in undergraduate organic chemistry, covering the structure, the properties, the mechanism of reactions, and the synthetic utility of carbonyl compounds; C.R. NOLLER, *Chemistry of Organic Compounds*, 3rd ed. (1965), a textbook, designed for a high-level first-year course in organic chemistry, that includes material relating to aldehydes and ketones; R.C. FUSON, *Reactions of Organic Compounds* (1962), a textbook for the advanced student that includes material relating to aldehydes and ketones; E.S. GOULD, *Mechanism and Structure in Organic Chemistry* (1959); J.S. HINE, *Physical Organic Chemistry*, 2nd ed. (1962); J. MARCH, *Advanced Organic Chemistry: Reactions, Mechanisms, and Structure* (1968); and E.E. ROYALS, "The Chemistry of Carbonyl Compounds," in *Advanced Organic Chemistry*, pt. 2 (1954), advanced textbooks including material relating to the mechanisms of the reactions of carbonyl compounds; W.P. JENCKS, "Carbonyl- and Acyl-group Reactions," in *Catalysis in Chemistry and Enzymology*, pp. 463–554 (1969), a detailed discussion of the mechanism of various types of carbonyl reactions; S. PATAI (ed.), *The Chemistry of the Carbonyl Group* (1966) and *The Chemistry of Carboxylic Acids and Esters* (1969), collections of detailed essays covering various aspects of carbonyl chemistry.

(C.D.G.)

Alembert, Jean Le Rond d'

Jean Le Rond d'Alembert, French mathematician, philosopher, and writer, achieved fame as a mathematician and scientist before acquiring a considerable reputation as a contributor to the famous *Encyclopédie* and as one of the leading figures in the French Enlightenment.

By courtesy of the Bibliotheque Nationale, Paris



D'Alembert, pastel by Maurice-Quentin de La Tour (1704–88).

The illegitimate son of a famous hostess, Mme de Tencin, and one of her lovers, the chevalier Destouches, he was born on November 17, 1717, and abandoned on the steps of the Parisian church of Saint-Jean-le-Rond, from which he derived his Christian names. Although Mme de Tencin never recognized her son, Destouches eventually sought out the child and entrusted him to a glazier's wife, whom d'Alembert always treated as his mother. Through his father's influence, he was admitted to a prestigious Jansenist school, enrolling first as Jean-Baptiste Daremberg and subsequently changing his name, perhaps for reasons of euphony, to d'Alembert. Although Destouches never disclosed his identity as father of the child, he left his son an annuity of 1,200 livres. D'Alembert's teachers at first hoped to train him for theology, being perhaps encouraged by a commentary he wrote on St. Paul's Letter to the Romans, but they inspired in him only a lifelong aversion to the subject. He spent two years studying law and became an advocate in 1738, although he never practiced. After taking up medicine for a year, he finally devoted himself to mathematics—"the only occupation," he said later, "which really interested me." Apart from some private lessons, d'Alembert was almost entirely self-taught.

Mathematics. In 1739 he read his first paper to the Academy of Sciences, of which he became a member in 1741. In 1743, at the age of 26, he published his important *Traité de dynamique*, a fundamental treatise on dynamics containing the famous "d'Alembert's principle," which states that Newton's third law of motion (for every action there is an equal and opposite reaction) is true for bodies that are free to move as well as for bodies rigidly fixed. Other mathematical works followed very rapidly; in 1744 he applied his principle to the theory of equilibrium and motion of fluids, in his *Traité de l'équilibre et du mouvement des fluides.* This discovery was followed by the development of partial differential equations, a branch of the theory of calculus, the first papers on which were published in his *Réflexions sur la cause générale des vents* (1747). It won him a prize at the Berlin Academy, to which he was elected the same year. In 1747 he applied his new calculus to the problem of vibrating chords, in his *Recherches sur les cordes vibrantes;* in 1749 he furnished a method of applying his principles to the motion of any body of a given shape; and in 1749 he found an explanation of the precession of the equinoxes (a gradual change in the position of the Earth's orbit), determined its characteristics, and explained the phenomenon of the nutation (nodding) of the earth's axis, in *Recherches sur la précession des équinoxes et sur la nutation de l'axe de la terre.* In 1752 he published *Essai d'une nouvelle théorie de la résistance des fluides,* an essay containing various original ideas and new observations. In it he considered air as an incompressible elastic fluid composed of small particles and, carrying over from the principles of solid body mechanics the view that resistance is related to loss of momentum on impact of moving bodies, he produced the surprising result that the resistance of the particles was zero. D'Alembert was himself dissatisfied with the result; the conclusion is known as "d'Alembert's paradox" and is not accepted by modern physicists. In the *Memoirs* of the Berlin Academy he published findings of his research on integral calculus—which devises relationships of variables by means of rates of change of their numerical value—a branch of mathematical science that is greatly indebted to him. In his *Recherches sur différens points importans du système du monde* (1754–56) he perfected the solution of the problem of the perturbations (variations of orbit) of the planets that he had presented to the academy some years before. From 1761–80 he published eight volumes of his *Opuscules mathématiques.*

D'Alembert's principle

The Encyclopédie. Meanwhile, d'Alembert began an active social life and frequented well-known salons, where he acquired a considerable reputation as a witty conversationalist and mimic. Like his fellow Philosophes —those thinkers, writters, and scientists who believed in the sovereignty of reason and nature (as opposed to authority and revelation) and rebelled against old dogmas and institutions—he turned to the improvement of society. A rationalist thinker in the free-thinking tradition, he opposed religion and stood for tolerance and free discussion; in politics the Philosophes sought a liberal monarchy with an "enlightened" king who would supplant the old aristocracy with a new, intellectual aristocracy. Believing in man's need to rely on his own powers, they promulgated a new social morality to replace Christian ethics. Science, the only real source of knowledge, had to be popularized for the benefit of the people, and it was in this tradition that he became associated with the *Encyclopédie* about 1746. When the original idea of a translation into French of Ephraim Chambers' English *Cyclopædia* was replaced by that of a new work under the general editorship of the Philosophe Denis Diderot, d'Alembert was made editor of the mathematical and scientific articles. In fact, he not only helped with the general editorship and contributed articles on other subjects but also tried to secure support for the enterprise in influential circles. He wrote the *Discours préliminaire* that introduced the first volume of the work in 1751. This was a remarkable attempt to present a unified view of contemporary knowledge, tracing the development and interrelationship of its various branches and showing how they formed coherent parts of a single structure; the second section of the *Discours* was devoted to the intellectual history of Europe from the time of the Renaissance. In 1752 d'Alembert wrote a preface to Volume III, which was a vigorous rejoinder to the *Encyclopédie's* critics, while an *Éloge de Montesquieu,* which served as the preface to Volume V (1755), skillfully but somewhat disingenuously presented Montesquieu as one of the *Encyclopédie's* supporters. Montesquieu had, in fact, refused an invitation to write the articles "Democracy" and "Despotism," and the promised article on "Taste" remained unfinished at his death in 1755.

The Philosophes

In 1756 d'Alembert went to stay with Voltaire at Geneva, where he also collected information for an *Encyclopédie* article, "Genève," which praised the doctrines and practices of the Genevan pastors. When it appeared in 1757, it aroused angry protests in Geneva because it affirmed that many of the ministers no longer believed in Christ's divinity and also advocated (probably at Voltaire's instigation) the establishment of a theatre. This article prompted Rousseau, who had contributed the ar-

ticles on music to the *Encyclopédie,* to argue in his *Lettre à d'Alembert sur les spectacles* (1758) that the theatre is invariably a corrupting influence. D'Alembert himself replied with an incisive but not unfriendly *Lettre à J.-J. Rousseau, citoyen de Genève.* Gradually discouraged by the growing difficulties of the enterprise, d'Alembert gave up his share of the editorship at the beginning of 1758, thereafter limiting his commitment to the production of mathematical and scientific articles.

Later literary, scientific, and philosophical work. His earlier literary and philosophical activity, however, led to the publication of his *Mélanges de littérature, d'histoire et de philosophie* (1753). This work contained the impressive *Essai sur les gens de lettres,* which exhorted writers to pursue "liberty, truth and poverty" and also urged aristocratic patrons to respect the talents and independence of such writers.

Académie Française

Largely as a result of the persistent campaigning of Mme du Deffand, a prominent hostess to writers and scientists, d'Alembert was elected to the French Academy in 1754; he proved himself to be a zealous member, working hard to enhance the dignity of the institution in the eyes of the public and striving steadfastly for the election of members sympathetic to the cause of the Philosophes. His personal position became even more influential in 1772 when he was made permanent secretary. One of his functions was the continuation of the *Histoire des membres de l'Académie;* this involved writing the biographies of all the members who had died between 1700 and 1772. He paid tribute to his predecessors by means of *Éloges* that were delivered at public sessions of the academy. Though of limited literary value, they throw interesting light on his attitude toward many contemporary problems, and also reveal his desire to establish a link between the Academy and the public.

From 1752 onward, Frederick II of Prussia repeatedly tried to persuade d'Alembert to become president of the Berlin Academy, but the philosopher contented himself with a brief visit to the King at the Rhine village of Wesel in 1755 and a longer stay at Potsdam in 1763. For many years he gave the King advice on the running of the academy and the appointment of new members. In 1762 another monarch, the empress Catherine II of Russia, invited d'Alembert to become tutor to her son, the grand duke Paul; this offer also was refused. Apart from fearing the harmful effects of foreign residence upon his health and personal position, d'Alembert did not wish to be separated from the intellectual life of Paris.

Although as a skeptic, d'Alembert willingly supported the Philosophes' hostility to Christianity, he was too cautious to become openly aggressive. The expulsion of the Jesuits from France, however, prompted him to publish in 1765, at first anonymously, and then in his own name, *Sur la destruction des Jésuites en France, par un auteur désintéressé;* he there tried to show that the Jesuits, in spite of their qualities as scholars and educators, had destroyed themselves through their inordinate love of power.

Music theory

During these years d'Alembert's interests included musical theory. His *Éléments de musique* of 1752 was an attempt to expound the principles of the composer Jean-Philippe Rameau (1683–1764), who had consolidated contemporary musical development into a harmonic system that dominated Western music until about 1900. In 1754 d'Alembert published an essay expressing his thoughts on music in general—and French music in particular—entitled *Réflexions sur la musique en général et sur la musique française en particulier.* He also published in his mathematical *opuscules* treatises on acoustics, the physics of sound, and he contributed several articles on music to the *Encyclopédie.* In 1765 a serious illness compelled him to leave his foster-mother's house, and he eventually went to live in the house of Julie de Lespinasse, with whom he fell in love. He was the leading intellectual figure in her salon, which became an important recruiting centre for the French Academy. Although they may have been intimate for a short time, d'Alembert soon had to be satisfied with the role of steadfast friend. He discovered the extent of her passionate involvement with other men only after Julie's death in 1776. He transferred his home to an apartment at the Louvre—to which he was entitled as secretary to the Academy—where he died on October 29, 1783.

Assessment. Posterity has not confirmed the judgment of those contemporaries who placed d'Alembert's reputation next to Voltaire's. In spite of his original contributions to the mathematical sciences, intellectual timidity prevented his literary and philosophical work from attaining true greatness. Nevertheless, his scientific background enabled him to elaborate a philosophy of science that, though inspired by the rationalist ideal of the ultimate unity of all knowledge, established "principles" making possible the interconnection of the various branches of science. Moreover, d'Alembert was a typical 18th-century Philosophe, for in both his life and his work he tried to invest the name with dignity and serious meaning. In his personal life he was simple and frugal, never seeking wealth and yet dispensing charity whenever possible, always watchful of his integrity and independence, and constantly using his influence, both at home and abroad, to encourage the advance of "enlightenment."

BIBLIOGRAPHY. The main editions of d'Alembert's collected works are: *Oeuvres philosophiques, historiques et littéraires de d'Alembert,* ed. by J.F. BASTIEN, 18 vol. (1805); and *Oeuvres complètes de d'Alembert,* ed. by BOSSANGE and BELIN, 5 vol. (1821, reprinted 1967). These may be supplemented by d'Alembert's *Mélanges de littérature, d'histoire et de philosophie,* 2 vol. (1753), 4 vol. (1759), 5th vol. (1767), which contain a few pieces omitted from the collected editions; his *Histoire des membres de l'Académie française morts depuis 1700 jusqu'en 1771,* 6 vol. (1785–87, reprinted 1970); *Oeuvres posthumes de d'Alembert,* ed. by C. POUGENS, 2 vol. (1799–1821); and *Oeuvres et correspondances inédites de d'Alembert,* ed. by C. HENRY (1887). General studies are: JOSEPH BERTRAND, *D'Alembert* (1889); and RONALD GRIMSLEY, *Jean d'Alembert, 1717–83* (1963). More specialized are: M. MULLER, *Essai sur la philosophie de Jean d'Aiembert* (1926); JOHN N. PAPPAS, *Voltaire and d'Alembert* (1962); and THOMAS L. HANKINS, *Jean d'Alembert: Science and the Enlightenment* (1970).

(R.Gr.)

Aleppo

Aleppo (Arabic Ḥalab, Turkish Halep) is now the principal city of northern Syria; under Ottoman rule, it was the greatest commercial centre of the Near East. The city is built at an altitude of 390 metres (1,279 feet) in a shallow hollow crossed by a small stream, the Quwayq. The climate is dry (annual rainfall averages under 18 inches) but healthful. Modern Aleppo is the capital of the Aleppo Administrative District.

History. In prehistoric times the site's rocky elevation and surrounding resources gave rise to early rural settlement. The city is first mentioned at the end of the 3rd millennium BC, when a dynasty of Hittite origin ruled in Aleppo, then called Halap. The Mitannians controlled Aleppo *c.* 1650 BC, and it was subjugated by Egypt in the 15th century BC. For a while the Egyptians and Mitanni were able to withstand the Hittites, but they eventually retook the city. With the collapse of the Hittite Empire (*c.* 1200 BC) Aleppo seems to have reverted to a small rural market town. At the end of the 9th century, the town was subject to the kingdom of Urartu centred on Lake Van in eastern Anatolia, but in 742 BC the Assyrians under Tiglath-pileser III conquered Urartu and annexed Aleppo and northern Syria in 738.

The Hittite period

Cyrus II the Great in 539 BC integrated Aleppo into the Achaemenian Empire of Persia. After the death of Alexander the Great (323 BC), Aleppo was ruled by the Seleucids, who renamed it Beroea (Beroia). In the 1st century BC, the city was incorporated in the Roman province of Syria. The Byzantine period was characterized by great commercial activity and the construction of many Christian and Jewish monuments. In AD 540, Aleppo was pillaged and burned in the course of a Persian invasion launched by the Sāsānian king Khosrow I.

The Arabs occupied Aleppo without resistance in 637. Part of the population was converted to Islām, but the

city played little part in the early history of the empire of the Caliphate. At the beginning of the 10th century, the house of Ḥamdān established itself in Aleppo, which then became an independent principality. It reached its zenith under Sayf ad-Dawlah (943–967), who had a brilliant court and surrounded himself with men of letters.

In 962, Aleppo was sacked by the Byzantine emperor Nicephorus II Phocas and temporarily paid tribute to the empire. It then fell successively to the Fāṭimids in 1015 and the Mirdāsids in 1023. In 1070 the Seljuq, Alp-Arslān, conquered the town; the ʿUqaylids established themselves there in 1079, but a minor Seljuq dynasty replaced them in 1086. Too weak to resist the invading crusaders, Aleppo paid them tribute in 1100, but successfully repelled their attacks in 1118 and 1124. In 1129 Aleppo submitted to the prince (*atabeg*) of al-Mawsil (Mosul) ʿImād ad-Dīn Zangī, who succeeded in making it a bulwark against the crusaders. After 1146 his son Nūr ad-Dīn (Nureddin) made it the centre of a renaissance of the Sunnī sect of Islām. In 1183 the Ayyūbid Saladin acquired the city and gave it to his son aẓ-Ẓāhīr Ghāzī. From the middle of the 13th century Aleppo was placed on the defensive. It fell to the Mongols under Hülegü in January 1260 and was later incorporated into the Mamlūk Empire of Egypt. Aleppo was the victim of attacks throughout the 14th century and was ravaged by Timur Lenk (Tamerlane) in 1400. From the middle of the 15th century it underwent a remarkable commercial expansion, becoming the principal market in the Levant in the following century.

Commercial expansion

In 1516 Aleppo was incorporated into the empire of the Ottoman Turks and, despite Ottoman–Iranian conflicts in the area, its economic prosperity continued until the end of the 18th century.

In 1831 the Egyptian general Ibrahim Pasha occupied Aleppo, but pressure from the West forced him to evacuate Syria in 1839. Between 1850 and 1862 Aleppo was agitated by disturbances against Ottoman rule. Full economic recovery came after 1880, with the railroad.

In October 1918, Fayṣal I occupied Aleppo with the Allied forces, and it was thereafter linked with Syria, which remained a French mandate until 1941. The French mandatory administration created an independent state of Aleppo in 1920 but abolished it in 1923. Under the mandate Aleppo expanded and became an industrial town and a political and intellectual centre that rivalled Damascus. This development continued during World War II and after independence.

Remains and the modern city. Located at the crossroads of great commercial routes, about 60 miles from both the Mediterranean Sea and the Euphrates River, on a site commanding a height suitable for fortification, the present town is the result of the long course of development that began in the 3rd millennium BC. The mound of al-ʿAqabah marks the area of first settlement. The streets of the old town outline the checkered Hellenistic plan that constituted the framework of the Arab town. An influx of population in the Byzantine period brought about the haphazard construction of new districts—Baḥsita, Farafrā, al-Asfaris, and Jallūm—that stand in marked contrast to the well-planned Hellenistic city.

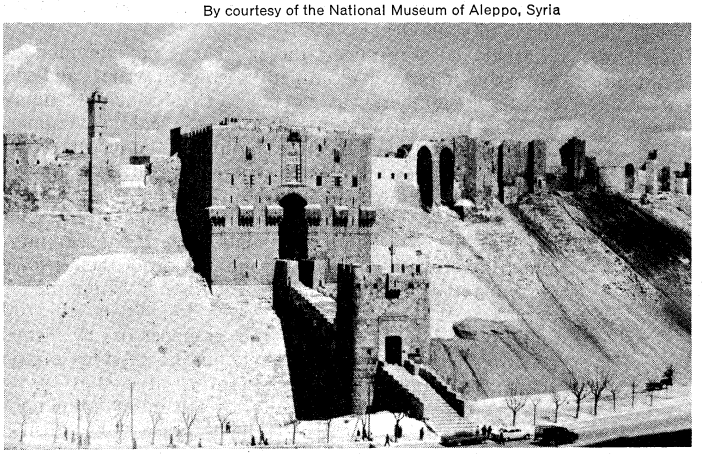
From this period date the remains of the Cathedral of St. Helena, converted into a mosque in 1125 and into the *madrasah*, or theological school, Ḥallāwāyah in 1149.

From the beginning of the Muslim period the Great Umayyad Mosque, built in 715, has been preserved, of which the minaret (late 11th century) is particularly notable. It was under the Ayyūbid dynasty in the 13th century that building flourished. Among the *madrasah*s are the Sulṭanīyah (1223) and, outside the walls of the city, al-Firdaws (1235). The ancient fortifications were often restored, and the ancient foundations on the western side of the city are still visible. These were surmounted by courses of stone in the Middle Ages, at which time the gates Bāb Anṭākiya and Bāb Qinnasrīn were given a new defensive disposition. The most remarkable monument of the town is the citadel, the silhouette of which is visible from some distance. Its present appearance dates from the reign of aẓ-Ẓāhīr Ghāzī (13th century), who remodelled the workmanship, covered the embankment with a glacis of stone, and built an important defensive work at its entrance. Some of the most beautiful and monumental covered bazaars in the Middle East are found in Aleppo and date from the Ottoman period: the Khān as-Sabūn (16th century) and the Khān al-Wazīr (17th century). At the end of the 19th century the life of the city became decentralized, and European influence made itself felt in the layout of such new quarters as al-ʿAzīzīyah, al-Jamīlīyah, al-Talal, al-Mushāriqa, Kuttab, and al-Judaydah. New peripheral districts have been added more recently.

Covered bazaars

Modern Aleppo is an industrial city with a population of about 500,000 (1970). Its principal industries are silk weaving and cotton printing, and subsidiary industries include the preparation of hides, wool, dried fruit, and nuts. Aleppo is an active intellectual centre with a university. The new National Museum (1969) contains material found in northern Syria at the ancient sites of Tall Aḥmar, Tall Ḥalaf, Arslān Ṭāsh and Ḥamāh, as well as the still active sites of Ugarit (Ras Shamra), Māriʿ (Tall Ḥarīrī), and ʿAyn Dārah.

By courtesy of the National Museum of Aleppo, Syria



The 13th-century citadel at Aleppo.

Administrative district. The Aleppo Administrative District has an area of 6,232 square miles and a population of close to a million. It is a rich agricultural area, extending to the Euphrates in the east and the Turkish frontier on the west. The main products of the region are grapes, wheat, barley, fruit, nuts, and sesame.

BIBLIOGRAPHY. JEAN SAUVAGET, *Alep, essai sur le développement d'une grande ville syrienne des origines au milieu du XIXe siècle* (1941), is the most important monograph on the northern Syrian metropolis with a large bibliography and an atlas of plates. The article "Ḥalab" in the *Encyclopaedia of Islam*, new ed., vol. 3 (1965), is a useful summary of all necessary information about the history and archaeology of the city. Quite different is *Corpus Inscriptionum Arabicarum*, pt. 2, *Syrie du Nord, inscriptions et monuments d'Alep*, 3 vol. (1954–56), in which E. HERZFELD published documents chronologically classified and accompanied by a commentary. S. SAWWAF, *Aleppo: Past and Present*, 3rd Eng. ed. (1965), is an account of the monuments of Aleppo and its neighbourhood with a short guide to the National Museum. IRA M. LAPIDUS in his *Muslim Cities in the Later Middle Ages* (1967), studies the social structures and political processes during the Mamlūk Period (1250–1517), centring his book on Damascus and Aleppo. A. RUSSELL, *The Natural History of Aleppo and Parts Adjacent*, 2nd ed., 2 vol. (1794), a basic work with plates, gives a description of the city and a good analysis of social life in the 18th century. See also HERBERT L. BODMAN, JR., *Political Factions in Aleppo, 1760–1826* (1963), a monograph on the political situation; and N. CHEHADE, "Aleppo," in M. BERGER (ed.), *The New Metropolis in the Arab World* (1963), information about the contemporary economic and social structures with statistics.

(N.E.)

Alexander III, Pope

One of the most vigorous exponents of papal authority in the 12th century, Rolando Bandinelli reigned as Pope

Alexander III from 1159 to 1181, frequently quarrelling with European rulers such as Henry II of England and Emperor Frederick Barbarossa. Bandinelli was born *c.* 1105 in Siena, Italy. After studies in theology and law, he became professor of law at Bologna and emerged as an important legal scholar and theologian. He wrote a commentary on the *Decretum Gratiani* and a book of *Sentences*, or theological opinions. He rose rapidly in the church during the pontificate of Pope Eugenius III and, during the reign of Pope Adrian IV, served as chief papal negotiator with Emperor Frederick Barbarossa.

Quarrel with the empire

In the complex politics of the 12th century, Rolando emerged as a man of keen judgment and shrewd understanding. His intellect was subtle and his instincts diplomatic. He belonged to that group of cardinals in the Roman Curia who feared the growing strength of the Holy Roman Empire in Italy and inclined toward the Norman kingdom of Sicily as a means of redressing the balance of power. He participated in the drawing up of the Concordat of Benevento (1156) between the papacy and King William I of Sicily. He revealed his fear of the empire still further in the following year at Besançon (1157), where he referred to the empire as a "benefice" of the papacy. The term aroused a storm of controversy with the imperial chancellor Rainald of Dassel, who argued that the term implied that the empire was a fief of the church and thus was an insult to the Emperor. Rolando and the Pope maintained that it meant only "benefit," but they could hardly have been unaware of the ambiguity of the term. Most likely, they intended its use as a warning to Frederick Barbarossa.

The papal election of 1159, in which the majority of the cardinals chose Rolando as pope under the name Alexander III, witnessed a strong effort on the part of Frederick to secure the election of a candidate favourable to his policies. A minority of the cardinals chose Cardinal Octavian (who took the name Victor IV), thus beginning a line of antipopes. Alexander, faced by strong imperial opposition in Italy, fled to France in April 1162. This move prevented a total victory by the Emperor and enabled Alexander to build support in France and England, where he gained the recognition of kings Louis VII and Henry II. During this period Alexander also continued to hold the loyalty of most of the clergy in Italy, especially in the south, and many in Germany. He continued to press forward the program of church reform begun in the previous century under the leadership of Pope Gregory VII. He supported Thomas Becket, archbishop of Canterbury, in his dispute with King Henry II of England on the issue of the legal status of the clergy, despite the risk that he would lose much-needed royal support. And he condemned certain propositions of Henry's Constitutions of Clarendon. If Alexander's efforts on Becket's behalf were cautious, he did not compromise the principles on which the Archbishop's case was based. After the murder of Becket, Alexander found Henry easier to deal with and was able to reach some agreement.

Medieval church-state relations

Papal relations with the empire in the 12th century revolved around the problems, both theoretical and practical, created by two autonomous powers—one spiritual, the other temporal—vying for authority in the lives of men. The church claimed primary responsibility over moral decisions; secular authorities were attempting to carve out for themselves a sphere of competence over political matters. There was no clear-cut distinction between the two areas, though constant efforts were being made to define them. The important fact is that during the 11th and early 12th centuries medieval society had become increasingly a dualistic society, recognizing two sources of authority and attempting to reconcile them. Alexander found himself playing a large role in the political arena in defense of what he regarded as the legitimate authority of the church. The conflict with Frederick Barbarossa, which consumed most of his efforts in the 1160s and 1170s, was perceived by him as a defense of the papacy, on which the liberty of the church rested.

Following the return of Alexander III to Rome in 1165, which was the result of a more favourable political climate in Italy caused by the temporary absence of Frederick Barbarossa, the conflict entered its critical period. In 1166, Frederick returned to Italy and forced the Pope into exile once more. He entered Rome, where he received the imperial crown from his current antipope, Paschal III. Alexander now turned to the communes of northern Italy for support, finding in many of them a deep concern over the protection of their independence from the empire that united them to his cause. The result was the formation of the Lombard League, which provided the Pope with the support essential to carry on his conflict with Barbarossa.

Alexander was unwilling, however, to take extreme measures against the Emperor, whom he saw as the legitimate secular leader of Christendom. He rejected the notion proposed by the Byzantine emperor Manuel I Comnenus of a reunification of East and West under Byzantine rule and, instead, placed greater reliance on the Normans of southern Italy and the Lombard cities. It was this policy that was ultimately to prevail and to lay the foundation for the policies followed by the papal Curia in the 13th century. Frederick found himself increasingly isolated in Italy and at odds with powerful elements in Germany. His decisive defeat by the Lombards at Legnano (1176) paved the way for the Peace of Venice (1177), which closed this phase of the struggle.

Assessment

In his rule of the church, Alexander stood in the reform tradition and at the beginning of a line of lawyer-popes that would culminate in Innocent III. His concern for education was typical of his age, which witnessed the early development of the universities, in which he played a role both as teacher and later as Pope (by ordering that the license to teach should be conferred on worthy candidates without charge). He presided over important synods at Beauvais (1160) and Tours (1163), as well as over the third Lateran Council (1179). He was among the first popes to express concern over the spread of the Catharist or Albigensian heresy in southern France.

With his death on August 30, 1181, the church found itself in a more assured position than at his accession. His policies served as important signposts to his successors. Yet he had succeeded throughout his troubled reign in preserving the image of his moderation and in winning respect even from his enemies. He was not merely one of the abler medieval popes but also one who won greater prestige for his office by his conduct.

BIBLIOGRAPHY. A brief and useful account of the pontificate of Alexander III in English may be found in MARSHALL BALDWIN, *Alexander III and the Twelfth Century* (1966). The standard biography is MARCEL PACAUT, *Alexandre III* (1956), in French. Pacaut's views are also available in his study of *Frédéric Barberousse* (1967; Eng. trans., 1970). Of fundamental importance is A. FLICHE and V. MARTIN, *Histoire de l'Église, depuis les origines jusqu'à nos jours*, vol. 9 (1953). H. MANN, *The Lives of the Popes in the Middle Ages*, 2nd ed., vol. 10 (1925), contains an account of the pontificate but has been superseded by Pacaut.

(J.M.Po.)

Alexander VI, Pope

Worldly, corrupt, and ambitious, Pope Alexander VI (Rodrigo Borgia; reigned 1492–1503) typifies the decadence of the Renaissance papacy. His neglect of the spiritual interests of the church contributed to the decline of papal prestige that led to the Protestant Reformation.

Rodrigo Borgia (Spanish Borja) was born in Játiva, Spain, in 1431, into the prominent and powerful Borgia family. His uncle Alonso de Borgia, bishop of Valencia (later cardinal), supervised his education and endowed him with ecclesiastical benefices while still in his teens. Rodrigo studied law at Bologna, and on February 22, 1456, he was created a cardinal by his uncle, now Pope Calixtus III. As vice chancellor of the Roman Church, Rodrigo amassed enormous wealth and, despite a severe rebuke from Pope Pius II, lived as a Renaissance prince. He patronized the arts and fathered a number of children for whom he provided livings, mainly in Spain. Of a Roman noblewoman, Vanozza Catanei, he had four sub-

sequently legitimized offspring—Juan, Cesare, Jofré, and Lucrezia—whose complicated careers troubled his pontificate.

Alinari



Alexander VI, detail of a fresco by Pinturicchio, 1492–94. In the Vatican.

Election to the papacy

Despite the shadow of simony that surrounded the disposal of his benefices among the papal electors, Rodrigo emerged from a tumultuous conclave on the night of August 10–11, 1492, as Pope Alexander VI and received the acclaim of the Roman populace. He embarked upon a reform of papal finances and a vigorous pursuit of the war against the Ottoman Turks. His position was menaced by the French king Charles VIII, who invaded Italy in 1494 to vindicate his claim to the kingdom of Naples. Charles, at the instigation of a rival cardinal of the influential della Rovere family, threatened the Pope with deposition and the convocation of a reform council. Politically isolated, Alexander sought assistance from the Turkish sovereign, Bayezid II. In the course of the Pope's meeting with King Charles in Rome in early 1495, however, he received the traditional obeisance from the French monarch. He still refused to support the King's claim to Naples and, by an alliance with Milan, Venice, and the Holy Roman Emperor, eventually forced the French to withdraw from Italy.

In September 1493 Alexander created his teenaged son Cesare a cardinal, along with Alessandro Farnese (the brother of the papal favourite Giulia la Bella and the future Pope Paul III). In the course of his pontificate Alexander appointed 47 cardinals to further his complicated dynastic, ecclesiastical, and political policies. His son Juan was made duke of Gandia (Spain) and was married to Maria Enriquez, the cousin of King Ferdinand IV of Castile; Jofré was married to Sancia, the granddaughter of the King of Naples; and Lucrezia was given first to Giovanni Sforza of Milan, and, when that marriage was annulled by papal decree on the grounds of impotence, she was married to Alfonso of Aragon. Upon his assassination Lucrezia received as a third husband Alfonso I d'Este, duke of Ferrara.

Alexander's attempt at reform

Tragedy struck the papal household on June 14, 1497, when Alexander's favourite son, Juan, was murdered. Gravely afflicted, Alexander announced a reform program and called for measures to restrain the luxury of the papal court, reorganize the Apostolic Chancery, and repress simony and concubinage. The Pontiff had shown great forebearance in dealing with the Dominican friar Girolamo Savonarola, who usurped political control in Florence in 1494, condemned the evils of the papal court, and called for the Pope's deposition, and, even before the friar's downfall in May 1498, theologians and men of affairs had expressed support for the papacy. Meanwhile, however, Alexander had returned to a policy of political intrigue.

Cesare resigned the cardinalate in 1498 and married Charlotte d'Albret in order to cement the Borgia alliance with the French king Louis XII, whose request for a marriage annulment was granted by the Pope. By a ruthless policy of siege and assassination, Cesare brought the north of Italy under his control; he conquered the duchies of Romagna, Umbria, and Emilia and earned the admiration of Niccolò Machiavelli, who used Cesare as the model for his classic on politics, *The Prince*. In Rome, Alexander destroyed the power of the Orsini and Colonna families and concluded an alliance with Spain, granting Isabella and Ferdinand the title of Catholic sovereigns. He negotiated the Treaty of Tordesillas (1494), which divided the New World into Portugese and Spanish zones of exploration and which conferred on the monarchs the *patronato real*—royal control over the church in the mission lands they colonized.

Patron of the arts

As a patron of the arts, Alexander erected a centre for the University of Rome, restored the Castel Sant 'Angelo, built the monumental mansion of the Apostolic Chancery, embellished the Vatican palaces, and persuaded Michelangelo to draw plans for the rebuilding of St. Peter's Basilica. He proclaimed the year 1500 a Holy Year of Jubilee and authorized its celebration with great pomp.

Embroiled in political and family manoeuvring to the end, the Pope was taken with a fever and, after confessing and receiving the last sacraments, died in Rome on August 18, 1503.

Recent attempts to whitewash Alexander's private conduct have proved abortive. While his religious convictions cannot be challenged, scandal accompanied his activities throughout his career. Even from a Renaissance viewpoint, his relentless pursuit of political goals and unremitting efforts to aggrandize his family were seen as excessive. Neither as corrupt as depicted by Machiavelli and by gossip nor as useful to the church's expansion as apologists would make him, Alexander VI holds a high place on the list of the so-called bad popes.

BIBLIOGRAPHY. LUDWIG VON PASTOR, *Geschichte der Päpste seit dem Ausgang des Mittelalters* (Eng. trans., *The History of the Popes, from the Close of the Middle Ages*, 2nd ed., vol. 5 and 6, ed. by F.I. ANTROBUS, 1901), an honest, objective, documented biography based on Vatican sources; ORESTES FERRARA, *El Papa Borgia* (Eng. trans., *The Borgia Pope, Alexander the Sixth*, 1940), an interesting but abortive attempt to exonorate papal misdeeds; MICHAEL DE LA BEDOYERE, *The Meddlesome Friar and the Wayward Pope: The Story of the Conflict Between Savonarola and Alexander VI* (1958), a vigorous, popular account; G. SORANZO, *Il tempo di Alessandro VI papa e di Fra Girolamo Savonarola* (1960), a serious, critical study; PHILIP HUGHES, *A History of the Church*, vol. 3, pp. 403–414 (1947), a brief, reliable account.

(F.X.M.)

Alexander III the Great

Alexander the Great (Alexander III), king of Macedonia, overthrew the Persian Empire, carried Macedonian arms to India, and laid the foundations for the Hellenistic world of territorial kingdoms. Already in his lifetime the subject of fabulous stories, he later became the hero of a full-scale legend bearing only the sketchiest resemblance to his historical career.

LIFE

He was born in 356 BC at Pella in Macedonia, the son of Philip II and Olympias (daughter of King Neoptolemus of Epirus). From age 13 to 16 he was taught by Aristotle, who inspired him with an interest in philosophy, medicine, and scientific investigation; but he was later to advance beyond his teacher's narrow precept that non-Greeks should be treated as slaves. Left in charge of Macedonia in 340 during Philip's attack on Byzantium, Alexander defeated the Maedi, a Thracian people; two years later he commanded the left wing at the Battle of Chaeronea, in which Philip defeated the allied Greek states, and displayed personal courage in breaking the Sacred Band of Thebes. A year later Philip divorced Olympias; and after a quarrel at a feast held to celebrate his father's new marriage, Alexander and his mother fled to Epirus, and Alexander later went to Illyria. Shortly afterward father and son were reconciled and Alexander returned; but his position as heir was jeopardized.



Alexander III the Great, portrait head on a coin of Lysimachus (355–281 BC). In the British Museum.

By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.

Alexander succeeds Philip

In 336, however, on Philip's assassination, Alexander, acclaimed by the army, succeeded without opposition. He at once executed the princes of Lyncestis, alleged to be behind Philip's murder, along with all possible rivals and the whole of the faction opposed to him. He then marched south, recovered a wavering Thessaly, and at an assembly of the Greek League at Corinth was appointed generalissimo for the forthcoming invasion of Asia, already planned and initiated by Philip. Returning to Macedonia by way of Delphi (where the Pythian priestess acclaimed him "invincible"), he advanced into Thrace in spring 335 and, after forcing the Shipka Pass and crushing the Triballi, crossed the Danube to disperse the Getae; turning west, he then defeated and shattered a coalition of Illyrians who had invaded Macedonia. Meanwhile, a rumour of his death had precipitated a revolt of Theban democrats; other Greek states favoured Thebes, and the Athenians, urged on by Demosthenes, voted help. In 14 days Alexander marched 240 miles from Pelion (near modern Korce, Albania) in Illyria to Thebes. When the Thebans refused to surrender, he made an entry and razed their city to the ground, sparing only temples and Pindar's house; 6,000 were killed and all survivors sold into slavery. The other Greek states were cowed by this severity, and Alexander could afford to treat Athens leniently. Macedonian garrisons were left in Corinth, Chalcis, and the Cadmea (the citadel of Thebes).

Beginnings of the Persian expedition. From his accession Alexander had set his mind on the Persian expedition. He had grown up to the idea. Moreover, he needed the wealth of Persia if he was to maintain the army built by Philip and pay off the 500 talents he owed.

Exploits of the Ten Thousand

The exploits of the Ten Thousand, Greek soldiers of fortune, and of Agesilaus of Sparta, in successfully campaigning in Persian territory had revealed the vulnerability of the Persian Empire. With a good cavalry force Alexander could expect to defeat any Persian army. In spring 334 he crossed the Dardanelles, leaving Antipater, who had already faithfully served his father, as his deputy in Europe with over 13,000 men; he himself commanded about 30,000 foot and over 5,000 cavalry, of whom nearly 14,000 were Macedonians and about 7,000 allies sent by the Greek League. This army was to prove remarkable for its balanced combination of arms. Much work fell on the lightarmed Cretan and Macedonian archers, Thracians, and the Agrianian javelin men. But in pitched battle the striking force was the cavalry, and the core of the army, should the issue still remain undecided after the cavalry charge, was the infantry phalanx, 9,000 strong, armed with 13-foot spears and shields, and the 3,000 men of the royal battalions, the hypaspists. Alexander's second in command was Parmenio, who had secured a foothold in Asia Minor during Philip's lifetime; many of his family and supporters were entrenched in positions of responsibility. The army was accompanied by surveyors, engineers, architects, scientists, court officials, and historians; from the outset Alexander seems to have envisaged an unlimited operation.

After visiting Ilium (Troy), a romantic gesture inspired by Homer, he confronted his first Persian army, led by three satraps, at the Granicus (modern Kocabaş) River, near the Sea of Marmara (May/June 334). The Persian plan to tempt Alexander across the river and kill him in the melee almost succeeded; but the Persian line broke, and Alexander's victory was complete. Darius' Greek mercenaries were largely massacred, but 2,000 survivors were sent back to Macedonia in chains. This victory exposed western Asia Minor to the Macedonians, and most cities hastened to open their gates. The tyrants were expelled and (in contrast to Macedonian policy in Greece) democracies were installed. Alexander thus underlined his Panhellenic policy, already symbolized in the sending of 300 panoplies (sets of armour) taken at the Granicus as an offering dedicated to Athena at Athens by "Alexander son of Philip and the Greeks (except the Spartans) from the barbarians who inhabit Asia." (This formula, cited by the Greek historian Arrian in his history of Alexander's campaigns, is noteworthy for its omission of any reference to Macedonia.) But the cities remained de facto under Alexander, and his appointment of Calas as satrap of Hellespontine Phrygia reflected his claim to succeed the Great King of Persia. When Miletus, encouraged by the proximity of the Persian fleet, resisted, Alexander took it by assault; but, refusing a naval battle, he disbanded his own costly navy and announced that he would "defeat the Persian fleet on land," by occupying the coastal cities. In Caria, Halicarnassus resisted and was stormed; but Ada, the widow and sister of the satrap Idrieus, adopted Alexander as her son and, after expelling her brother Pixodarus, Alexander restored her to her satrapy. Some parts of Caria held out, however, until 332.

Asia Minor and the Battle of Issus. In winter 334–333 Alexander conquered western Asia Minor, subduing the hill tribes of Lycia and Pisidia; and in spring 333 he advanced along the coastal road to Perga, passing the cliffs of Mt. Climax, thanks to a fortunate change of wind. The fall in the level of the sea was interpreted as a mark of divine favour by Alexander's flatterers, including the historian Callisthenes. At Gordium in Phrygia, tradition records his cutting of the Gordian knot, which could only be loosed by the man who was to rule Asia; but this story may be apocryphal or at least distorted. At this point Alexander benefitted from the sudden death of Memnon, the competent Greek commander of the Persian fleet. From Gordium he pushed on to Ancyra (modern Ankara) and thence south through Cappadocia and the Cilician Gates (modern Külek Boğazi); a fever held him up for a time in Cilicia. Meanwhile, Darius with his Grand Army had advanced northward on the eastern side of Mt. Amanus. Intelligence on both sides was faulty, and Alexander was already encamped by Myriandrus (near modern Iskenderun, Turkey) when he learned that Darius was astride his line of communications at Issus, north of Alexander's position (autumn 333). Turning, Alexander found Darius drawn up along the Pinarus River. In the battle that followed, Alexander won a decisive victory. The struggle turned into a Persian rout and Darius fled, leaving his family in Alexander's hands; the women were treated with chivalrous care.

The Gordian knot

Conquest of the Mediterranean coast and Egypt. From Issus Alexander marched south into Syria and Phoenicia, his object being to isolate the Persian fleet from its bases and so to destroy it as an effective fighting force. The Phoenician cities Marathus and Aradus came over quietly, and Parmenio was sent ahead to secure Damascus and its rich booty, including Darius' war chest. In reply to a letter from Darius offering peace, Alexander replied arrogantly, recapitulating the historic wrongs of Greece and demanding unconditional surrender to himself as lord of Asia. After taking Byblos (modern Jubayl) and Sidon (Arabic Ṣaydā), he met with a check at Tyre, where he was refused entry into the island city. He thereupon prepared to use all methods of siegecraft to take it, but the Tyrians resisted, holding out for seven months. In the meantime (winter 333–332) the Persians had counter-

attacked by land in Asia Minor—where they were defeated by Antigonus, the satrap of Greater Phrygia—and by sea, recapturing a number of cities and islands.

While the siege of Tyre was in progress, Darius sent a new offer: he would pay a huge ransom of 10,000 talents for his family and cede all his lands west of the Euphrates. "I would accept," Parmenio is reported to have said, "were I Alexander"; "I too," was the famous retort, "were I Parmenio." The storming of Tyre in July 332 was Alexander's greatest military achievement; it was attended with great carnage and the sale of the women and children into slavery. Leaving Parmenio in Syria, Alexander advanced south without opposition until he reached Gaza on its high mound; there bitter resistance halted him for two months, and he sustained a serious shoulder wound during a sortie. There is no basis for the tradition that he turned aside to visit Jerusalem.

The siege of Tyre

In November 332 he reached Egypt. The people welcomed him as their deliverer, and the Persian satrap Mazaces wisely surrendered. At Memphis Alexander sacrificed to Apis, the Greek term for Hapi, the sacred Egyptian bull, and was crowned with the traditional double crown of the pharaohs; the native priests were placated and their religion encouraged. He spent the winter organizing Egypt, where he employed Egyptian governors, keeping the army under a separate Macedonian command. He founded the city of Alexandria near the western arm of the Nile on a fine site between the sea and Lake Mareotis, protected by the island of Pharos, and had it laid out by the Rhodian architect Deinocrates. He is also said to have sent an expedition to discover the causes of the flooding of the Nile. From Alexandria he marched along the coast to Paraetonium and from there inland to visit the celebrated oracle of the god Amon (at Sīwah); the difficult journey was later embroidered with flattering legends. On his reaching the oracle in its oasis, the priest gave him the traditional salutation of a pharaoh, as son of Amon; Alexander consulted the god on the success of his expedition but revealed the reply to no one. Later the incident was to contribute to the story that he was the son of Zeus and, thus, to his "deification." In spring 331 he returned to Tyre, appointed a Macedonian satrap for Syria, and prepared to advance into Mesopotamia. His conquest of Egypt had completed his control of the whole eastern Mediterranean coast.

In July 331 Alexander was at Thapsacus on the Euphrates. Instead of taking the direct route down the river to Babylon, he made across northern Mesopotamia toward the Tigris, and Darius, learning of this move from an advance force sent under Mazaeus to the Euphrates crossing, marched up the Tigris to oppose him. The decisive battle of the war was fought on the plain of Gaugamela between Nineveh and Arbela. Alexander pursued the defeated Persian forces for 35 miles to Arbela, but Darius escaped with his Bactrian cavalry and Greek mercenaries into Media.

Battle of Gaugamela

Alexander now occupied Babylon, city and province; Mazaeus, who surrendered it, was confirmed as satrap in conjunction with a Macedonian troop commander, and quite exceptionally was granted the right to coin. As in Egypt, the local priesthood was encouraged. Susa, the capital, also surrendered, releasing huge treasures amounting to 50,000 gold talents; here Alexander established Darius' family in comfort. Crushing the mountain tribe of the Ouxians, he now pressed on over the Zagros range into Persia proper and, successfully turning the Pass of the Persian Gates, held by the satrap Ariobarzanes, he entered Persepolis and Pasargadae. At Persepolis he ceremonially burned down the palace of Xerxes, as a symbol that the Panhellenic war of revenge was at an end; for such seems the probable significance of an act that tradition later explained as a drunken frolic inspired by Thaïs, an Athenian courtesan. In spring 330 Alexander marched north into Media and occupied its capital Ecbatana. The Thessalians and Greek allies were sent home; henceforward he was waging a purely personal war.

As Mazaeus' appointment indicated, Alexander's views on the empire were changing. He had come to envisage a joint ruling people consisting of Macedonians and Persians, and this served to augment the misunderstanding that now arose between him and his people. Before continuing his pursuit of Darius, who had retreated into Bactria, he assembled all the Persian treasure and entrusted it to Harpalus, who was to hold it at Ecbatana as chief treasurer. Parmenio was also left behind in Media to control communications; the presence of this older man had perhaps become irksome.

In midsummer 330 Alexander set out for the eastern provinces at a high speed via Rhagae (modern Rayy, near Tehrān) and the Caspian Gates, where he learned that Bessus, the satrap of Bactria, had deposed Darius. After a skirmish near modern Shāhrūd, the usurper had Darius stabbed and left him to die. Alexander sent his body for burial with due honours in the royal tombs at Persepolis.

Death of Darius

Campaign eastward, to Central Asia. Darius' death left no obstacle to Alexander's claim to be Great King, and a Rhodian inscription of this year (330) calls him "lord of Asia"—*i.e.*, of the Persian Empire; soon afterward his Asian coins carry the title of king. Crossing the Elburz Mountains to the Caspian, he seized Zadracarta in Hyrcania and received the submission of a group of satraps and Persian notables, some of whom he confirmed in their offices; in a diversion westward, perhaps to modern Āmol, he reduced the Mardi, a mountain people who inhabited the Elburz Mountains. He also accepted the surrender of Darius' Greek mercenaries. His advance eastward was now rapid. In Aria he reduced Satibarzanes, who had offered submission only to revolt, and he founded Alexandria of the Arians (modern Herāt). At Phrada in Drangiana (either near modern Nad-e ʿAli in Seistan or farther north at Farah), he at last took steps to destroy Parmenio and his family. Philotas, Parmenio's son, commander of the elite Companion cavalry, was implicated in an alleged plot against Alexander's life, condemned by the army, and executed; and a secret message was sent to Cleander, Parmenio's second in command, who obediently assassinated him. This ruthless action excited widespread horror but strengthened Alexander's position relative to his critics and those whom he regarded as his father's men. All Parmenio's adherents were now eliminated and men close to Alexander promoted. The Companion cavalry was reorganized in two sections, each containing four squadrons (now known as hipparchies); one group was commanded by Alexander's oldest friend, Hephaestion, the other by Cleitus, an older man. From Phrada, Alexander pressed on during the winter of 330–329 up the valley of the Helmand River, through Arachosia, and over the mountains past the site of modern Kabul into the country of the Paropamisadae, where he founded Alexandria by the Caucasus.

Bessus was now in Bactria raising a national revolt in the eastern satrapies with the usurped title of Great King. Crossing the Hindu Kush northward over the Khawak Pass (11,650 feet), Alexander brought his army, despite food shortages, to Drapsaca (sometimes identified with modern Banu [Andarab], probably farther north at Qunduz); outflanked, Bessus fled beyond the Oxus (modern Amu Darya), and Alexander, marching west to Bactra-Zariaspa (modern Balkh [Wazirabad] in Afghanistan), appointed loyal satraps in Bactria and Aria. Crossing the Oxus, he sent his general Ptolemy in pursuit of Bessus, who had meanwhile been overthrown by the Sogdian Spitamenes. Bessus was captured, flogged, and sent to Bactra, where he was later mutilated after the Persian manner (losing his nose and ears); in due course he was publicly executed at Ecbatana.

Revolt in Bactria and Sogdiana

From Maracanda (modern Samarkand) Alexander advanced by way of Cyropolis to the Jaxartes (modern Syrdarya), the boundary of the Persian Empire. There he broke the opposition of the Scythian nomads by his use of catapults and, after defeating them in a battle on the north bank of the river, pursued them into the interior. On the site of modern Leninabad (Khojent) on the Jaxartes, he founded a city, Alexandria Eschate, "the farthest." Meanwhile, Spitamenes had raised all Sogdiana

in revolt behind him, bringing in the Massagetai, a people of the Śaka confederacy. It took Alexander until the autumn of 328 to crush the most determined opponent he encountered in his campaigns. Later in the same year he attacked Oxyartes and the remaining barons who held out in the hills of Paraetacene (modern Tadzhikistan); volunteers seized the crag on which Oxyartes had his stronghold, and among the captives was his daughter, Roxana. In reconciliation Alexander married her, and the rest of his opponents were either won over or crushed.

An incident that occurred at Maracanda widened the breach between Alexander and many of his Macedonians. He murdered Cleitus, one of his most trusted commanders, in a drunken quarrel; but his excessive display of remorse led the army to pass a decree convicting Cleitus posthumously of treason. The event marked a step in Alexander's progress toward Eastern absolutism, and this growing attitude found its outward expression in his use of Persian royal dress. Shortly afterward, at Bactra, he attempted to impose the Persian court ceremonial, involving prostration (*proskynesis*), on the Greeks and Macedonians too; but to them this custom, habitual for Persians entering the king's presence, implied an act of worship and was intolerable before a man. Even Callisthenes, historian and nephew of Aristotle, whose ostentatious flattery had perhaps encouraged Alexander to see himself in the role of a god, disappointingly refused to abase himself. Macedonian laughter caused the experiment to founder, and Alexander was wise enough to abandon it. Shortly afterward, however, Callisthenes was held to be privy to a conspiracy among the royal pages and was executed (or died in prison; accounts vary); resentment of this action alienated much sympathy from Alexander within the Peripatetic school of philosophers, with which Callisthenes had close connections.

Progress toward absolutism

Invasion of India. In early summer 327 Alexander left Bactria with a reinforced army under a reorganized command. If Plutarch's figure of 120,000 men has any reality, however, it must include all kinds of auxiliary services, together with muleteers, camel drivers, medical corps, peddlers, entertainers, women, and children; the fighting strength perhaps stood at about 35,000. Recrossing the Hindu Kush, probably by Bamian and the Ghorband Valley, Alexander divided his forces. Half the army with the baggage under Hephaestion and Perdiccas, both cavalry commanders, was sent through the Khyber Pass, while he himself led the rest, together with his siege train, through the hills to the north. His advance through Swāt and Gandhāra was marked by the storming of the almost impregnable pinnacle of Aornos, the modern Pir-Sar, a few miles west of the Indus and north of the Buner River, an impressive feat of siegecraft. In spring 326, crossing the Indus near Attock, Alexander entered Taxila, whose ruler, Taxiles, furnished elephants and troops in return for aid against his rival Porus, who ruled the lands between the Hydaspes (modern Jhelum) and the Acesines (modern Chenab). In June Alexander fought his last great battle on the left bank of the Hydaspes. He founded two cities there, Alexandria Nicaea (to celebrate his victory) and Bucephala (named after his horse Bucephalus, which died there); and Porus became his ally.

Battle of the Hydaspes

How much Alexander knew of India beyond the Hyphasis (probably the modern Beas) is uncertain; there is no conclusive proof that he had heard of the Ganges. But he was anxious to press on farther, and he had advanced to the Hyphasis when his army mutinied, refusing to go farther in the tropical rain; they were weary in body and spirit, and Coenus, one of Alexander's four chief marshals, acted as their spokesman. On finding the army adamant, Alexander agreed to turn back.

On the Hyphasis he erected 12 altars to the 12 Olympian gods, and on the Hydaspes he built a fleet of 800 to 1,000 ships. Leaving Porus, he then proceeded down the river and into the Indus, with half his forces on shipboard and half marching in three columns down the two banks. The fleet was commanded by Nearchus, and Alexander's own captain was Onesicritus; both later wrote accounts of the campaign. The march was attended with much fighting and heavy, pitiless slaughter; at the storming of one town of the Malli near the Hydraotes (Ravi) River Alexander received a severe wound which left him weakened.

On reaching Patala, located at the head of the Indus delta, he built a harbour and docks and explored both arms of the Indus, which probably then ran into the Rann of Kutch. He planned to lead part of his forces back by land, while the rest in perhaps 100 to 150 ships under the command of Nearchus, a Cretan explorer, made a voyage of exploration along the Persian Gulf. Local opposition led Nearchus to set sail in September (325), and he was held up for three weeks until he could pick up the northeast monsoon in late October. In September Alexander too set out along the coast through Gedrosia (modern Baluchistan), but he was soon compelled by mountainous country to turn inland, thus failing in his project to establish food depots for the fleet. Craterus, a high-ranking officer, already had been sent off with the baggage and siege train, the elephants, and the sick and wounded, together with three battalions of the phalanx, by way of the Mulla Pass, Quetta, and Kandahar into the Helmand Valley; from there he was to march through Drangiana to rejoin the main army on the Amanis (modern Minab) River in Carmania. Alexander's march through Gedrosia proved disastrous; waterless desert and shortage of food and fuel caused great suffering, and many, especially women and children, perished in a sudden monsoon flood while encamped in a wadi. At length, at the Amanis, he was rejoined by Nearchus and the fleet, which also had suffered losses.

Return from India

Consolidation of the empire. Alexander now proceeded farther with the policy of replacing senior officials and executing defaulting governors on which he had already embarked before leaving India. Between 326 and 324 over a third of his satraps were superseded and six were put to death, including the Persian satraps of Persis, Susiana, Carmania, and Paraetacene; three generals in Media, including Cleander, the brother of Coenus (who had died a little earlier), were accused of extortion and summoned to Carmania, where they were arrested, tried, and executed. How far the rigour that from now onward Alexander displayed against his governors represents exemplary punishment for gross maladministration during his absence and how far the elimination of men he had come to distrust (as in the case of Philotas and Parmenio) is debatable; but the ancient sources generally favourable to him comment adversely on his severity.

In spring 324 he was back in Susa, capital of Elam and administrative centre of the Persian Empire; the story of his journey through Carmania in a drunken revel, dressed as Dionysus, is embroidered, if not wholly apocryphal. He found that his treasurer, Harpalus, evidently fearing punishment for peculation, had absconded with 6,000 mercenaries and 5,000 talents to Greece; arrested in Athens, he escaped and later was murdered in Crete. At Susa Alexander held a feast to celebrate the seizure of the Persian Empire, at which, in furtherance of his policy of fusing Macedonians and Persians into one master race, he and 80 of his officers took Persian wives; he and Hephaestion married Darius' daughters Barsine (also called Stateira) and Drypetis respectively, and 10,000 of his soldiers with native wives were given generous dowries.

Attempt at creating a master race

This policy of racial fusion brought increasing friction to Alexander's relations with his Macedonians, who had no sympathy for his changed concept of the empire. His determination to incorporate Persians on equal terms in the army and the administration of the provinces was bitterly resented. This discontent was now fanned by the arrival of 30,000 native youths who had received a Macedonian military training and by the introduction of Orientals from Bactria, Sogdiana, Arachosia, and other parts of the empire into the Companion cavalry; whether Orientals had previously served with the Companions is uncertain, but if so they must have formed separate squadrons. In addition, Persian nobles had been accepted into the royal cavalry bodyguard. Peucestas, the new governor of Persis, gave this policy full support to flatter Alex-

ander; but most Macedonians saw it as a threat to their own privileged position. The issue came to a head at Opis (324), when Alexander's decision to send home Macedonian veterans under Craterus was interpreted as a move toward transferring the seat of power to Asia. There was an open mutiny involving all but the royal bodyguard; but when Alexander dismissed his whole army and enrolled Persians instead, the opposition broke down. An emotional scene of reconciliation was followed by a vast banquet with 9,000 guests to celebrate the ending of the misunderstanding and the partnership in government of Macedonians and Persians—but not, as has been argued, the incorporation of all the subject peoples as partners in the commonwealth. Ten thousand veterans were now sent back to Macedonia with gifts, and the crisis was surmounted.

In summer 324 Alexander attempted to solve another problem, that of the wandering mercenaries, of whom there were thousands in Asia and Greece, many of them political exiles from their own cities. A decree brought by Nicanor to Europe and proclaimed at Olympia (September 324) required the Greek cities of the Greek League to receive back all exiles and their families (except the Thebans), a measure that implied some modification of the oligarchic regimes maintained in the Greek cities by Alexander's governor Antipater. Alexander now planned to recall Antipater and supersede him by Craterus; but he was to die before this could be done.

Last year and death

In autumn 324 Hephaestion died in Ecbatana, and Alexander indulged in extravagant mourning for his closest friend; he was given a royal funeral in Babylon with a pyre costing 10,000 talents. His post of chiliarch (grand vizier) was left unfilled. It was probably in connection with a general order now sent out to the Greeks to honour Hephaestion as a hero that Alexander linked the demand that he himself should be accorded divine honours. For a long time his mind had dwelt on ideas of godhead. Greek thought drew no very decided line of demarcation between god and man, for legend offered more than one example of men who, by their achievements, acquired divine status. Alexander had on several occasions encouraged favourable comparison of his own accomplishments with those of Dionysus or Heracles. He now seems to have become convinced of the reality of his own divinity and to have required its acceptance by others. There is no reason to assume that his demand had any political background (divine status gave its possessor no particular rights in a Greek city); it was rather a symptom of growing megalomania and emotional instability. The cities perforce complied, but often ironically: the Spartan decree read, "Since Alexander wishes to be a god, let him be a god."

In the winter of 324 Alexander carried out a savage punitive expedition against the Cossaeans in the hills of Luristan. The following spring at Babylon he received complimentary embassies from the Libyans and from the Bruttians, Etruscans, and Lucanians of Italy; but the story that embassies also came from more distant peoples, such as Carthaginians, Celts, Iberians, and even Romans, is a later invention. Representatives of the cities of Greece also came, garlanded as befitted Alexander's divine status. Following up Nearchus' voyage, he now founded an Alexandria at the mouth of the Tigris and made plans to develop sea communications with India, for which an expedition along the Arabian coast was to be a preliminary. He also dispatched Heracleides, an officer, to explore the Hyrcanian (*i.e.*, Caspian) Sea. Suddenly, while busy with plans to improve the irrigation of the Euphrates and to settle the coast of the Persian Gulf, Alexander was taken ill after a prolonged banquet and drinking bout; ten days later, on June 13, 323, he died in his 33rd year; he had reigned for twelve years and eight months. His body, diverted to Egypt by Ptolemy, the later king, was eventually placed in a golden coffin in Alexandria. Both in Egypt and elsewhere in the Greek cities he received divine honours.

No heir had been appointed to the throne, and his generals adopted Philip II's half-witted illegitimate son, Philip Arrhidaeus, and Alexander's posthumous son by Roxana, Alexander IV, as kings, sharing out the satrapies among themselves, after much bargaining. The empire could hardly survive Alexander's death as a unit. Both kings were murdered, Arrhidaeus in 317 and Alexander in 310/309. The provinces became independent kingdoms, and the generals, following Antigonus' lead in 306, took the title of king.

EVALUATION

Of Alexander's plans little reliable information survives. The far-reaching schemes for the conquest of the western Mediterranean and the setting up of a universal monarchy, recorded by Diodorus, a 1st-century Greek historian, are probably based on a later forgery; if not, they were at once jettisoned by his successors and the army. Had he lived, he would no doubt have completed the conquest of Asia Minor, where Paphlagonia, Cappadocia, and Armenia still maintained an effective independence. But in his later years Alexander's aims seem to have been directed toward exploration, in particular of Arabia and the Caspian.

Achievement

In the organization of his empire, Alexander had been content in many spheres to improvise and adapt what he found. His financial policy is an exception; though the details cannot be wholly recovered, it is clear that he set up a central organization with collectors perhaps independent of the local satraps. That this proved a failure was partly due to weaknesses in the character of Harpalus, a Macedonian feudal prince. But the establishment of a new coinage with a silver standard based on that of Athens in place of the old bimetallic system current both in Macedonia and in Persia helped trade everywhere and, combined with the release of vast amounts of bullion from the Persian treasuries, gave a much-needed fillip to the economy of the whole Mediterranean area.

Alexander's foundation of new cities—Plutarch speaks of over 70—initiated a new chapter in Greek expansion. No doubt many of the colonists, by no means volunteers, deserted these cities, and marriages with native women led to some dilution of Greek ways; but the Greek (rather than Macedonian) influence remained strong in most of them, and since the process was carried further by Alexander's Seleucid successors the spread of Hellenic thought and customs over much of Asia as far as Bactria and India was one of the more striking effects of Alexander's conquests.

His plans for racial fusion, on the other hand, were a failure. The Iranian satraps were perhaps not efficient, for out of 18, ten were removed or executed—with what justice it is no longer possible to say. But, more important, the Macedonians, leaders and men alike, rejected the idea, and in the later Seleucid Empire the Greek and Macedonian element was to be clearly dominant.

How far Alexander would have succeeded in the difficult task of coordinating his vast dominions, had he lived, is hard to determine. The only link between the many units that went to make up an empire more disparate than that of the Habsburgs, and far larger, was his own person; and his death came before he could tackle this problem.

What had so far held it all together was his own dynamic personality. He combined an iron will and ability to drive himself and his men to the utmost with a supple and flexible mind; he knew when to draw back and change his policy, though he did this reluctantly. He was imaginative and not without romantic impulses; figures like Achilles, Heracles, and Dionysus were often in his mind, and the salutation at the oracle of Amon clearly influenced his thoughts and ambitions ever afterward. He was swift in anger, and under the strain of his long campaigns this side of his character grew more pronounced. Ruthless and self-willed, he had increasing recourse to terror, showing no hesitation in eliminating men whom he had ceased to trust, either with or without the pretense of a fair trial. Years after his death, Cassander, son of Antipater, a regent of the Macedonian Empire under Alexander, could not pass his statue at Delphi without shuddering. Yet he maintained the loyalty of his men,

who followed him to the Hyphasis without complaining and continued to believe in him throughout all hardships. Only when his whim would have taken them still farther into unknown India did he fail to get his way.

Generalship

As a general Alexander is among the greatest the world has known. He showed unusual versatility both in the combination of different arms and in adapting his tactics to the challenge of enemies who commanded novel forms of warfare—the Śaka nomads, the Indian hill tribes, or Porus with his elephants. His strategy was skillful and imaginative, and he knew how to exploit the chances that arise in every battle and may be decisive for victory or defeat; he also drew the last advantage from victory by relentless pursuit. His use of cavalry was so effective that he rarely had to fall back upon his infantry to deliver the crushing blow.

Alexander's short reign marks a decisive moment in the history of Europe and Asia. His expedition and his own personal interest in scientific investigation brought many advances in the knowledge of geography and natural history. His career led to the moving of the great centres of civilization eastward and initiated the new age of the Greek territorial monarchies; it spread Hellenism in a vast colonizing wave throughout the Near East and created, if not politically, at least economically and culturally, a single world stretching from Gibraltar to the Punjab, open to trade and social intercourse and with a considerable overlay of common civilization and the Greek *koinē* as a lingua franca. It is not untrue to say that the Roman Empire, the spread of Christianity as a world religion, and the long centuries of Byzantium were all in some degree the fruits of Alexander's achievement.

BIBLIOGRAPHY. W.W. TARN, *Alexander the Great*, 2 vol. (1948), provides a narrative favourable to Alexander and a survey of the ancient sources; for cogent criticism, see E. BADIAN, "Alexander the Great and the Unity of Mankind," *Historia*, 7:425–444 (1958). A.R. BURN, *Alexander the Great and the Hellenistic Empire* (1947), is more dispassionate. Controversial issues are discussed in G.T. GRIFFITH (ed.), *Alexander the Great: The Main Problems* (1966), reprinting articles by various scholars, and the essays published in *Greece and Rome*, 12:113–228 (1965). The original sources for Alexander are lost, and the problem is to recover them from the secondary authorities: DIODORUS, bk. xvii; QUINTUS CURTIUS RUFUS; PLUTARCH, *Life of Alexander;* ARRIAN, *Anabasis* and *Indica;* JUSTINUS' abridgement of TROGUS. On this see L. PEARSON, *The Lost Histories of Alexander the Great* (1960). For Alexander as a general, see J.F.C. FULLER, *The Generalship of Alexander the Great* (1960); and E.W. MARSDEN, *The Campaign of Gaugamela* (1964).

(F.W.W.)

Alexander I of Russia

Alexander I, emperor of Russia from 1801 to 1825, was the prime mover in the coalition that finally defeated Napoleon I. Attracted first to Western liberalism and later to mystical religion, Alexander promoted a few internal reforms early in his reign; in the post-Napoleonic era, he was known chiefly for his quixotic efforts to form a Christian union among the rulers of Europe.

Early life and upbringing. The first child of Grand Duke Pavel Petrovich (later Paul I) and Grand Duchess Maria Fyodorovna, a princess of Württemberg-Montbéliard, Aleksandr Pavlovich was born on December 23 (December 12, old style), 1777. His grandmother, the reigning empress Catherine II (the Great), took him from his parents and raised him herself to prepare him to succeed her. She was determined to disinherit her own son Pavel, who repelled her by his instability and who hated her and considered her a usurper.

A friend and disciple of the philosophers of the French Enlightenment, Catherine invited Denis Diderot, the encyclopaedist, to become Alexander's private tutor. When he declined, she chose Frédéric-César La Harpe, a Swiss citizen, a republican by conviction and an excellent educator. He inspired deep affection in his pupil and permanently shaped his flexible and open mind.

As an adolescent, Alexander was allowed to visit his father at Gatchina, on the outskirts of St. Petersburg, away from the court. There, Pavel had created a ridiculous little kingdom where he devoted himself to military exercises and parades. Alexander received his military training at Gatchina under the direction of a tough and rigid officer, Aleksey Arakcheyev, who was faithfully attached to him and whom Alexander loved throughout his life. Alexander thus passed from Catherine's gracious court to Pavel's severe discipline. Obliged to change his personality as one changes clothing, he learned early to play opposing roles, to dissimulate, and to adapt himself to false situations. These metamorphoses may explain how he became, in the words of a diplomat of the period, "as sharp as a needle's point and as deceptive as the foam of the sea."



Alexander I, miniature by Jean-Baptiste Isabey, *c.* 1814. In the collection of Mrs. Merriweather Post, Hillwood, Washington, D.C.
By courtesy of Mrs. Merriweather Post, Hillwood, Washington, D.C.

Alexander's education was not continued after he was 16, when his grandmother married him to Princess Louise of Baden-Durlach, who was 14, in 1793. The precocious marriage had been arranged to guarantee descendants to the Romanov dynasty, and it was unhappy from the beginning. The sweet and charming girl who became Yelisaveta Alekseyevna was loved by everyone except her husband.

Catherine had already written the manifesto that deprived her son of his rights and designated her grandson as the heir to the throne, when she died suddenly on November 17 (November 6, O.S.), 1796. Alexander, who knew of it, did not dare to disclose the manifesto, and Pavel became emperor.

Early domestic policy. Paul I's reign was a dark period for Russia. The monarch's tyrannical and bizarre behaviour led to a plot against him by certain nobles and military men, and he was assassinated during the night of March 23 (March 11, O.S.), 1801. Alexander became tsar the next day. The plotters had let him in on the secret, assuring him they would not kill his father but would only demand his abdication. Alexander believed them or, at least, wished to believe that all would go well. This incident reveals another aspect of Alexander's personality. He had a capacity to see only what he wanted to see and, having become emperor through a crime, was forever burdened by remorse and haunted by a terrible sense of sin.

Ascent to the throne

After the darkness into which Paul had plunged Russia, Alexander appeared to his subjects as a radiant dawn. He was handsome, strong, pleasant, humane, and full of enthusiasm. He wanted his reign to be a happy one and dreamed of great and necessary reforms. With four friends, who were of noble families but motivated by liberal ideas—Prince Adam Czartoryski, Count Pavel

Stroganov, Count Viktor Kochubey, and Nikolay Novosiltsev—he formed the Private Committee (*Neglasny Comitet*). Its avowed purpose was to frame "good laws, which are the source of the well-being of the Nation."

Alexander and his close advisers corrected many of the injustices of the preceding reign and made many administrative improvements, such as the creation of Cabinet ministries. Their principal achievement was the initiation of a vast plan for public education, which involved the formation of many schools of different types, institutions for training teachers, and the founding of three new universities. Nevertheless, despite the humanitarian ideas inculcated in him by La Harpe and despite his own wish to make his people happy, Alexander lacked the energy necessary to carry out the most urgent reform, the abolition of serfdom. The institution of serfdom was, in the Tsar's own words, "a degradation and a misfortune for the Nation," which kept Russia in a disastrously backward state. But to liberate the serfs, who composed three-quarters of the total population, would inevitably arouse the hostility of their noble masters, who did not want to lose the slaves on whom their wealth and comfort depended. Serfdom was a continuing burden on the Russians. It prevented modernization of the country, which was at least a century behind the rest of Europe.

Failures as a reformer

Out of a sincere desire to innovate, Alexander had even considered a constitution and "the limitation of the autocracy," but he recoiled before the danger of imposing sudden change on a nobility that rejected it. Moreover, he was a visionary who could not transform his dreams into reality. Because of his unstable personality, he would become intoxicated by the notion of grand projects, while balking at carrying them out. He has been compared to a man who looks at the sea but is afraid to rush in or to a lover of art who pauses to admire a picture and passes on. Finally, the "Western" theoretical education of Alexander and his young friends had not prepared them for gaining a clear vision of the realities of Russian life.

Early foreign policy. Displaying an astonishing inconstancy, Alexander soon abandoned his internal reforms to devote himself to foreign policy, to which he would commit the major portion of his reign. Sensitive to fluctuations in continental politics, he was a "European" who hoped for peace and unity among nations. He felt that he was called to be a mediator, like his grandmother, who had been called the "Arbiter of Europe."

As soon as he came to power, Alexander resealed an alliance with England that had been broken by Paul I. He nonetheless maintained good relations with France in the hope of "moderating" Bonaparte by restraining his spirit of conquest, whereby Alexander could achieve his generous project of "general pacification." He signed a treaty of friendship with Prussia. A feeling of chivalry attached him to its king, Frederick William III, and to Queen Louise. Later, he got on good terms with Austria. His idealism persuaded him that this network of alliances would lead to a European federation.

First war against Napoleon

But Napoleon had other ideas. His territorial encroachments, desire for world hegemony, and his coronation in 1804 as emperor forced Alexander to declare war against him. Assuming the role of commander in chief, he relied on the Austrian generals and scorned the wise counsel of the Russian general Prince Kutuzov, who was a shrewd strategist. The Russians and Austrians were defeated at Austerlitz, in Moravia, on December 2, 1805, and the emperor Francis II was forced to sign the peace treaty, since his territory was occupied by the enemy. Russia remained intact behind its frontiers. Moreover, Napoleon wanted to spare the Tsar; he hoped to gain his friendship and to divide the world with him. Such a notion did not occur to Alexander, who was determined to take revenge.

In 1806 Napoleon defeated Prussia at Jena and Auerstädt. Despite the warnings of both his mother and his advisers, the Tsar rushed to the aid of his friend. The battles were fought in east Prussia. After a partial success at Eylau, the Russian army, under General Bennigsen, was decimated at Friedland, on June 14, 1807. Then occurred the meeting (June 25) of the two emperors on a raft in the middle of the Niémen off Tilsit (now Sovetsk). The sequel of these events demonstrates that, in the course of the Tilsit interview, it was the Tsar of Russia who deceived the Emperor of the French. Seeking to gain time he used his charm to play the admiring friend and to convince Napoleon of his sincerity. He accepted all the victor's conditions, promising to break with England; to adhere to the Continental System set up by Napoleon to isolate and weaken Great Britain; and to recognize the creation of the Grand Duchy of Warsaw, formed from the part of Poland given to Prussia during the Partition of 1795. In "recompense" Napoleon gave Alexander liberty to expand at the expense of Sweden and Turkey.

From the Peace of Tilsit to the 1812 Invasion. Most Russians were angered and humiliated by the Tilsit Alliance; they thought that breaking off trade with England would inevitably create a disastrous economic situation, but Alexander kept his plans secret and bided his time. He reorganized and strengthened his armies with the competent aid of Arakcheyev, the instructor from Gatchina who had become his indispensable colleague. Meanwhile, the monarch's popularity dropped; all levels of the population accused him of having uselessly sacrificed Russian blood and of ruining the country.

Aborted legal reforms

Alexander once again turned his attention to internal reforms. He placed responsibility for them on a remarkable legal writer, Mikhail Mikhaylovich Speransky. Of modest origins, Speransky's talent caused him to rise rapidly. He conceived a vast plan for total reorganization of Russian legal structures and authored a complete collection and a systematically coordinated digest of Russian laws. Only a very small part of his great plan was applied, for once again Alexander withdrew from any practical fulfillment, partly because foreign events distracted him from rebuilding his empire on new foundations.

Despite the strong Russian reaction against France, the Tsar again met Napoleon, at Erfurt in Saxony, in 1808, where he showed himself to have become distant from his Tilsit ally. When a new war broke out between France and Austria in 1809, Alexander, despite his commitments, did not intervene in Napoleon's behalf, contenting himself with feigning a military advance. Although he continued to assure Napoleon of his friendship, Franco-Russian relations cooled. Napoleon reproached the Tsar for trading with England under cover of neutral vessels and for refusing him the hand of his sister, the grand duchess Anna Pavlovna. For his part Alexander tried in vain to obtain from Napoleon a commitment not to create an independent Kingdom of Poland. When Napoleon annexed the German territories on the Baltic, including the Grand Duchy of Oldenburg, a fief of the Tsar's brother-in-law, Alexander protested against what he considered a personal offense.

All of this was only a pretext for military preparations on both sides. A violent shift of opinion against Napoleon appeared in Russia. The hostility toward France among the court, the nobility, and the dignitaries compelled Alexander to dismiss his legal adviser Speransky, a great admirer of Napoleon and his Code, and send him into exile. Changing his opinions yet again, the Tsar adopted the reactionary ideas of a patriotic group dominated by his favourite sister, the grand duchess Yekaterina Pavlovna. He judged that, under the conditions then prevailing, Russia had best keep her traditional institutions.

The defeat of Napoleon. Napoleon and his Grand Army of 600,000 men invaded Russia on June 24, 1812. The conflict that ensued was justly called the Patriotic War by the Russians; in it, the strong resistance and outstanding endurance of an entire people were displayed. The war transformed Alexander, suffusing him with energy and determination. The French advanced as rapidly as the Russians retreated, drawing them away from their bases. Napoleon thought that, once Moscow was taken, the Tsar would capitulate. But after the bloody Battle of

Borodino, Moscow was burned by order of its governor, Count Fyodor Rostopchin, and Alexander did not sue for peace. The conqueror had to camp in a ruined city where he could not remain. The Tsar, meanwhile, under pressure of public opinion, had named Kutuzov, whom he detested, supreme commander. The old warrior, through brilliant strategy and with the aid of heroic partisans, pursued the enemy and drove him from the country. The retreat from Russia was a terrible disaster for Napoleon; combined with his reverses in Spain, it precipitated his downfall.

Alexander had declared, "Napoleon or I: from now on we cannot reign together!" He said that the burning of Moscow had "illuminated his soul." He called Europe to arms, to rescue the people who had been enslaved by Napoleon's conquests. His enthusiasm, perseverance, and steadfast determination to triumph aroused the King of Prussia and the Emperor of Austria, and the enheartened allies were victorious at Leipzig in October 1813. This "Battle of Nations" could have been decisive, but Alexander wanted his revenge: no peace until he reached Paris. He entered Paris triumphantly in March 1814. Napoleon abdicated. The Tsar reluctantly accepted the restoration of the Bourbons, for whom he had little esteem, and imposed a constitutional charter on the new ruler, Louis XVIII. Alexander showed his generosity toward France, alleviating her condition as a defeated country and protesting that he had made war on Napoleon and not on the French people, whom he admired.

The most powerful sovereign in Europe

He had become the most powerful sovereign in Europe and the arbiter of her destinies, as he had wished. He inspired the covening of the greatest international congress in history in Vienna, in the autumn of 1814. It was a time of sumptuous feasts and also of diplomatic intrigues and bitter quarrels. The Tsar's allies, whom he had saved, now feared his power and opposed the annexation of Poland to Russia. It was his only claim in reward for what he had done, and he was determined to achieve it.

When Napoleon returned from his exile in Elba and regained the throne, the war resumed, ending with his final defeat by the allies at Waterloo on June 18, 1815. Again the victorious sovereigns met in Paris to frame a peace treaty, and once again Alexander intervened on behalf of France.

The final decade. This period marked a turning point for the Tsar. Since the invasion of his country, he had become religious; he read the Bible daily and prayed often. It was his frequent visits with the pietistic visionary Barbara Juliane Krüdener in Paris that turned him into a mystic. She was an impassioned woman who considered herself a prophetess sent to the Tsar by God, and, if her personal influence was of brief duration, Alexander nevertheless retained his newly found evangelical fervour and came to profess a nondogmatic "universal religion" strongly influenced by Quaker and Moravian beliefs.

Alexander obtained Poland, set it up as a kingdom with himself as king, and gave it a constitution, declaring his attachment to "free institutions" and his desire to "extend them throughout all the countries dependent on him." These words awakened great hopes in Russia, but, when the Tsar returned home after a long absence, he was no longer thinking of reform. He devoted his entire attention to the Russian Bible Society and to an unfortunate innovation, the military colonies, by which he attempted to settle soldiers and their families on the land so that they might enjoy more stable lives. These colonies brought great suffering to Russian soldiers and peasants. His old military mentor Arakcheyev was blamed for it, but the idea was the Tsar's alone. Arakcheyev was, nonetheless, very powerful, for Alexander placed the governing of the empire in his hands during his absences.

The Holy Alliance

After the Second Treaty of Paris, Alexander I, inspired by piety, formed the Holy Alliance, which was supposed to bring about a peace based on Christian love to the monarchs and peoples of Europe. It is possible to see in the alliance the beginnings of a European federation, but it would have been a federation with evangelical and ecumenical, rather than political, foundations.

The idealistic Tsar's vision came to a sad end, for the alliance became a league of monarchs against their peoples. Its members, following up the congress with additional meetings at Aix-la-Chapelle, Troppau, Laibach (Ljubljana), and Verona, revealed themselves as the champions of despotism and the defenders of an order maintained by arms. When a series of uprisings against despotic regimes in Italy and Spain broke out, the "holy allies" responded with bloody repression. Alexander himself was badly shaken by the mutiny of his Semyonovsky regiment and thought he detected the presence of revolutionary radicalism even in his own empire.

This marked the end of his liberal dreams, for, from then on, all revolt appeared to him as a rebellion against God. He shocked Russia by refusing to support the Greeks, his coreligionists, when they rose against Turkish tyranny, maintaining they were rebels like any others. The Austrian chancellor, Prince Metternich, to whom the Tsar abandoned the conduct of European affairs, shamelessly exploited Alexander's state of mind.

After his return to Russia, he left everything in Arakcheyev's hands. For Alexander, it was a period of lassitude, discouragement, and dark thoughts. For Russia, it was a period of reaction, obscurantism, and struggle against real and imagined subversion. Alexander thought he saw "the reign of Satan" everywhere. In opposition, secret societies spread, composed of young men, mostly from the military, who sought to regenerate and liberalize the country. Plots were made. Alexander was warned of them, but he refused to act decisively. His crown weighed heavily on him, and he did not hide from his family and close friends his desire to abdicate.

The Empress was ill, and Alexander decided to take her to Taganrog, on the Azov Sea. This dismal, windy townlet was a strange watering place. The royal pair, however, who had been so long estranged, enjoyed a calm happiness there. Soon after, during a tour of inspection in the Crimea, Alexander contracted pneumonia or malaria and died on his return to Taganrog, on November 19, 1825.

Death and final legend

The Tsar's sudden death, his mysticism, and the bewilderment and the blunders of his entourage all went into the creation of the legend of his "departure" to a Siberian retreat. The refusal to open the Tsar's coffin after his death has only served to deepen the mystery.

The end of Alexander's reign was tragic. He did not fulfill the promise of his glorious beginning, for, although he was a warmhearted and amiable sovereign, he remained a dilettante, with his head full of illusions and unfinished projects. Humanitarian and romantic, he was more a European than a Russian and thus better suited to deal with foreign affairs than to govern his empire. After reaching the peak of his energy and power, he fell back as though exhausted by the effort. Melancholy and disappointed in himself, he became indifferent toward the real world and turned to religious meditation. Russia was the loser.

BIBLIOGRAPHY. N. SCHILDER, *Imperator Aleksandr Pervy*, 4 vol. (1890–1904), in Russian; GRAND-DUKE NICOLAS MIKHAILOVITCH, *Le Tsar Alexandre Ier*, 2 vol. (1900); and A. VANDAL, *Napoléon et Alexandre*, 3 vol. (1891–96), though old, are still the three most important biographies written since the authors—and especially the two first mentioned—had access to many State and family papers. See also the *Correspondance de l'Empereur Alexandre avec sa soeur, la Grand-Duchesse Catherine Pavlovna* (1910; Eng. trans., *Scenes of Russian Court Life: Being the Correspondence of Alexander I with His Sister Catherine*, 1917). Two good studies in English are MARTHA E. ALMEDINGEN, *The Emperor Alexander I* (1964); and MICHAEL JENKINS, *Arakcheev: Grand Vizier of the Russian Empire* (1969).

(D.Ol.)

Alexander II of Russia

Of Alexander II of Russia, the emperor who for his part in freeing the serfs became known as the "Tsar Liberator," it can be said that he was a great historical figure without being a great man, that what he did was more important than what he was. His Great Reforms indeed rank in importance with those of Peter the Great and



Alexander II, portrait by an unknown artist, 19th century. In the collection of Mrs. Merriweather Post, Hillwood, Washington, D.C.
By courtesy of Mrs. Merriweather Post, Hillwood, Washington, D.C.

Lenin, yet the impact of his personality was much inferior to theirs. The Tsar's place in history—a substantial one—is due almost entirely to his position as the absolute ruler of a vast empire at a critical stage in its development. Under Russia's system of autocratic government, far-reaching changes came to be carried out in many departments of Russian life, but they were shaped far more by the demands of the forces pressing for change than by his own initiative.

Early influences

The future Alexander II was born Aleksandr Nikolayevich in Moscow on April 17 (old style), 1818, the eldest son of the grand duke Nikolay Pavlovich (who, in 1825, became the emperor Nicholas I) and his wife, Alexandra Fyodorovna (who, before her marriage to the Grand Duke and baptism into the Orthodox Church, had been the princess Charlotte of Prussia). Alexander's youth and early manhood were overshadowed by the overpowering personality of his dominating father, from whose authoritarian principles of government he was never to free himself. But at the same time, at the instigation of his mother, responsibility for the boy's moral and intellectual development was entrusted to the poet Vasily Zhukovsky, a humanitarian liberal and romantic. Alexander, a rather lazy boy of average intelligence, retained throughout his life traces of his old tutor's romantic sensibility. The tensions created by the conflicting influences of Nicholas I and Zhukovsky left their mark on the future emperor's personality. Alexander II, like his uncle Alexander I before him (who was educated by a Swiss republican tutor, a follower of Rousseau), was to turn into a "liberalizing," or at any rate humanitarian, autocrat.

Alexander succeeded to the throne at the age of 36, following the death of his father in February 1855, at the height of the Crimean War. The war had revealed Russia's glaring backwardness in comparison with more advanced nations like England and France. Russian defeats, which had set the seal of final discredit on the oppressive regime of Nicholas I, had provoked among Russia's educated elite a general desire for drastic change. It was under the impact of this widespread urge that the Tsar embarked upon a series of reforms designed, through "modernization," to bring Russia into line with the more advanced Western countries.

Among the earliest concerns of the new emperor (once peace had been concluded in Paris in the spring of 1856 on terms considered harsh by the Russian public) was the improvement of communications. Russia at this time had only one railway line of significance, that linking the two capitals of St. Petersburg and Moscow. At Alexander's accession there were less than 600 miles (965 kilometres) of track; when he died in 1881, some 14,000 miles (22,525 kilometres) of railway were in operation. In Russia, as elsewhere, railway construction, in its turn, meant a general quickening of economic life in a hitherto predominantly feudal agricultural society. Joint-stock companies developed, as did banking and credit institutions. The movement of grain, Russia's major article of export, was facilitated.

Emancipation of the serfs

The same effect was achieved by another measure of modernization, the abolition of serfdom. In the face of bitter opposition from landowning interests, Alexander II, overcoming his natural indolence, took an active personal part in the arduous legislative labours that on February 19, 1861, culminated in the Emancipation Act. By a stroke of the autocrat's pen, tens of millions of human chattels were given their personal freedom. By means of a long-drawn-out redemption operation, moreover, they were also endowed with modest allotments of land. Although for a variety of reasons the reform failed in its ultimate object of creating an economically viable class of peasant proprietors, its psychological impact was immense. It has been described as "the greatest social movement since the French Revolution" and constituted a major step in the freeing of labour in Russia. Yet at the same time, it helped to undermine the already shaken economic foundations of Russia's landowning class.

The abolition of serfdom brought in its train a drastic overhaul of some of Russia's archaic administrative institutions. The most crying abuses of the old judicial system were remedied by the judicial statute of 1864. Russia, for the first time, was given a judicial system that in important respects could stand comparison with those of Western countries (in fact, in many particulars it followed that of France). Local government in its turn was remodelled by the statute of 1864, setting up elective local assemblies known as *zemstva*. Their gradual introduction extended the area of self-government, improved local welfare (education, hygiene, medical care, local crafts, agronomy), and brought the first rays of enlightenment to the benighted Russian villages. Before long *zemstvo* village schools powerfully supported the spread of rural literacy. Meanwhile, Dmitry Milyutin, an enlightened minister of war, was carrying out an extensive series of reforms affecting nearly every branch of the Russian military organization. The educative role of military service was underlined by a marked improvement of military schools. The army statute of 1874 introduced conscription for the first time, making young men of all classes liable to military service.

Modernization of Russia

The keynote of these reforms—and there were many lesser ones affecting various aspects of Russian life—was the modernization of Russia, its release from feudalism, and acceptance of Western culture and technology. Their aim and results were the reduction of class privilege, humanitarian progress, and economic development. Moreover, Alexander, from the moment of his accession, had instituted a political "thaw." Political prisoners had been released, and Siberian exiles allowed to return. The personally tolerant Emperor had removed or mitigated the heavy disabilities weighing on religious minorities, particularly Jews and sectarians. Restrictions on foreign travel had been lifted. Barbarous medieval punishments were abolished. The severity of Russian rule in Poland was relaxed. Yet, notwithstanding these measures, it would be wrong, as is sometimes done, to describe Alexander II as a liberal. He was in fact a firm upholder of autocratic principles, sincerely convinced both of his duty to maintain the God-given autocratic power he had inherited and of Russia's unreadiness for constitutional or representative government. Practical experience only strengthened these convictions. Thus, the relaxation of Russian rule in Poland led to patriotic street demonstrations, attempted assassinations, and, finally, in 1863, to a national uprising that was only suppressed with some difficulty—and under threat of Western intervention on behalf of the Poles. Even more serious, from the Tsar's point of view, was the spread of nihilistic doctrines among Russian youth, producing radical leaflets, secret societies, and the beginnings of a revolutionary movement. The government, after 1862, had reacted increasingly with repressive police measures. A climax was reached in the spring of 1866, when Dmitry Karakozov,

Failure of liberal policies

a young revolutionary, attempted to kill the Emperor. Alexander—who bore himself gallantly in the face of great danger—escaped almost by a miracle. The attempt, however, left its mark by completing his conversion to conservatism. For the next eight years, the Tsar's leading minister—maintaining his influence at least partially by frightening him with real and imaginary dangers—was Pyotr Shuvalov, the head of the secret police.

The period of reaction following Karakozov's attempt coincided with a turning point in Alexander's personal life, the beginning of his liaison with princess Yekaterina Dolgorukaya, a young girl to whom the aging Emperor had become passionately attached. The affair, which it was impossible to conceal, absorbed the Tsar's energies while weakening his authority both in his own family circle (his wife, the former princess Marie of Hesse-Darmstadt, had borne him six sons and two daughters) and in St. Petersburg society. His sense of guilt, moreover, made him vulnerable to the pressures of the Pan-Slav nationalists, who used the ailing and bigoted Empress at their advocate when in 1876 Serbia became involved in war with the Ottoman Empire. Although decidedly a man of peace, Alexander became the reluctant champion of the oppressed Slav peoples and in 1877 finally declared war on Turkey. Following initial setbacks, Russian arms eventually triumphed; and, early in 1878, the vanguard of the Russian armies stood encamped on the shores of the Sea of Marmora. The prime reward of Russian victory—seriously reduced by the European powers at the Congress of Berlin—was the independence of Bulgaria from Turkey. Appropriately, that country still honours Alexander II among its "founding fathers" with a statue in the heart of its capital, Sofia.

Comparative military failure in 1877, aggravated by comparative diplomatic failure at the conference table, ushered in a major crisis in the Russian state. Beginning in 1879, there was a resurgence of revolutionary terrorism soon concentrated on the person of the Tsar himself. Following unsuccessful attempts to shoot him, to derail his train, and finally to blow up the Winter Palace in St. Petersburg itself, Alexander, who under personal attack had shown unflinching courage based on a fatalist philosophy, entrusted supreme power to a temporary dictator. The minister of the interior, Count Mikhail Loris-Melikov, was charged with exterminating the terrorist organization (calling itself People's Will) while at the same time conciliating moderate opinion, which had become alienated by the repressive policies pursued since 1866. At the same time, following the death of the Empress in 1880, the Tsar had privately married Yekaterina Dolgorukaya (who had borne him three children) and was planning to proclaim her his consort. To make this step palatable to the Russian public, he intended to couple the announcement with a modest concession to constitutionalist aspirations. There were to be two legislative commissions including indirectly elected representatives. This so-called Loris-Melikov Constitution, if implemented, might possibly have become the germ of constitutional development in Russia. But on the day when, after much hesitation, the Tsar finally signed the proclamation announcing his intentions (March 1, 1881), he was mortally wounded by bombs in a plot sponsored by People's Will.

Assessment

Alexander's achievement, notwithstanding its limitations, was considerable. The modernization of Russian institutions, though piecemeal, was extensive. In Alexander's reign, Russia built the base needed for emergence into capitalism and industrialization later in the century. At the same time, Russian expansion, especially in Asia, steadily gathered momentum. The sale of Alaska to the U.S. in 1867 was far outweighed in importance by the acquisition of the Maritime Provinces from China (1858 and 1860) and the founding of Vladivostok as Russia's far eastern capital (1860), the definitive subjugation of the Caucasus (in the 1860s), and the conquest of central Asia (Khiva, Bokhara, Turkestan) in the 1870s. The contribution of the reign to the development of what was to be described as Russia's "cotton imperialism" was immense. Here also, the reign of Alexander paved the way for the later phases of Russian imperialism in Asia.

Alexander's importance lies chiefly in his efforts to assist Russia's emergence from the past. To some extent, he was, of course, the representative of forces—intellectual, economic, and political—that were stronger than himself or, indeed, any one individual. After the Crimean War, the modernization of Russia had indeed become imperative if Russia was to retain its position as a major European power. But even within the context of a wider movement, the role of Alexander II, through his position as autocratic ruler, was a highly important one. The Great Reforms, both in what they achieved and in what they failed to do, bear the imprint of his personality. Unfortunately, however, by placing great power in the hands of the influential reactionary minister K.P. Pobedonostsev—whom he appointed minister for church affairs (procurator of the Holy Synod) and entrusted with the education of his son and heir, the future Alexander III—Alexander II, perhaps unwittingly, did much to frustrate his own reforming policies and to set Russia finally on the road to revolution.

BIBLIOGRAPHY. S.S. TATISHCHEV, *Imperator Aleksandr II*, 2 vol. (1903), the standard life of Alexander II, is the pre-revolutionary official biography. The fullest modern biography is C. DE GRUNWALD, *Le Tsar Alexandre II et son temps* (1963). A short, concise life of the emperor is W.E. MOSSE, *Alexander II and the Modernization of Russia* (1958). Two popular biographies are S. GRAHAM, *Alexander II: Tsar of Russia* (1935); and MARTHA E. ALMEDINGEN, *The Emperor Alexander II* (1962).

(W.E.Mo.)

Alexander III of Russia

Alexander III (Aleksandr Aleksandrovich), emperor of Russia from 1881 to 1894, was born in St. Petersburg on March 10 (new style; February 26, old style), 1845 second son of Alexander II and of Maria Aleksandrovna (Marie of Hesse-Darmstadt). In disposition he bore little resemblance to his softhearted, impressionable father and still less to his refined, chivalrous, yet complex granduncle, Alexander I. He gloried in the idea of being of the same rough texture as the great majority of his subjects. His straightforward manner savoured sometimes of gruffness, while his unadorned method of expressing himself harmonized well with his roughhewn, immobile features.

Early years

During the first 20 years of his life, Alexander had no prospect of succeeding to the throne. He received only the perfunctory training given to grand dukes of that period, which did not go much beyond primary and secondary instruction, acquaintance with French, English, and German, and military drill. When he became heir apparent on the death of his elder brother Nikolay in 1865, he began to study the principles of law and administration under the jurist and political philosopher K.P. Pobedonostsev, who influenced the character of his reign by instilling into his mind hatred for representative government and the belief that zeal for Orthodoxy ought to be cultivated by every tsar.

The tsarevitch Nikolay, on his deathbed, had expressed a wish that his fiancée, Princess Dagmar of Denmark, thenceforward known as Maria Fyodorovna, should marry his successor. The marriage proved a most happy one. During his years as heir apparent—from 1865 to 1881—Alexander let it be known that certain of his ideas did not coincide with the principles of the existing government. He deprecated undue foreign influence in general and German influence in particular. His father, however, occasionally ridiculed the exaggerations of the Slavophiles and based his foreign policy on the Prussian alliance. The antagonism between father and son first appeared publicly during the Franco-German War, when the tsar sympathized with Prussia and the tsarevitch Alexander with the French. It reappeared in an intermittent fashion during the years 1875–79, when the disintegration of the Ottoman Empire posed serious problems for Europe. At first the tsarevitch was more Slavophile than the government, but he was disabused of his illusions during the Russo-Turkish War of 1877–78, when he commanded the left wing of the invading army. He was a conscientious commander, but he was mortified when



Alexander III, portrait by an unknown artist, 19th century. In the collection of Mrs. Merriweather Post, Hillwood, Washington, D.C.
By courtesy of Mrs. Merriweather Post, Hillwood, Washington, D.C.

most of what Russia had obtained by the Treaty of San Stefano was taken away at the Congress of Berlin under the chairmanship of the German chancellor Otto von Bismarck. To this disappointment, moreover, Bismarck shortly afterward added the German alliance with Austria for the express purpose of counteracting Russian designs in eastern Europe. Although the existence of the Austro-German alliance was not disclosed to the Russians until 1887, the tsarevitch reached the conclusion that for Russia the best thing to do was to prepare for future contingencies by a radical scheme of military and naval reorganization.

On March 13 (N.S.; March 1, O.S.), 1881, Alexander II was assassinated, and the following day autocratic power passed to his son. In the last years of his reign, Alexander II had been much disturbed by the spread of Nihilist conspiracies. On the very day of his death he signed an *ukaz* creating a number of consultative commissions that might have been transformed eventually into a representative assembly. Alexander III cancelled the *ukaz* before it was published and in the manifesto announcing his accession stated that he had no intention of limiting the autocratic power he had inherited. All the internal reforms that he initiated were intended to correct what he considered the too liberal tendencies of the previous reign. In his opinion, Russia was to be saved from anarchical disorders and revolutionary agitation not by the parliamentary institutions and so-called liberalism of western Europe but by the three principles of Orthodoxy, autocracy, and *narodnost*, or belief in the Russian people.

Attempt to create a unified nation

Alexander's political ideal was a nation containing only one nationality, one language, one religion, and one form of administration; and he did his utmost to prepare for the realization of this ideal by imposing the Russian language and Russian schools on his German, Polish, and Finnish subjects, by fostering Orthodoxy at the expense of other confessions, by persecuting the Jews, and by destroying the remnants of German, Polish, and Swedish institutions in the outlying provinces. In the other provinces he clipped the feeble wings of the *zemstvo* (an elective local administration resembling the county and parish councils in England) and placed the autonomous administration of the peasant communes under the supervision of landed proprietors appointed by the government. At the same time, he sought to strengthen and centralize the imperial administration and to bring it more under his personal control. In foreign affairs he was emphatically a man of peace but not a partisan of the doctrine of peace at any price. Though indignant at the conduct of Bismarck toward Russia, he avoided an open rupture with Germany and even revived for a time the Alliance of the Three Emperors between the rulers of Germany, Russia, and Austria. It was only in the last years of his reign, especially after the accession of William II as German emperor in 1888, that Alexander adopted a more hostile attitude toward Germany. The termination of the Russo-German alliance in 1890 drove Alexander reluctantly into an alliance with France, a country that he strongly disliked as the breeding place of revolutions. In Central Asian affairs he followed the traditional policy of gradually extending Russian domination without provoking a conflict with Great Britain, and he never allowed the bellicose partisans of a forward policy to get out of hand.

As a whole, Alexander's reign cannot be regarded as one of the eventful periods of Russian history; but it is arguable that under his hard, unsympathetic rule the country made some progress. He died at Livadiya, in the Crimea, on November 1 (N.S.; October 20, O.S.), 1894.

BIBLIOGRAPHY. HUGH SETON-WATSON, *The Russian Empire, 1801–1917* (1967), massive information clearly organized and objectively presented; *The Decline of Imperial Russia, 1855–1914* (1952), good factual record with little attention to personalities; M.T. FLORINSKY, *Russia: A History and an Interpretation*, vol. 2 (1955), a compact detailed account, factual and analytical; K.P. POBEDONOSTSEV, *Mémoires politiques, correspondance officielle et documents inédits relatifs à l'histoire du règne de l'empereur Alexandre III* (1927), indispensable revealing documents; B.E. NOLDE, *L'Alliance franco-russe* (1936), an excellent survey, based on primary sources, on the evolution of Russia's foreign policy.

(M.T.F.)

Alexander Nevsky

Alexander Nevsky, a 13th-century Russian leader, halted the eastward drive of the Germans and Swedes but collaborated with the Mongols in imposing their rule on Russia. As grand prince of Vladimir he unified, with the conquerors' aid, the numerous politically fragmented and feuding independent northern Russian principalities.

Prince of Novgorod

Alexander was born around 1220, the son of Yaroslav II Vsevolodovich, grand prince of Vladimir, the foremost among the Russian rulers. In 1236 Alexander was elected prince—a figure who functioned as little more than military commander—of the city of Novgorod. In 1239 he married the daughter of the Prince of Polotsk.

When in 1240 the Swedes invaded Russia to punish the Novgorodians for encroaching on Finnish tribes and to bar Russia's access to the sea, Alexander defeated the Swedes at the confluence of the Rivers Izhora and Neva —hence his name Nevsky, "of the Neva." His standing enhanced by his victory, he apparently began to intervene in the affairs of the city and was expelled by the Novgorodians a few months later.

Urged by Pope Gregory IX to "Christianize" the Baltic region, the Teutonic Knights shortly thereafter invaded Russia, and Novgorod invited Alexander to return. After a number of battles, Alexander decisively defeated the Germans in the famous "massacre on the ice" in April 1242 on a narrow channel between Lakes Chud (Peipus) and Pskov. Alexander, who continued to fight both the Swedes and Germans and eventually stopped their eastward expansion, also won many victories over the pagan Lithuanians and the Finnic peoples.

In the east, however, Mongol armies were conquering most of the politically fragmented Russian lands. Alexander's father, the grand prince Yaroslav, agreed to serve the new rulers of Russia but died in September 1246 of poisoning after his return from a visit to the Great Khan in Mongolia. When, in the ensuing struggle for the grand princely throne, Alexander and his younger brother Andrew appealed to Khan Batu of the Mongol Golden Horde, he sent them to the Great Khan. Violating Russian customs of seniority, the Great Khan appointed Andrew grand prince of Vladimir and Alexander prince of Kiev—probably because Alexander was Batu's favourite and Batu was in disfavour with the Great Khan.

Collaboration with Mongols

When Andrew began to conspire against the Mongol overlords with other Russian princes and western nations, Alexander went to Saray on the Volga and denounced his brother to Sartak, Batu's son, who sent an army to depose Andrew and installed Alexander as grand prince.

Henceforth, for over a century, no northeastern Russian prince challenged the Mongol conquest. Alexander proceeded to restore Russia by building fortifications and churches and promulgating laws. As grand prince, he continued to rule Novgorod through his son Vasily, thus changing the constitutional basis of rule in Novgorod from personal sovereignty by invitation to institutional sovereignty by the principal Russian ruler. When, in 1255, Novgorod, tiring of grand princely rule, expelled Vasily and invited an opponent of Mongol hegemony, Alexander assembled an army and reinstalled his son.

In 1257 the Mongols, in order to levy taxes, took a census in most of Russia. It encountered little opposition, but when news of the impending enumeration reached Novgorod an uprising broke out. In 1258 Alexander, fearing that the Mongols would punish all of Russia for the Novgorodian revolt, helped force Novgorod to submit to the census and to Mongol taxation. This completed the process of imposing the Mongol yoke on northern Russia.

In 1262 uprisings broke out in many towns against the Muslim tax farmers of the Golden Horde, and Alexander made a fourth journey to Saray to avert reprisals. He succeeded in his mission, as well as in obtaining exemption for Russians from a draft of men for a planned invasion of Iran. Returning home, Alexander died on November 14, 1263, in Gorodets on the Volga. After his death Russia once more disintegrated into many feuding principalities. His personal power, based upon support of the princes, boyars, and clergy, as well as the fear of Mongols, could not be transmitted to any other man, including his weak sons.

Assessment

Whether Alexander was a quisling in his dealings with the Mongol conquerors is a question seldom posed by Russian historians, because some Russian princes had for centuries concluded alliances with Turkic steppe nomads in order to gain advantage in domestic rivalries. Because Alexander was a willing collaborator, he may have reduced the common people's suffering by interceding for them with the Khan. He was supported by the church, which thrived under Mongol protection and tax exemption and feared the anti-Mongol princes who negotiated with the papacy. For these reasons, Alexander by 1381 was elevated to the status of a local saint and was canonized by the Russian Orthodox Church in 1547. Alexander's son Daniel founded the house of Moscow, which subsequently reunited the northern Russian lands and ruled until 1598. Alexander was one of the great military commanders of his time, who protected Russia's western frontier against invasion by Swedes or Germans. This image of him was popular in northwestern Russia and has in succeeding centuries been adduced for propaganda purposes. Thus, after the conclusion of the war with Sweden, the Order of Alexander Nevsky was created in 1725, and during World War II (in July 1942), when Germany had deeply penetrated into Russia, Stalin pronounced Alexander Nevsky a national hero and established a military order in his name.

BIBLIOGRAPHY. There is no book-length study of Nevsky in English. Additional information may be found in A.E. PRESNIAKOV, *The Formation of the Great Russian State: A Study of Russian History in the Thirteenth to Fifteenth Centuries* (1970; orig. pub. in Russian, 1918); and GEORGE VERNADSKY, *A History of Russia*, vol. 3, *The Mongols and Russia* (1953).

(R.He.)

Alexandria

Alexandria (Arabic al-Iskandarīyah), capital of Egypt from its founding by Alexander the Great in 332 BC to AD 642, when it was subdued by the Arabs, is Egypt's second largest city and chief seaport. The city's built-up area—133 square miles—in the early 1970s contained most of the more than 2,000,000 residents of its larger *muḥāfaẓah* (governorate [1,035 square miles]). It lies on the Mediterranean Sea at the western edge of the Nile Delta, 129 miles northwest of Cairo.

History. *Founding and Greek period.* The city was founded by Alexander the Great when he spent the winter of 332–331 BC in Egypt. To the ancient Egyptian town of Rakotis (dating back to 1500 BC), situated at the mouth of the westernmost Nile distributary (long since silted up), he added a western suburb, called Neapolis, which he intended to supplant Naukratis as a centre of Greek culture in Egypt and to serve as a naval base. Alexander occupied the island of Pharos and had a walled city marked out on the mainland to include Rakotis. A few months later he left Egypt and did not return to his city except for entombment. His viceroy, Cleomenes, continued the creation of Alexandria. Inheriting the trade of the Phoenician city of Tyre and becoming the centre of increased commerce between Europe and the East, in less than a century the city surpassed even Carthage in size. A centre not only of Hellenism but of Semitic learning as well, the city became a focal point of the highest developments of Greek scholarship and science, containing the great research institute museum founded by the Ptolemies and a library reputed to contain 500,000 volumes. The Septuagint translation of the Old Testament into Greek was produced there. These institutions survived until the main museum and library were destroyed during the civil war of the 3rd century AD; a subsidiary library was burned by Christians in AD 391.

The Alexandrian libraries

The Roman period. As a free Greek city Alexandria retained its own senate into Roman times, when it became the leading provincial capital, second only to Rome. Although under Roman influence for more than a century, it was not until 80 BC that it passed formally under Roman jurisdiction. Key to the granary of the empire and focal pivot for imperial ambitions, it was the scene of many Roman dramas. In Alexandria, Caesar was courted by Cleopatra in 47 BC and then mobbed by the masses. He was followed by Mark Antony, for whose favour the city paid dearly to Octavian (later Augustus), who founded his own rival city nearby in suburban ar-Raml (ar-Ramlah) and reduced Alexandria's function to exporting grain for his imperial coffers. The city survived, despite these and later afflictions (such as a general massacre at the emperor Caracalla's behest in AD 215), and, with the conversions in the empire to Christianity, it took on new importance as a centre of theology and church government. Tradition has it that St. Mark made his first Christian convert in Alexandria in AD 45, thus opening a new phase in history. Arianism (a Christian theological movement initiated by Arius in the 4th century that holds that Christ is not divine but a created being) was formulated in Alexandria, and there St. Athanasius, bishop of Alexandria (died 373), opponent of both heresy and pagan reaction, worked and triumphed.

The Islāmic period. From its inception, Alexandria had represented an alien culture, and, as the peace of the empire broke up, its hold over the Egyptian interior grew weaker and it declined in commercial importance and in population. In 616 it was taken from the Byzantines by the Persians, and in 642 by the Arabs, who, meeting their only significant opposition there, had to retake it in 646. The establishment of a new Islāmic capital at Fusṭāṭ (near what was later to be the site of Cairo), and the preference of desert Arabs for land routes reduced Alexandria to a minor naval base. The city knew brief periods of revival, however, whenever interest in the Mediterranean was rekindled (notably during 12th-century crusader incursions and 13th-century Mamlūk sea campaigns and when interest heightened in the East–West spice trade, over which Egypt had a monopoly until the end of the 15th century). Throughout the Middle Ages, the Islāmic settlement, about a mile long and more than a half a mile wide, was confined to the mainland on a fraction of the area formerly occupied by the Roman city. Its walls, rebuilt in the 13th and 14th centuries, were not removed until 1811. The final blow to Alexandria was dealt by the realigning of the spice routes after the Portuguese discovered the southern sea route to India in 1498. This commercial disaster was followed by the Turkish conquest of Egypt (1517). Within the next century the settlement within the walls became deserted, and a new "Turkish City" was built on the neck of land

(formed by silting) between the mainland and the former island of Pharos. Neglected by the Turks, the canal connecting Alexandria with the Rosetta branch of the Nile closed, and Nile-borne commerce was diverted to Rosetta (Arabic Rashīd). By 1777 Alexandria had become an insignificant fishing town with a population of 6,000; it was further reduced in size and importance because of battles between Mamlūks, French, and British during Napoleon's Egyptian campaign fought from 1798 onward.

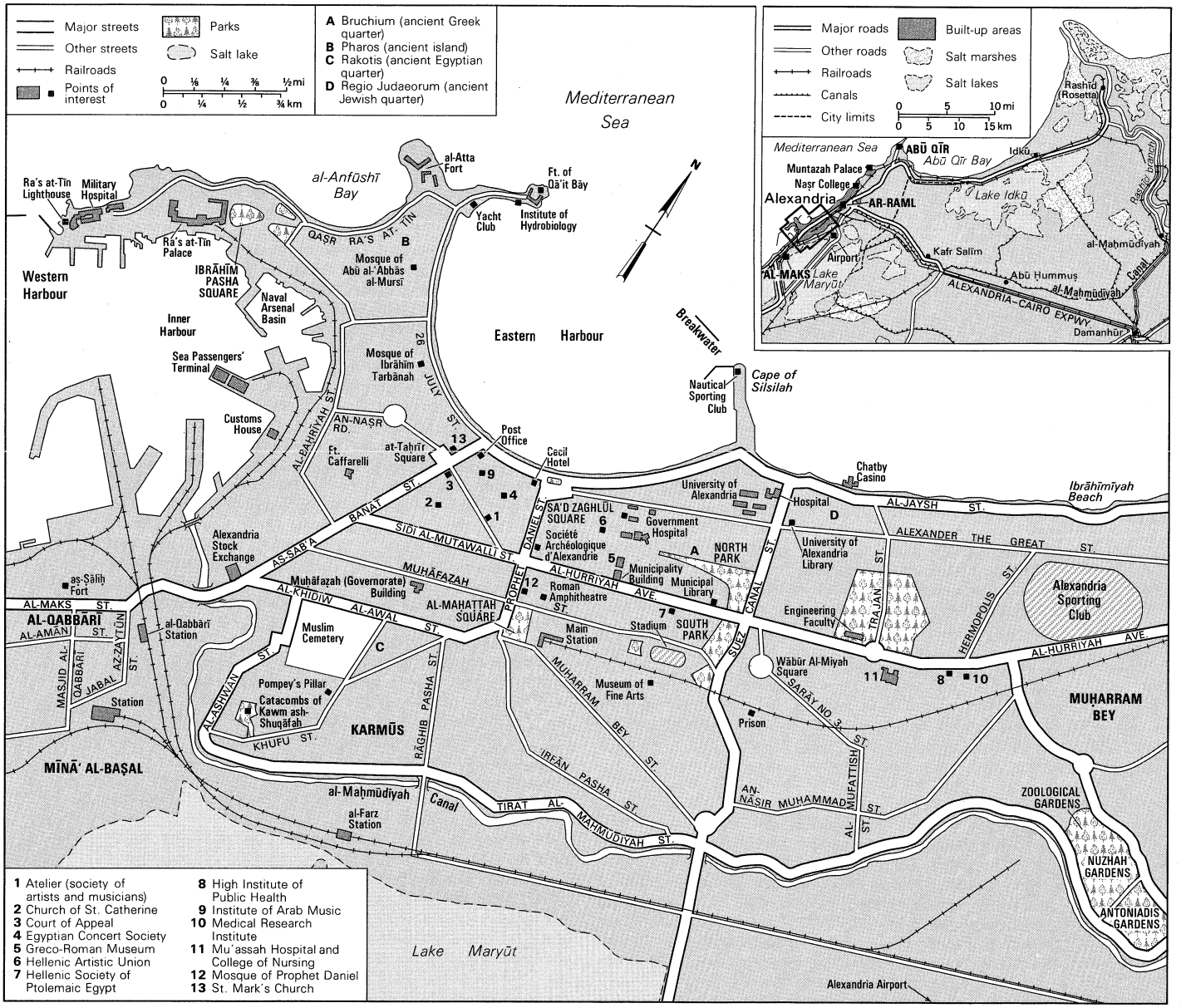
The city's modern renascence

Modern period. Alexandria's rebirth began when Muḥammad ʿAlī Pasha was Ottoman viceroy of Egypt from 1805 to 1848. To further his ambitions to modernize Egypt, he reopened Alexandria's access to the Nile via the 45 mile-long Maḥmūdīyah Canal (completed between 1818 and 1820), built new docks and an arsenal (1828 to 1833), and located industry there. Particularly after the completion of a railroad linking Suez with Alexandria via Cairo, the opening of the Suez Canal (1869), and the construction of new harbour works (1871–78), the rapidly growing city attracted numerous foreign merchants and speculators who enjoyed legally sanctioned privileges to exploit both the expanding traffic to India and the export trade to Europe of Egyptian cotton.

Population increased from about 12,000 in 1828 to 233,000 by 1882, when the city was bombarded and then invaded by the United Kingdom, thus beginning more than 50 years of colonial rule. Throughout the subsequent period, Alexandria retained its position as the second (summer) capital of Egypt and as the stronghold of foreign commercial interests. In World War I Alexandria was the chief Allied naval base in the eastern Mediterranean. Between the two world wars the city shared with Cairo a burst of industrial development, and in the 1920s the population rose to more than 500,000. World War II heightened Alexandria's economic and military importance, and, despite repeated German bombardment, internal migration raised the population to nearly 1,000,000 by the late 1940s. British forces left the city in 1946.

During the Revolution of 1952 Alexandria played its role; from there the deposed King Farouk was saluted as his boat steamed out of the harbour. By the early 1970s the steadily growing population had doubled, and, in addition to functioning as the country's chief port and major industrial centre (20 percent of Egypt's large industrial firms and 25 percent of its industrial workers are in Alexandria), the city had become an increasingly popular seaside resort.

The modern city. *The city site.* Alexandria is built on the strip of land separating a low salt lake (Maryūt, or



Central Alexandria and (inset) its metropolitan area.

Mareotis, eight feet below sea level and now partly drained and cultivated) from the Mediterranean and on a T-shaped promontory. The stem of the T was originally a mole (the Heptastadium) leading to the island of Pharos, which formed the crosspiece of the T. In the course of centuries the mole was widened by the accumulation of silt until it became an isthmus half a mile wide, on which a part of the city is built. The isthmus forms two harbours: the old, or great, harbour on the east and the present harbour on the west.

The Pharos

On the eastern tip of the original island stood the great lighthouse, one of the Seven Wonders of the World, known as the Pharos of Alexandria. It was begun by the first Ptolemy and completed by his successor about 280 BC. The prototype of all the lighthouses in the world, it was reputed to be 400 feet (122 metres) high. It was still standing in the 12th century, but in 1477 the sultan Qā'it Bāy built a fort and mosque from its ruins. The eastern tip, or cape, is still known as Pharos or as Qā'itbāy, from the fort and mosque, while at the western end is the cape of Ra's at-Tīn (Cape of Figs).

Plan of the city. The city, stretching in a long, narrow band along the coast, originally included the royal precinct and quarters for the Greek, Egyptian, and Jewish inhabitants. The customs house and the chief warehouses are by the western harbour, but the principal buildings of the city are in the eastern and southeastern quarters. One of the main features of modern Alexandria is the seacoast road, al-Jaysh (Army) Street, completed in 1934, when it was called the Corniche. It runs eastward for about 15 miles, with bathing beaches stretching along most of its length.

Muḥammad ʿAlī Square serves as the focal point of the city. Known in the 19th century first as the Place des Consuls and then as the Grand Square, it had always been colloquially called al-Manshia. In 1952 its name was officially changed to Mīdān at-Taḥrīr (Liberation Square). On its northern side is the Anglican Church of St. Mark's, built on land originally given to the British community by Muḥammad ʿAlī in 1839 as a free grant. Other imposing buildings, including the law courts, border the square.

Farther east and giving onto the sea is the Saʿd Zaghlūl Square, named for the nationalist leader who died in 1927. Nearby are a large hotel, the Cecil, and the regional office for the Eastern Mediterranean of the World Health Organization. Southeast from Mīdān at-Taḥrīr is the shopping and commercial part of the town.

The northern quarter of Alexandria is Oriental in style. The yacht club that overlooked the harbour at Ra's at-Tīn is now the Egyptian naval officers' mess, and the yacht club has been moved to the eastern harbour. Also located on the promontory is the palace built by Muḥammad ʿAlī, which, since the abdication of King Farouk in 1952, has been open to the public. The eastern bay is rocky and exposed and is used only by fishing craft and yachts. The large suburb of ar-Raml stretches eastward from Alexandria in an unbroken line for approximately 16 miles.

Between Mīdān at-Taḥrir and the western harbour, one of the poorest quarters of the city, is an open space in the centre of which is Ft. Caffarelli, so named because it was fortified by François Caffarelli, a general in Napoleon's army.

The port. The harbour is on the west of Pharos and is partly formed by a breakwater (built 1871–73 and extended 1906–07) two miles long. Opposite is Ra's at-Tīn, with its 180-foot-high lighthouse built by Muḥammad ʿAlī. Another breakwater starts from the western (al-Qabbārī) side; the opening between the two is about half a mile. The enclosed water is divided into an outer and inner harbour by a mole 1,000 yards long. The inner harbour is lined for two and one-half miles by quays that can accommodate ships with drafts of up to 28 feet. New wharves for coke and grain imports were completed in 1960. The outer harbour, which is about three times larger than the inner harbour, has a dry dock and quays and jetties.

The city is the headquarters of most of the merchants and companies engaged in the development of the Nile Delta. In the early 1970s, more than 90 percent of the total exports and imports of the country passed through Alexandria's port.

Transport and communications. Alexandria is linked to other Egyptian centres by railway, telegraph and telephone lines, roads, and air services. It also has a good canal connection with the Nile. Two main roads run to Cairo, one through the western desert and the Alexan-

Carl Frank



At-Taḥrīr Square on the Mediterranean coast of Alexandria. The statue in the square is of Muhammad ʿAlī Pasha.

dria–Cairo expressway through the delta. A third road leads west along the coastal desert, past al-ʿAlamayn to the Libyan frontier, and connecting with the North African coastal road. Rail connections are with Cairo, Suez, the Nile Delta, and Libya.

The civil airport is situated one and three-quarter miles south of the city. Electric trams and buses, whose lines all converge at Mīdān at-Taḥrīr, link the suburbs and different quarters of the town together. Buses also run to outlying towns, along the coastal desert road to as-Sallūm on the Libyan frontier, and to Cairo through delta and desert.

Climate. In Alexandria the climate is dominated by the Mediterranean. From December to March the weather is variable, with predominantly cool, sunny days but occasional heavy falls of rain or hail and stormy winds. The coldest month is January with a mean day temperature of 64° F (18° C). In summer the heat is moderated by sea breezes, but there is much humidity; in August, the hottest month, the average day temperature is 87° F (31° C).

Population. The population of Alexandria in 1966 was 1,801,000 and by the early 1970s had grown to more than 2,000,000. Although the city once contained a sizable minority of foreigners and Jews, by 1947 these groups constituted only about 10 percent (of which 31,000 were Greeks, 21,000 Jews, 12,000 Italians, and 10,000 British) of the population. Emigration of minorities has continued since then. Many Jews departed in 1948, 1956, and 1967; Greeks emigrated when adversely affected by post-revolutionary decrees placing business firms in Egyptian hands; and western European minorities declined after the 1956 invasion of Suez. Foreigners now constitute less than 2 percent of the population. The major religious subdivisions are between Egyptian Muslims (about 85 percent) and Egyptian Copts.

Buildings and monuments. Little of ancient Alexandria remains. In the southern part of the city are Pompey's Pillar at the site of the Sarapeum, or temple of the god Sarapis, and the catacombs. The polished granite shaft of Pompey's Pillar (actually dedicated to Diocletian soon after 297) stands on the highest spot in Alexandria and is about 88 feet high, including the pedestal. Close by are the famous Hadrianic catacombs of Kawm ash-Shuqāfah containing a very beautiful central tomb surrounded by many passages and containing small chambers and a large triclinium. The two ancient Egyptian obelisks called Cleopatra's Needles that formerly faced the great (eastern) harbour were removed to London and New York City in the 19th century. The site of the Pharos of Alexandria at the eastern tip of the island has long been under the sea, but the ruins of Qāʾit Bāy's fort and mosque remain.

Muḥammad ʿAlī Pasha is commemorated by an equestrian statue in Mīdān at-Taḥrīr and Saʿd Pasha Zaghlūl by a statue in the square named after him. South of al-Ḥurrīyah Avenue (formerly Rossette Street), one of the main east–west streets, is a mosque containing the tomb of Sāʿīd Pasha and those of other royal figures.

Economic life. The industrial region of al-Qabbārī contains asphalt works, oil, rice, and paper mills. On either side of the canal are the warehouses of wholesale dealers in cotton, wool, sugar, grain, onions, and other products. Close by is the Mīnāʾ al-Baṣal area, which contains the cotton exchange. Beyond al-Qabbārī to the southwest is al-Maks, with its salt and tanning industries and stone quarries.

Commerce and industry

The area south and southeast of Alexandria has been developed for a variety of industries, including cotton textile and dyeing works; factories for ginning and cotton-seed-oil pressing; paper, chocolate and confectionery works; factories for food processing and canning; leather, soap, and soda works; plastics, pottery, and car-assembly workshops. An oil refinery has been established, and electric power available to the city has been much increased.

In the Abīs district, four miles south-southeast of the city, an area of some 24,000 acres was drained and reclaimed from marsh, beginning in the mid-1950s. Model villages were built and landless peasants from overcrowded areas resettled and given a new start. Their crops are marketed in Alexandria. In the desert to the west extensive agricultural experiments were also undertaken; olives, vines, figs, and tomatoes have been successfully grown.

By far the largest and most valuable export from Egypt via Alexandria is raw cotton, followed by cereals, vegetables and fruit, wooden and plastic products, and textiles. Tea, coffee, timber, raw wool, and machinery are imported. The inhabitants of Alexandria are mostly occupied in trade and transport of merchandise in all its branches, in harbour work and industries, in shipping, and in fishing.

Government. The first fully constituted municipality in Egypt was established in Alexandria in 1895, built upon an earlier organization formed by the foreign merchants. Home rule there preceded that in Cairo, the capital, by more than half a century. Today Alexandria (city and agricultural hinterland) constitutes a *muḥāfaẓah* or governorate, and its government is responsible to the Ministry of Local Government.

Education. Since the 1952 revolution, Egypt's educational system has comprised four types of schools: primary, postprimary or preparatory, secondary, and technical. Attendance is compulsory for children between six and 12 years of age. East of Alexandria is Naṣr College (formerly Victoria College), a boys' boarding school run on English public school lines, although under Egyptian management since the Suez conflict of 1956–57. There is also a girls' school run on English lines. Also to the east of the city is the University of Alexandria (formerly Farouk University), founded in 1942, which has eight faculties and four affiliated institutes. Close by is the Mu'assah Hospital and College of Nursing, the hospital being among the best equipped in Europe or the Middle East.

Cultural life. Near Mīdān at-Taḥrīr is the Greco-Roman Museum, with its excellent collection of antiquities. The Municipal Library houses about 70,000 volumes in Arabic and European languages. Beside it is an art gallery. The University of Alexandria Library has about 1,200,000 volumes. There are two fine botanical gardens in the city—the Nuzhah Gardens and the Antoniadis Park.

Near the railway station is a sports stadium accommodating 25,000 spectators which serves as the site of the Arab Olympic Games. To the northeast, near the university, are two sports clubs, which include polo fields, golf links, tennis courts, a racecourse, and other facilities. On the coast, al-Jaysh Street ends at the former royal palace of Muntazah, which is now a casino where only foreigners are admitted; the former royal gardens and bathing beaches, however, are open to all on payment. In addition to the large Said Darwish theatre, there are seven cinemas.

BIBLIOGRAPHY. ANTHONY DE COSSON, *Mareotis* (1935), is an authoritative, well-written history and description of the Maryut district, which contains varied aspects of Alexandrian history and life from ancient to recent times, with sources given. PERCIVAL G. ELGOOD, *The Ptolemies of Egypt* (1938), is a scholarly but lively history of the Ptolemies of Alexandria from Ptolemy I Soter (died 283 BC) to Cleopatra (died 30 BC). E.M. FORSTER, *Alexandria: A History and a Guide*, 2nd ed. (1938), a short, thorough, and readable work, includes a detailed account of the Alexandria Museum and describes the town quarter by quarter as it was in the early 1920s. MARY ROWLATT, *A Family in Egypt* (1956), descriptive of Egypt generally, contains aspects of 19th- and 20th-century European life in Alexandria, based on family letters. The same author's *Founders of Modern Egypt* (1962) is a detailed history of events during the 1870s and 1880s in Egypt, the culmination of which took place in Alexandria. Also of interest is HASSAN EL SAATY and GORDON HIRABAYASHI, *Industrialization in Alexandria* (1959).

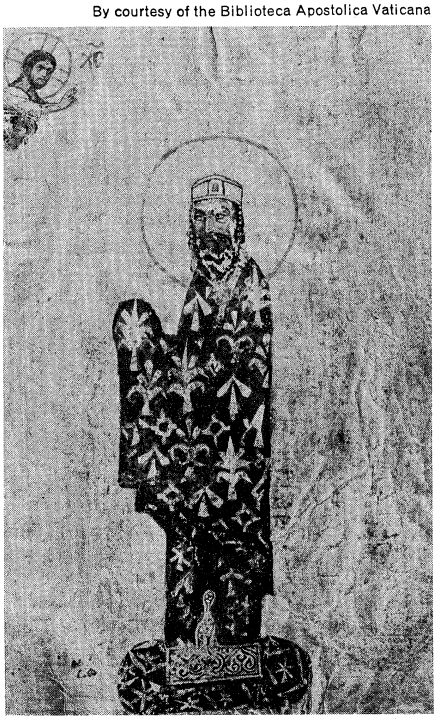
(Ma.R.)

Alexius I Comnenus

Alexius I Comnenus, Byzantine emperor at the time of the First Crusade, founded the Comnenian dynasty and

partially restored the strength of the empire after its defeats by the Normans and Turks in the 11th century.

By courtesy of the Biblioteca Apostolica Vaticana



Alexius I Comnenus, illumination from a Greek manuscript. In the Vatican Library (Cod. Vat. Gr. 666).

Alexius (1048–1118), third son of John Comnenus and nephew of Isaac I (emperor 1057–59), came of a distinguished Byzantine landed family and was one of the military magnates who had long urged more effective defense measures, particularly against the Turks' encroaching on Byzantine provinces in eastern and central Anatolia. From 1068 to 1081 he gave able military service during the short reigns of Romanus IV, Michael VII, and Nicephorus III. Then, with the support of his brother Isaac and his mother, the formidable Anna Dalassena, and with that of the powerful Ducas family, to which his wife Irene belonged, he seized the Byzantine throne from Nicephorus III.

Alexius was crowned on April 4, 1081. After more than 50 years of ineffective or short-lived rulers, Alexius, in the words of Anna Comnena, his daughter and biographer, found the empire "at its last gasp," but his military ability and diplomatic gifts enabled him to retrieve the situation. He drove back the south Italian Normans, headed by Robert Guiscard, who were invading western Greece (1081–82). This victory was achieved with Venetian naval help, bought at the cost of granting Venice extensive trading privileges in the Byzantine Empire. In 1091 he defeated the Pechenegs, Turkic nomads who had been continually surging over the Danube into the Balkans, hitherto defying all attempts at placation or subjugation. Alexius halted the further encroachment of the Seljuq Turks, who had already established the Sultanate of Rūm (or Konya) in central Anatolia. He made agreements with Sulaymān ibn Qutalmïsh of Konya (1081) and subsequently with his son Qïlïch Arslan (1093), as well as with other Muslim rulers on Byzantium's eastern border.

Military successes

At home, Alexius' administrative policy of strengthening the central authority and building up professional military and naval forces resulted in increased Byzantine strength in western and southern Anatolia, and in east Mediterranean waters. But he was unable or unwilling to make any effective attempt to limit the considerable powers of the landed magnates who had threatened unity of the empire in the past. Indeed, he strengthened their position by further concessions and he had to reward services, military and otherwise, by granting fiscal rights over specified areas. This method, which was to be increasingly employed by his successors, inevitably weakened central revenues and imperial authority. He repressed heresy and maintained the traditional imperial role of protecting the Eastern Orthodox Church, but he did not hesitate to seize ecclesiastical treasure when in financial need. He was subsequently called to account for this by the church.

To later generations Alexius appeared as the ruler who pulled the empire together at a crucial time, thus enabling it to survive until 1204, and in part until 1453, but modern scholars tend to regard him, together with his successors John II (reigned 1118–43) and Manuel I (reigned 1143–80), as effecting only stopgap measures. But judgments of Alexius must be tempered by allowing for the extent to which he was handicapped by the inherited internal weaknesses of the Byzantine state, and even more by the series of crises precipitated by the western European crusaders from 1097 onward (see CRUSADES). The crusading movement, motivated partly by a desire to recapture the holy city of Jerusalem, partly by the hope of acquiring new territory, increasingly encroached on Byzantine preserves and frustrated Alexius' foreign policy, which was primarily directed toward the re-establishment of imperial authority in Anatolia. Alexius had to see his relations with Muslim powers disrupted on occasion, and former valued Byzantine possessions, such as Antioch, pass into the hands of arrogant Western princelings, who even introduced Latin Christianity in place of Greek. Thus, it was during Alexius' reign that the last phase of the clash between the Latin West and the Greek East was inaugurated. He did regain some control over western Anatolia; he also advanced into the southeast Taurus region, securing much of the fertile coastal plain around Adana and Tarsus, as well as penetrating further south along the Syrian coast. But neither Alexius nor succeeding Comnenian emperors were able to establish permanent control over the Latin crusader principalities. Nor was the Byzantine Empire immune from further Norman attacks on its western islands and provinces—as in 1107–08, when Alexius successfully repulsed Bohemond I of Antioch's assault on Avlona in western Greece. Continual Latin, particularly Norman, attacks, constant thrusts from Muslim principalities, the rising power of Hungary and the Balkan principalities—all conspired to surround Byzantium with potentially hostile forces. Even Alexius' diplomacy, whatever its apparent success, could not avert the continual erosion that ultimately led to the Ottoman conquest.

Effects of the First Crusade

BIBLIOGRAPHY. F. CHALANDON, *Essai sur le règne d'Alexis Ier Comnène, 1081–1118* (1900), still the principal full-length study, now in need of some reconsideration; G. OSTROGORSKY, *History of the Byzantine State*, 2nd Eng. ed. (1968), a useful brief account with bibliography; J.M. HUSSEY (ed.), *Cambridge Medieval History*, new ed., vol. 4, pt. 1 (1966), with full bibliography, and *Church and Learning in the Byzantine Empire, 867–1185* (1937, reprinted 1963); H. AHRWEILER, *Byzance et la mer* (1966), for Alexius' administrative and naval policy; *History of the Crusades*, 2nd ed., vol. 1 ed. by M.W. BALDWIN, vol. 2 ed. by R.L. WOLFF and H.W. HAZARD (1969), with some bibliography; S. RUNCIMAN, *History of the Crusades*, vol. 1 (1951), a vivid presentation, particularly of the Byzantine point of view; ANNA COMNENA, *The Alexiad*, trans. by E.R.A. SEWTER (1969), a valuable and lively contemporary source by Alexius' daughter.

(J.M.H.)

Alfonso V the Magnanimous, of Aragon

The figure of Alfonso the Magnanimous, king of Aragon from 1416 to 1458, was praised, respected, and admired by the writers of his own time and also by those of the next generation. The latter were still close enough to him to draw upon a living tradition but were free of the desire to flatter that affected his contemporaries. Among Alfonso's apologists were the Italian Humanist scholars Antonio Beccadelli, Aeneas Sylvius Piccolomini (Pope Pius II), Vespasiano da Bisticci, and Giovanni Pontano. They praised Alfonso for his Humanist education and for his love of books and fine arts, for his delight in hunt-

ing, dancing, tournaments, and good clothes, and for his charity, clemency, and deep religious faith. He has been regarded by some scholars as a brilliant Renaissance prince and a great sovereign, but, in general, modern Spanish historians are less enthusiastic about Alfonso and blame him for occupying himself with amorous adventures in Naples while neglecting his duties to his peninsular territories.



Alfonso V, bronze medal by Pisanello, 1448–49. In the Bargello, Florence.
Alinari—Giraudon

Alfonso was one of the seven princes of Aragon about whom the poet Jorge Manrique wrote; they were the children of Fernando de Antequera (second son of King John I of Castile and afterward Ferdinand I of Aragon) and his aunt Leonor de Alburquerque. Leonor was six years older than her husband and nephew, but the masculine descendants of this unequal match possessed infinite vitality and made their mark on the history of Aragon.

Alfonso was born in 1396 and brought up in the brilliant Castilian court at Medina del Campo; but when he was 16, his father became king of Aragon, and he himself went to live there. Three years later (1415) he married his cousin María, the daughter of Henry III of Castile, but she produced no children, and they were separated for many years. The marriage was a failure and perhaps helps to explain Alfonso's reluctance to return to his peninsular kingdoms after he had conquered Naples, where he was encouraged to remain by his mistress, Lucrezia de Alagno.

He succeeded his father as king of Aragon in 1416 and, at the beginning of his reign, had political difficulties with both Catalans and Aragonese, because he retained some Castilian counsellors and deprived the *justicia*, the supreme law officer of Aragon, of his position.

Mediterranean expansion

From the moment of his accession, Alfonso continued the traditional Aragonese policy of Mediterranean expansion. Thus, in 1420 he set out with a fleet to pacify Sardinia and Sicily and to attack the Genoese possession of Corsica. The queen of Naples, Joan II, then sought his help against Louis III of Anjou and adopted him as her son and heir. Alfonso was received as a liberator in Naples on July 5, 1421; but the volatile character of the Queen, who soon afterward began to make overtures to Louis of Anjou, obliged Alfonso in 1423 to return to Catalonia to seek reinforcements.

After intervening in the internal politics of Castile to defend the interests of his brothers Henry and John in the near civil war that existed during the weak rule of John II, Alfonso set out again for Italy, from where, as it turned out, he was never to return. He was receiving tempting offers (1432) to intervene again in Naples and spent two years in Sicily preparing his fleet and army. His opportunity seemed to come in 1435, after the deaths of Louis III of Anjou and Queen Joan II, but while blockading the port of Gaeta, a key citadel from which to launch an attack on Naples, he was defeated off the island of Ponza by a Genoese squadron. Alfonso was captured, with many others, and sent as a prisoner to Genoa and then to Milan, whose duke, Filippo Maria Visconti, ruled both cities. Alfonso, however, charmed his captor into an alliance and then continued his fight to gain possession of Naples against the opposition of Venice, Florence, and the Pope. He took Naples on July 2, 1442, and transferred his court there permanently in 1443. It became a brilliant centre of art and culture, fed by the fertile interaction of Italian Renaissance and Spanish Gothic influences and forming a cultural bridge between the two peninsulas of the western Mediterranean.

Alfonso engaged in much diplomatic and military activity in Africa, the Balkans, and the Eastern Mediterranean in order to protect his commerce with the East and to share in the defense of Christendom against the Turks. He helped the Knights of St. John defend Rhodes; allied himself with Hungary (1444), with Serbia (1447), and with Abyssinia (1450); and he fought against Egypt (1453–54). But he was not strong enough to do anything to prevent the fall of Constantinople to the Turks in 1453.

Unrest in Spain

Meanwhile, his Spanish dominions were suffering from serious unrest, the result of social and economic tensions to which no solution could be found by his viceroys, his queen, Maria, and his brother John of Navarre. In Catalonia the *remensa*, the peasantry, were vigorously seeking to be freed from feudal dues and received some support from the crown. In Majorca a popular rising, which led to fighting between the capital of the island and the rural population, had to be crushed by troops that Alfonso sent from Naples. And in Barcelona a serious class struggle caused so many disturbances in the city that Alfonso reformed the city government, allowing public offices to be distributed by lot. Meanwhile, the sporadic war with Castile both impoverished the Kingdom of Aragon and deprived Alfonso and his family of their ancestral estates in Castile. Only Valencia, with its flourishing economy, remained unharmed by the general crisis. A restless, energetic ruler to the last, Alfonso was engaged in an assault upon Genoa, which had recently surrendered to the French, when death surprised him in Ovo Castle at Naples, on June 27, 1458. In the Kingdom of Naples he was succeeded by his illegitimate son, Ferrante, and in his other states by his brother John (King John II of Aragon), who had been king of Navarre since 1425.

BIBLIOGRAPHY. The fundamental work on Alfonso V the Magnanimous, although now somewhat outdated, is J. AMETLLER Y VINYAS, *Alfonso V de Aragón en Italia y la crisis religiosa del siglo XV*, rev. ed., 3 vol. (1903–28). Also of value are C. BATLLE, *Estudios sobre Alfonso el Magnánimo con motivo del quinto centenario de su muerte* (1960); A. GIMENEZ SOLER, *Itinerario del rey don Alonso de Aragón el que ganó Nápoles* (1909); J.M. MADURELL MARIMON (ed.), *Mensajeros barceloneses en la corte de Nápoles de Alfonso V de Aragón* (1963); and A. SORIA ORTEGA, *Los humanistas de la corte de Alfonso el Magnánimo* (1956).

(E.Sa.)

Alfonso VI of Castile and Leon

The Spanish Christian king Alfonso VI ruled over Leon from 1065 until his death in 1109 and over Castile from 1072. He is remembered for his relationship with Spain's national hero, Rodrigo Díaz de Vivar (el Cid), and because his oppressive policies toward his Muslim vassals led to the invasion of Spain by the Almoravids in 1086. But if his reign was politically unsuccessful, Alfonso VI carried out an important cultural task by Europeanizing his dominions. Alfonso married Constance of Burgundy, and influences from across the Pyrenees showed themselves in the introduction of the Romanesque style in art, the adoption of the Roman instead of the Mozarabic liturgy, the replacement of Visigothic by Carolingian script, and the energetic support that Alfonso gave to Cluniac monasticism, as well as in his reconstruction and safeguarding of the pilgrim road to Santiago.

Alfonso was born sometime before June 1040, the second son of King Ferdinand I and his wife Sancha; he was educated by Raimundo, later the bishop of Palencia, and by Pedro Ansúrez, the count of Carrión. On his death in 1065, Ferdinand left to Alfonso the kingdom of Leon together with tribute paid by the Muslim king-



Alfonso VI, portrait miniature from a manuscript, 12th century. In Santiago Cathedral, Spain.
Archivo Mas, Barcelona

dom of Toledo. These possessions aroused the envious hostility of Alfonso's elder brother, Sancho II, who had inherited the kingdom of Castile and the tribute of Saragossa. Alfonso was defeated by his brother in two battles; after the defeat at Llantada (1068) he managed to retain his kingdom, but after that at Golpejera (1072) he was captured and exiled, living for a short while at the court of his vassal Ma'mūn, the Muslim king of Toledo. Soon Alfonso's sister Urraca stirred up a rebellion in Leon, and Sancho besieged her in the walled city of Zamora. During the siege, he was killed, perhaps at Urraca's instigation. She was clearly on Alfonso's side, and some modern historians have even suggested that they had an incestuous relationship.

Return from exile

With Sancho's death, Alfonso recovered his own kingdom of Leon and inherited (1072) that of Castile. He also occupied Galicia, which Sancho had kept from their younger brother García; Alfonso kept García in prison until his death. A late story, in which it is alleged that Alfonso took an oath in St. Gadea's Church, Burgos, that he had had no share in Sancho's murder, probably reflects Castilian reluctance to accept him as king.

Alfonso's reign now entered on a period of success. He seized the Rioja and the Basque provinces and received the feudal homage of Sancho Ramírez for the region of Navarre to the north of the Ebro River. By 1077 he had assumed the title *imperator totius Hispaniae* ("emperor of all Spain"), in which role other Christian kings accepted him. He then began the conquest of Toledo and, after a long siege, occupied it in May 1085. This was a vital conquest, which recovered for Christian Spain one of the most important historical, strategic, and cultural centres of the peninsula, one that had been in the possession of the Muslims since the early 8th century.

Enforced tribute of the Muslims

During this period Alfonso regularly exacted *parias*, heavy financial tributes, from the Muslim *ṭā'ifah* kingdoms in return for protection against their other enemies. By thus depriving them of their wealth, he hoped to weaken them so that they would eventually cede their independence without fighting. As a result of his exactions, Christian Spain was flooded with Muslim gold, which was spent on warfare and donations to shrines, churches, and monasteries. The demand for tribute caused the *ṭā'ifah* kings to tax their subjects heavily, producing popular discontent and disturbances and contributing to the weakness that caused the surrender of Toledo. The Muslim ruler al-Mu'tamid of Seville took a desperate decision and called for the help of Yūsuf ibn Tāshufīn, the Almoravid (Berber) emir of North Africa, and his Saharan tribes. The emir disembarked in Algeciras at the end of July 1086 and a few months later, on October 23 at Zallāqah, near Badajoz, inflicted a terrible defeat on Alfonso VI. Alfonso appealed for help to the rest of Christendom, and a small crusade was organized as a result; the crusaders did not reach Alfonso's lands but wasted their energies and resources in an unsuccessful siege of the Muslim outpost of Tudela.

The defeat at Zallāqah seriously lessened Alfonso's influence over the *ṭā'ifah* kingdoms. It also led to a reconciliation with Rodrigo Díaz, the Cid. Their relationship had been difficult from the start, because the Cid had taken a prominent part in the campaigns of Sancho against Alfonso; although the Cid at first attended Alfonso's court, suspicions deepened, and since 1081 he had been in the service of the Muslim king of Saragossa. Alfonso entrusted the Cid with the occupation and defense of eastern Spain, and Rodrigo carried out this task with great success. Between 1086 and 1109 Alfonso suffered constant defeat by the Almoravids; in the last battle, that of Uclés in 1108, he lost his only son, Sancho. It was characteristic of his indomitable spirit that he at once arranged for his daughter, Urraca, to marry Alfonso I of Aragon so that the war against the Almoravids should be continued after his death, even though it meant that Leon and Castile would be ruled by an Aragonese prince. He died in Toledo in 1109.

BIBLIOGRAPHY. The most important source for Alfonso VI and his age is RAMON MENENDEZ PIDAL, *La España del Cid*, 7th ed., 2 vol. (1969; Eng. trans. of 1st ed., *The Cid and His Spain*, 1934).

(E.Sa.)

Alfred the Great, of Wessex

Alfred the Great, king of Wessex, a Saxon kingdom in southwestern England, from 871 to 899, was the man whose courage, tenacity, and skill prevented England from falling to the Danes in the 9th century. By saving his own kingdom, he made it possible for his successor to reconquer the areas settled by the Danes. He was a remarkably enlightened ruler and was the only English king to bring about a great revival of learning not only by his encouragement but also by his own writings.



Alfred the Great, Anglo-Saxon coin, 9th century. In the British Museum.
By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co., Ltd.

Early scholarly interests

When he was born, in 849, it must have seemed unlikely that Alfred would become king, since he had four older brothers; he said that he never desired royal power. Perhaps a scholar's life would have contented him. His mother early aroused his interest in English poetry, and from his boyhood he also hankered after Latin learning, possibly stimulated by visits to Rome in 853 and 855. It is possible also that he was aware of and admired the great Frankish king Charlemagne, who had at the beginning of the century revived learning in his realm. Alfred had no opportunity to acquire the education he sought, however, until much later in life. From his youth

through most of his reign, he was occupied with the defense and government of his kingdom.

Alfred probably received the education in military arts normal for a young man of rank. He first appeared on active service in 868, when he and his brother, King Aethelred (Ethelred) I, went to help Burgred of Mercia (the kingdom between the Thames and the Humber) against a great Danish army that had landed in East Anglia in 865 and taken possession of Northumbria in 867. The Danes refused to give battle, and peace was made. In this year Alfred married Ealhswith, descended through her mother from Mercian kings. Late in 871, the Danes invaded Wessex, and Aethelred and Alfred fought several battles with them. Aethelred died in 871 and Alfred succeeded him. After an unsuccessful battle at Wilton he made peace. It was probably the quality of the West Saxon resistance that discouraged Danish attacks for five years.

In 876 the Danes again advanced on Wessex: they retired in 877 having accomplished little, but a surprise attack in January 878 came near to success. The Danes established themselves at Chippenham and the West Saxons submitted "except King Alfred." He harassed the Danes from a fort in the Somerset marshes, until seven weeks after Easter he secretly assembled an army, which defeated them at the Battle of Edington. They surrendered and their king, Guthrum, was baptized, Alfred standing as sponsor; the following year they settled in East Anglia.

Wessex secured

Wessex was never again in such danger. Alfred had a respite from fighting until 885, when he repelled an invasion of Kent by a Danish army, supported by the East Anglian Danes. In 886 he took the offensive and captured London, a success that brought all the English not under Danish rule to accept him as king. The possession of London also made possible the reconquest of the Danish territories in his son's reign, and Alfred may have been preparing for this, though he could make no further advance himself. He had to meet a serious attack by a large Danish force from the Continent in 892, and it was not until 896 that it gave up the struggle and retired to France.

The failure of the Danes to make any more advances against Alfred was largely a result of the defensive measures he undertook during the war. Old forts were strengthened and new ones built at strategic sites, and permanent arrangements were made for their manning. Alfred reorganized his army and used ships against the invaders as early as 875. Later he had larger ships built to his own design for use against the coastal raids that continued even after 896. Wise diplomacy also helped Alfred's defense. He maintained friendly relations with Mercia and Wales; Welsh rulers sought his support and supplied some troops for his army in 893.

Alfred succeeded in government as well as at war. He was a wise administrator, organizing his finances and the service due from his *thegns* (noble followers). He scrutinized the administration of justice and took steps to insure the protection of the weak from oppression by ignorant or corrupt judges. He promulgated an important code of laws, after studying the principles of lawgiving in the book of Exodus and the codes of Aethelbert of Kent, Ine of Wessex (688–94), and Offa of Mercia (757–96), again with special attention to the protection of the weak and dependent. While avoiding unnecessary changes in custom, he limited the practice of the blood feud and imposed heavy penalties for breach of oath or pledge.

Alfred's encouragement of scholarship

Alfred is most exceptional, however, not for his generalship or his administration but for his attitude to learning. He shared the contemporary view that Viking raids were a divine punishment for the people's sins, and he attributed these to the decline of learning, for only through learning could men acquire wisdom and live in accordance with God's will. Hence, in the lull from attack between 878 and 885, he invited scholars to his court from Mercia, Wales, and the Continent. He learned Latin himself and began to translate Latin books into English in 887. He directed that all young freemen of adequate means must learn to read English, and by his own translations and those of his helpers he made available English versions of "those books most necessary for all men to know," books that would lead them to wisdom and virtue. The *Ecclesiastical History of the English People*, by the English historian Bede, and the *7 Books of Histories Against the Pagans*, by Orosius, a 5th-century theologian—neither of which was translated by Alfred himself, though they have been credited to him—revealed the divine purpose in history. Alfred's translation of the *Pastoral Care* of St. Gregory I, the great 6th-century Pope, provided a manual for priests in the instruction of their flocks, and a translation by Bishop Werferth of Gregory's *Dialogues* supplied edifying reading on holy men. Alfred's rendering of the *Soliloquies* of the 5th-century theologian St. Augustine of Hippo, to which he added material from other works of the Fathers of the Church, discussed problems concerning faith and reason and the nature of eternal life. This translation deserves to be studied in its own right, as does his rendering of Boethius' *Consolation of Philosophy*. In considering what is true happiness and the relation of providence to faith and of predestination to free will, Alfred does not fully accept Boethius' position but depends more on the early Fathers. In both works, additions include parallels from contemporary conditions, sometimes revealing his views on the social order and the duties of kingship. Alfred wrote for the benefit of his people, but he was also deeply interested in theological problems for their own sake and commissioned the first of the translations, Gregory's *Dialogues*, "that in the midst of earthly troubles he might sometimes think of heavenly things." He may also have done a translation of the first fifty psalms. Though not Alfred's work, the Anglo-Saxon Chronicle, one of the greatest sources of information about Saxon England, which began to be circulated about 890, may have its origin in the intellectual interests awakened by the revival of learning under him. His reign also saw activity in building and in art, and foreign craftsmen were attracted to his court.

In one of his endeavours, however, Alfred had little success; he tried to revive monasticism, founding a monastery and a nunnery, but there was little enthusiasm in England for the monastic life until after the revivals on the Continent in the next century.

Asser's biography

Alfred died in 899. Alone of Anglo-Saxon kings, he inspired a full-length biography, written in 893, by the Welsh scholar Asser. This work contains much valuable information and it reveals that Alfred laboured throughout under the burden of recurrent, painful illness; and beneath Asser's rhetoric can be seen a man of attractive character, full of compassion, able to inspire affection, and intensely conscious of the responsibilities of kingly office. This picture is confirmed by Alfred's laws and writings.

Alfred was never forgotten: his memory lived on through the Middle Ages and in legend as that of a king who won victory in apparently hopeless circumstances and as a wise lawgiver. Some of his works were copied as late as the 12th century. Modern studies have increased knowledge of him but have not altered in its essentials the medieval conception of a great king.

BIBLIOGRAPHY. On ASSER'S *Life of King Alfred*, ed. by W.H. STEVENSON (1959 reprint with chapter on recent work by D. Whitelock), see D. WHITELOCK, *The Genuine Asser* (1968). The best modern account is F.M. STENTON, *Anglo-Saxon England*, 3rd ed. (1971); a popular account is E.S. DUCKETT, *Alfred the Great and His England* (1957). For other sources on the reign, see G.N. GARMONSWAY, *The Anglo-Saxon Chronicle* (1953); *English Historical Documents*, vol. 1, ed. by D. WHITELOCK (1955); and F.L. ATTENBOROUGH (ed.), *The Laws of the Earliest English Kings* (1922).

Editions of the Alfredian translations: King Alfred's West-Saxon Version of Gregory's Pastoral Care (1871) and *King Alfred's Orosius* (1883), both ed. by H. SWEET; *King Alfred's Old English Version of Boethius*, ed. by W.J. SEDGEFIELD (text 1899; modern English, 1900); *King Alfred's Old English Version of St. Augustine's Soliloquies*, ed. by H.L. HARGROVE (text 1902; modern English, 1904); *King Alfred's Version of St. Augustine's Soliloquies*, ed. by T.A. CARNICELLI (1969); *The Old English Version of Bede's Ecclesiastical History*, ed. by

T. MILLER (1890–98); Werferth's translation of Gregory's *Dialogues*, ed. by H. HECHT (1900). See also D. WHITELOCK, "The Prose of Alfred's Reign," in E.G. STANLEY (ed.), *Continuations and Beginnings* (1966), with bibliography; J.M. BATELY, "King Alfred and the Old English Translation of Orosius," *Anglia*, 88:433–460 (1970); K. OTTEN, *König Alfreds Boethius* (1964); and F. ANNE PAYNE, *King Alfred and Boethius* (1968).

(D.W.)

Algae

Algae include the pond scums of freshwater, the greenish stains on damp rocks and tree trunks, and the massive seaweeds of the oceans. These primitive plantlike organisms (some biologists call them plants, others group them with the protists) are capable of existing in places where few other living things can survive: in hot springs or in Arctic snow, in freshwater or marine environments. Most of them are free-living, only a few having intimate associations with other organisms, as symbionts (in lichens) or as parasites. The algae are a collection of phyla, or divisions, comparable to animal phyla. They exhibit a diversity of form ranging from one-celled organisms as small as the largest bacteria to elaborate giants, such as the kelps, comparable in size with flowering plants. They are invaluable in the economy of nature as primary producers of the food that ultimately sustains all aquatic life and as providers of the earth's store of oxygen. The marine planktonic algae, called the "grass of the sea," have been estimated to carry on about 90 percent of all photosynthetic activity on earth. Certain larger marine algae are a source of human food, food additives, and other products of industrial use. Since algae are common inhabitants of surface water, they often are troublesome in water supplies, producing disagreeable odours and "fishy" tastes and clogging filtration machinery. Because of the increase of mineral wastes that can be used as food by algae, many inland ponds and lakes are being smothered as a result of algal population explosions.

This article is divided into the following major sections:

I. General features
II. Economic importance
III. Natural history
IV. Form and function
V. Evolution and phylogeny
VI. Classification

I. General features

Although there are a number of colourless morphological derivatives among the simpler forms, algae normally contain the green pigment chlorophyll a, together with other pigments. They are not organized into true roots, stems, and leaves as are higher plants, but some brown and red algae simulate roots—*e.g.*, haptera in *Laminaria*, stems in *Sargassum*, leaves in *Levillea*. None has true tissues, particularly vascularized ones, though some thalli, especially in the brown algae, are complex.

DIVERSITY

Algae range in size from 3 microns to 62 metres (203 feet), and there are approximately 25,000 species. It is difficult to provide a precise definition of algae since they are a diverse group. The Cyanophyta, or blue-green algae, usually are grouped with the bacteria because they lack the nuclear, mitochondrial, and chloroplast membranes possessed by all other algae. Many of the unicellular flagellate forms also are classed as Protozoa (*q.v.*), but there may be one or more representatives that indicate a true algal ancestry.

Major groups

Algae range from single-celled (*e.g.*, *Chlamydomonas*) to colonial forms (*e.g.*, *Volvox*), branched filaments (*e.g.*, *Cladophora*, *Ectocarpus*), leafy forms (*e.g.*, *Ulva*, *Porphyra*), crusts such as *Ralfsia*, and large multicellular types (*Dictyota* and *Fucus*). In the three major algal groups, Chlorophyta (green algae), Phaeophyta (brown algae), and Rhodophyta (red algae), there are lime-encrusted genera and species that, in past eras, contributed to geological strata that subsequently emerged from the water. Other members are characterized by deposition of silica (*e.g.*, diatoms, stoneworts). Generally, the structure of the most complex members, Fucales and Laminariales, includes a central medulla, surrounded by a cortical region, and an outermost epidermal layer. In many Phaeophyta and all Rhodophyta the structure is basically filamentous. Growth occurs in a variety of ways, and reproductive bodies can be produced in unmodified vegetative cells, in special cells, or in thallus organs or outgrowths.

DISTRIBUTION

Marine algae, such as the common seaweeds, are most noticeable on rocky coastlines, where they form an almost continuous rock cover in the northern temperate zone. In the tropics, marine algae are not evident in the intertidal zone, where exposure is severe, but are abundant in the sublittoral zone. Here they are found on the floors of lagoons in association with coral reefs and island atolls, and they contribute to reef formation. In the Southern Hemisphere, algae are most conspicuous near the low-water mark and in the sublittoral zone. Algae extend into Arctic and Antarctic waters to considerable depths. Marine water and freshwater contain myriad minute, floating, motile representatives, the phytoplankton (see PLANKTON), and the great floating masses of the Sargasso weed (*Sargassum natans*) have been known for centuries. Algae are not conspicuous in freshwater unless it is polluted; *e.g.*, masses of green algae (*Cladophora*) have become a serious problem in Lake Erie (U.S.) each summer. Algae are found on leaves, especially in the tropics and subtropics, and on wood and stones in all parts of the world; algae also inhabit the soil, where they are capable of surviving dry conditions for long periods. One or two species capable of tolerating temperatures of 80° C live in and around hot springs, and a small number live in snow.

II. Economic importance

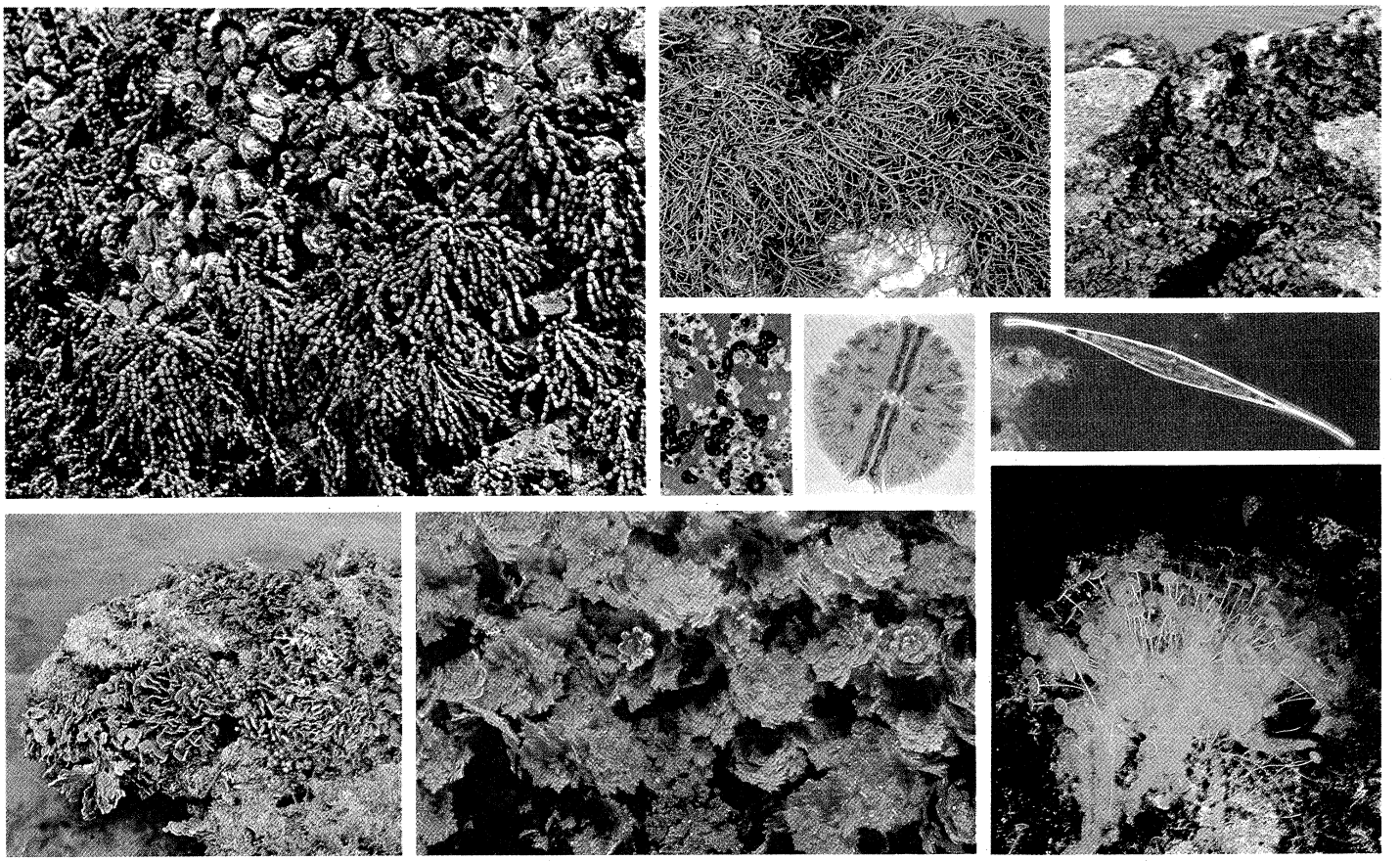
Algae have been important economically since the beginning of recorded history. The Chinese used them as food, and it is likely that the Japanese, Hawaiians, and inhabitants of some western European countries did also. Algae still are used as stock feed and as fertilizer. At one time in Great Britain substantial industries existed for the manufacture of soda and of iodine. The manufacture of agar, alginates, and carrageenan has replaced these earlier industries.

Some members of the Cyanophyta have important roles in the biosphere because they fix nitrogen directly from the air. *Chlorella*, *Scenedesmus*, *Chlamydomonas*, and other flagellates are important for final sewage purification in sewage oxidation ponds because they provide the oxygen that enables aerobic bacteria to complete the purification process. These flagellates are used also for basic research in metabolic mechanisms and for biological assays. The planktonic algae of the sea and freshwater are primary producers; without them no higher life forms could exist in these waters. Fossil algae are associated with the oil, dynamite, and lime industries. The capacity of certain algae to form close physiological associations with other plants and animals (symbiosis) is of great biological importance. Lichens are symbiotic associations between certain algae and fungi (see LICHEN). The dinoflagellates associated with coral polyps are essential to the formation of coral reefs (see CORAL ISLANDS, CORAL REEFS, AND ATOLLS).

Algae also have certain deleterious effects. Excessive algae in water can render it unpalatable, even unusable, for human consumption. Certain dinoflagellates and Cyanophyta produce algal blooms and liberate toxins that kill animals. There are parasitic algae, and one species, *Cephaleuros virescens*, is responsible for the disease known as tea rust that can cause severe damage on tea plantations.

COMMERCIAL PRODUCTS

Geloses. One important algal industry is concerned with geloses that can be extracted from some red sea-



Colour diversity in algae.
(Top left) *Hormosira* and mussels. (Top centre) *Gigartina*. (Top right) mostly *Colpomenia* on rock. (Centre left) *Rivularia* and barnacles *Chthamalus stellatus*. (Centre right) Desmids (magnification 85 x). (Centre far right) Desmid *Closterium* (magnified 250 x). (Bottom left) *Halimeda discoidea*. (Bottom centre) *Caulerpa*. (Bottom right) *Acetabularia mediterranea*.
(Top left) V.J. Chapman, (top centre) Alison Wilson, (top right) Robert W. Hoshaw—EB Inc., (centre left and bottom left) Douglas P. Wilson, (centre right and centre far right) Photo Researchers, (centre right) Winton Patnode, (centre far right) Eric Grave, (bottom centre) Richard H. Chesher, (bottom right) Raniero Maltini and Piero Solaini—SCALA

weeds. The gelose-producing algae, 78 species distributed among 30 genera, can be separated by the gelling characteristics of their extracts. The proportion of two polysaccharides (*i.e.*, complex carbohydrates), agarose and agaropectin, is responsible for the properties of agar extracts.

Uses of geloses

Agar. Agar manufacture originated in China, then spread to Japan; and, until World War II, Japan was virtually the sole manufacturer. Now the U.S.S.R., the U.S., Australia, and New Zealand also manufacture agar. The principal raw sources for agar production are species of *Gelidium*, *Gracilaria*, *Pterocladia*, *Acanthopeltis*, and *Ahnfeltia*. One important use of agar is in bacteriological work, in which it functions as a solidifying factor in growth media used for culturing micro-organisms. Agar is used also in canning fish, transporting cooked fish, sizing fabrics, making films, finishing leather, producing high-grade adhesives, cooking (soups, sauces, jellies), manufacturing ice creams and icings, clarifying beers and wines, manufacturing shoe stains and cosmetics (shaving soaps and creams), and in other minor industries.

Carrageen. Other edible algae on both sides of the North Atlantic are called Irish Moss or carrageen. The name applies principally to *Chondrus crispus*, but *Gigartina stellata* is included also. The main sources of carrageen are Ireland, France, New England, and Canada. In Canada the *Chondrus* harvest amounted to 5,500,000 pounds (2,750 tons) of dried seaweed in 1960. This seaweed, collected from boats by hand or with special rakes, can be used in the dried condition; generally, however, a gelatinous substance, known as carrageenan, is extracted from the dried preparation. Usually carrageenan is used commercially as the sodium, potassium, or calcium salt. Carrageenan extracts are important in the stabilization of emulsions, suspensions, and foams, and their uses parallel those of agar. The dried plants are used for making blancmanges and jellies and as clarifying agents for beers, coffee, honey, and wine. The extract of the red alga *Furcellaria fastigiata*, known as furcellaran, has the same uses as carrageenan but is especially valuable in jam making. Funoran from *Gloiopeltis furcata* (called "funori" in Japan) is a valuable gelan used as a glue.

Algins and alginates. Alginic acid is a major algal colloid. It occurs in commercially obtainable quantities in large brown rockweeds (*e.g.*, *Fucus*, *Ascophyllum nodosum*, *Laminaria*, *Macrocystis*), in which it is found in amounts ranging from 12–30 percent dry weight.

Sites for collecting and processing seaweed for alginates include the Pacific coast of the U.S. (*Macrocystis*); the Atlantic coast of the U.S. and Canada (*Fucus* and *Ascophyllum*); France, Great Britain, and Norway (*Laminaria* species, *Fucus*, *Ascophyllum*); Japan (*Laminaria*, *Eisenia*, *Ecklonia*); and Tasmania (*Macrocystis*). Other than those in Norway, most economically usable supplies in Europe have been exploited fully. The salts of alginic acid are of commercial importance; alkali salts can be spun like silk to produce artificial threads, and heavy metal salts, insoluble in water, form a plastic material. In the textile industry algin is associated with the printing of fabrics. Colloidal properties of alginates are utilized in industries concerned with cosmetics, car polishes, paints, and pharmaceuticals. Alginates are used also in the clarification of sugar solutions and mineral waters. The insoluble alginates are used in the preparation of films, gels, plastics, vulcanites, linoleum, and imitation leathers. The recent production of organic derivatives (propylene

Uses of alginates

glycol esters) from alginates has extended their usefulness, especially as natural gums.

Algae as food. *Animal consumption.* Although fresh seaweeds were eaten by horses, cattle, sheep, and pigs in the past, modern practice involves drying the seaweed and grinding it into a meal. The principal source is rockweed (*Fucus, Ascophyllum*), though *Laminaria* species are used also. Since seaweed does not provide a balanced diet it usually comprises only 10–15 percent of the total animal diet. The digestibility of the meal depends not only on the seaweed source but also on the condition of the animal to which it is given. Animals in poor condition seem able to utilize it better than healthy animals.

Laver, dulse, kombu

Human consumption. A principal natural algal food product is laver, or nori as it is termed in Japan, where its preparation is a major industry. Laver also is eaten in parts of Britain and Ireland, in New Zealand, and in Korea. The principal Japanese species is *Porphyra tenera* (a rhodophyte). The seaweed usually is obtained from cultivated areas. In the past it was grown on twigs planted in specific areas, but coconut-fibre nets or bamboo blinds now are used. The crop is increased by seeding the water with spores from a dwarf "conchocelis" phase that has been grown in culture tanks. Nori is required in such large quantities (about 130,000 tons annually) that the alga now is grown commercially in factories that use artificial seawater. Laver is useful in a variety of foods, including sandwiches, soups, and biscuits. Its limited nutritional value lies in its protein, vitamin, and mineral salt content; its widespread use in Japan is said to keep the incidence of goitre low.

Other than laver and the food products that utilize Irish Moss, agarophytes, and alginates, seaweed is seldom eaten by Western man. It is a staple food in Hawaii, where the various edible seaweeds (about 40) are known as different kinds of limu. Dulse (*Rhodymenia palmata*) is collected and packaged in Nova Scotia. In the Chad Republic of Central Africa some tribes collect and eat the dried mats of the blue-green alga *Spirulina platensis*. In Japan, major food products obtained from members of the Laminariales are known as kombu or kobu. About 15 different species are used, and, after drying, various treatments provide different kinds of kombu. One estimate (1955–60) for the average annual harvest indicated that 140,000 tons (280,000,000 pounds) of algae were collected for consumption as kombu. Wakame is made from the oarweed *Undaria pinnatifida*, and arame from *Eisenia bicyclis*. *Ulva*, *Enteromorpha*, and *Monostroma* also are used as food; *Monostroma* species are grown on nets in the same manner as *Porphyra*. Many other smaller brown and red seaweeds are collected, dried, and eaten in various ways.

There is little food value in the complex nitrogenous and carbohydrate compounds found in algae because humans cannot break down these complex structures into useful metabolites during digestion. Algae do have value as roughage, however; and their mineral, vitamin, and iodine content provide valuable nutrients.

Other products. Fossil diatom deposits (Kieselguhr) produce a hard powder used in filtration processes and in the cleaning and polishing of metals. One important use of diatomaceous earth is in dynamite manufacture; nitroglycerin absorbed by diatomaceous earth is safe to handle. The striations on living diatoms have been used for many years by lens manufacturers for testing lenses.

At one time seaweeds, especially the large brown ones, were collected, dried, and burned to produce an ash called kelp, from which soda and then iodine were extracted; small amounts of iodine still are produced in this way in Japan and possibly in the U.S.S.R. Algae are valuable in other ways to the drug industry. In addition to algal products used in drug preparation, short pieces of *Laminaria* are used in surgery for widening fistulas and wound entrances. An extract of the red seaweed, *Digenea simplex*, is marketed as a drug for the treatment of worms. Wherever seaweeds occur in great quantity they are used as fertilizer and usually are applied fresh, though sometimes a compost is used. The rich agricultural crops typical of northwestern France are attributed directly to regular use of seaweed fertilizer; coralline red algae (*Phymatolithon, Lithothamnion*) known as maerl are used. In recent years liquid extracts of brown rockweeds and oarweeds have been valuable in growing vegetables and glasshouse crops. These extracts are valuable for their nutrients, trace elements, and other growth-promoting substances.

Seaweeds used as fertilizer

A number of compounds with potential commercial value have been extracted from brown algae. Laminarin, the main food reserve of the Laminariales, may someday be useful as an anticoagulant. The sugar alcohol, mannitol, is another algal food reserve that may find use in the pharmaceutical industry and in the preparation of lacquers and plastics. Algae may come to play some role in the development of long-term life-support systems that will be necessary for space exploration. The amount of algae needed to produce a daily quota of oxygen for a man also absorbs his daily production of waste carbon dioxide. During this process, algae are produced, which possibly could be rendered palatable and digestible as food.

III. Natural history

REPRODUCTION

Reproduction in algae is vegetative, asexual, or sexual.

Vegetative reproduction. Vegetative reproduction includes multiplication by cell division or fragmentation. Vegetative propagules are produced regularly in certain algae (*e.g., Codium, Sphacelaria*). *Halimeda*, a tropical member in the Siphonales, possesses underground rhizoidal (rootlike) filaments for vegetative propagation. Regeneration, considered a means of vegetative reproduction, usually involves a basal rhizoidal or pad portion that survives an unfavourable season and gives rise to new shoots. Such organs are known in *Bryopsis* and in members of the Dasycladales and Phaeophyta; in the Rhodophyta the growth of a new adult *Bonnemaisonia aspara-*

Osamu Umebayashi



Algae farming in Japan.
(Left) Full-grown *Porphyra* leaves on culture net. (Right) Harvested *Gelidium* is dried at the beach.

goides from a dwarf prostrate "hymenoclonium" phase is also a form of vegetative reproduction.

Asexual reproduction. Asexual reproduction involves differentiated or specialized individual cells or a group of cells (*e.g.*, a short section of filament in some Cyanophyta can become covered by a thickened membrane and serve as a multicellular hormospore). Although many different kinds of spores are involved in asexual reproduction, motile zoospores are most common in algae. Zoospores, which have from one to four flagella and frequently have chloroplasts and a pigment spot, number from one to 64 or more per cell. In most cases they are haploid (n; see GENE), generally giving rise to a sexual haploid generation; in some cases, however, they serve as an accessory means for reproducing either the haploid or the diploid (2n) generation (*e.g.*, *Ectocarpus*, *Asperococcus*). Zoospores can be produced in specialized organs (*e.g.*, unilocular sporangia of Phaeophyta) or in cells similar to other vegetative cells (*e.g.*, *Ulothrix*). Meioswarmers found in some members of the Chaetophorales (*e.g.*, *Stigeoclonium*) and Ectocarpales (*e.g.*, *Ectocarpus padinae*) represent modified zoospores. Androspores that germinate into dwarf male plants of the Oedogoniales also are modified zoospores.

Zoospores

Typically zoospores are enclosed only in a cytoplasmic membrane, but some algae form nonmotile zoospores that are enclosed within a wall (aplanospores); others form spores with thick walls (hypnospores) that enable them to survive unfavourable conditions. A modified form of hypnospore is the spore produced in *Botrydium* in the "gongrosira" stage, when portions of the rhizoidal cytoplasm round off into spores. Another modification is the endospore of the Chamaesiphonales, in which all or part of the cytoplasm is cut off successively into small, rounded, nonmotile bodies.

Algae also form other types of spores. Sometimes a complete vegetative cell develops a thickened wall and accumulates food reserves; the cell is then called an akinete. In filamentous algae, akinetes may be produced singly or in series. Many Chlorococcales do not undergo cell division; cells multiply when their contents divide into small new cells, termed autospores. Some brown algae (Phaeophyta) and certain primitive red algae (Rhodophyta) possess monospores; each is produced in a vegetative cell. More advanced members of the Rhodophyta have polyspores produced in structures known as polysporangia. In some cases these are reproductive structures of a triploid (3n) generation; in others they are derived from tetraspores produced in tetrasporangia. Tetrasporangia are also found in the Dictyotales (Phaeophyta).

After fertilization in most Rhodophyta, a parasitic diploid generation (carposporophyte) is produced on the female plant and gives rise to carpospores that normally germinate to a free diploid plant. In species of *Prasiola* (Chlorophyta), it has been found that some cells (meiospores) near the edge of a thallus undergo reduction division and produce a haploid appendage. The same phenomenon occurs in the "chantransia" phase of *Batrachospermum*. Other somewhat anomalous phenomena are the asexual cysts produced in the cap of *Acetabularia* and the statospores or thick-walled cysts formed inside the cells of some diatoms and members of the Chrysophyceae. Finally, there is the thick-walled product of gamete fusion, the zygote or zygospore, that germinates either to form a new plant directly or to form zoospores or aplanospores.

Sexual reproduction. Sexual reproduction usually occurs at the end of the growing season or is induced under unfavourable environmental conditions; it involves the production of motile male gametes and motile, sometimes nonmotile, female gametes. Motile male gametes are formed from unspecialized vegetative cells or specialized cells called antheridia. In the Rhodophyta, nonmotile male gametes (spermatia) are produced in special cells (spermatangia) that are borne singly or in structures called sori. Motile or nonmotile female gametes are produced in unspecialized vegetative cells or in a special structure, the oogonium (carpogonium in the Rhodophyta).

Gametes

Gametes of the opposite sex must unite (syngamy) to form a zygote. When two uniting gametes are morphologically identical, they are termed isogametes or + and − gametes. If the female member of the pair is somewhat larger and more sluggish in movement, the gametes are called anisogametes. Evolutionary advancement has resulted in a loss of mobility by the female gamete; this is termed oogamy. Generally the ova are liberated into a liquid medium where fertilization takes place; however, in some groups the ovum is retained on the female plant.

LIFE HISTORY

When two haploid gametes fuse to form a diploid zygote the chromosome number (see CELL AND CELL DIVISION) is doubled and must be reduced before gametes can be formed again. In some algae a reduction division (meiosis) occurs at the time of zygote germination; in others there is a diploid generation, and reduction division occurs at the time of spore formation. In cases in which only the diploid generation is known (*e.g.*, Siphonales), the reduction division takes place at the time of gamete formation. In certain algae, either haploid generation, diploid generation, or both can reproduce through spores, and a regular alternation of generations does not occur; in other algae, however, alternation of generations occurs regularly. Algae have alternate generations that differ both morphologically and cytologically (*i.e.*, heteromorphic generations) as is the case in some Rhodophyta; in other algae there is alternation of like generations (*i.e.*, isomorphic generations). The following scheme (with typical examples) summarizes the different types of life histories found in the algae.

In the monogenetic haplophasic life history, in which the plant is haploid and the zygote is diploid (represented in Figure 1, top), the haploid generation predominates, and asexual reproduction can occur. Reduction division occurs when the zygote germinates and the haploid is reduplicated. Figure 1, middle, represents regular alternation of free isomorphic generations (diplohaplophasic and di- or trigenetic cycles). In *Polysiphonia* and most Florideophycidae, there is, in addition to alternate generations, a parasitic carposporophyte generation; but since this is attached to the female sexual plant, it does not represent a separate morphological generation. The independent tetrasporophyte is morphologically the same as the sexual plants. Alternation of heteromorphic generations (diplohaplophasic digenetic cycle) is illustrated in Figure 1, bottom. In the Laminariales, the diplophase generation is the macroscopic plant and the gametophytes (haploid) are microscopic. In the rhodophyte alga *Asparagopsis armata* and some other Bonnemaisoniales, the haplophase is dominant in the life history; the independent tetrasporophyte is dwarfed to such an extent that, previously, these forms were placed in separate genera. Another example of the diplohaplophasic life history is an alternation of two generations that are heteromorphic but about equal in size. In *Halicystis* the asexual generation formerly was thought to be a distinct genus, *Derbesia*, while in *Cutleria* the prostrate asexual plants formerly were known as *Aglaozonia*. In the final life-history type, there is one kind of diploid plant (diplophasic monogenetic cycle), represented by members of the Siphonales.

Morphological phases

Based upon morphological criteria, the most primitive algae should have had life histories involving one kind of haploid plant (monogenetic haplophasic type). During the evolutionary process, if reduction division of the zygote to the haploid state was delayed, a free-living diploid generation could have developed. Thereafter the haploid generation would have become less significant and ultimately would have disappeared. A simplified life history occurs in the rhodophyte *Phyllophora brodiaei*, in which there is no free diploid phase, and tetraspores are produced in nemathecia on the female plant. In *Ahnfeltia* there is no reduction or fertilization, and the nemathecia produce only monospores.

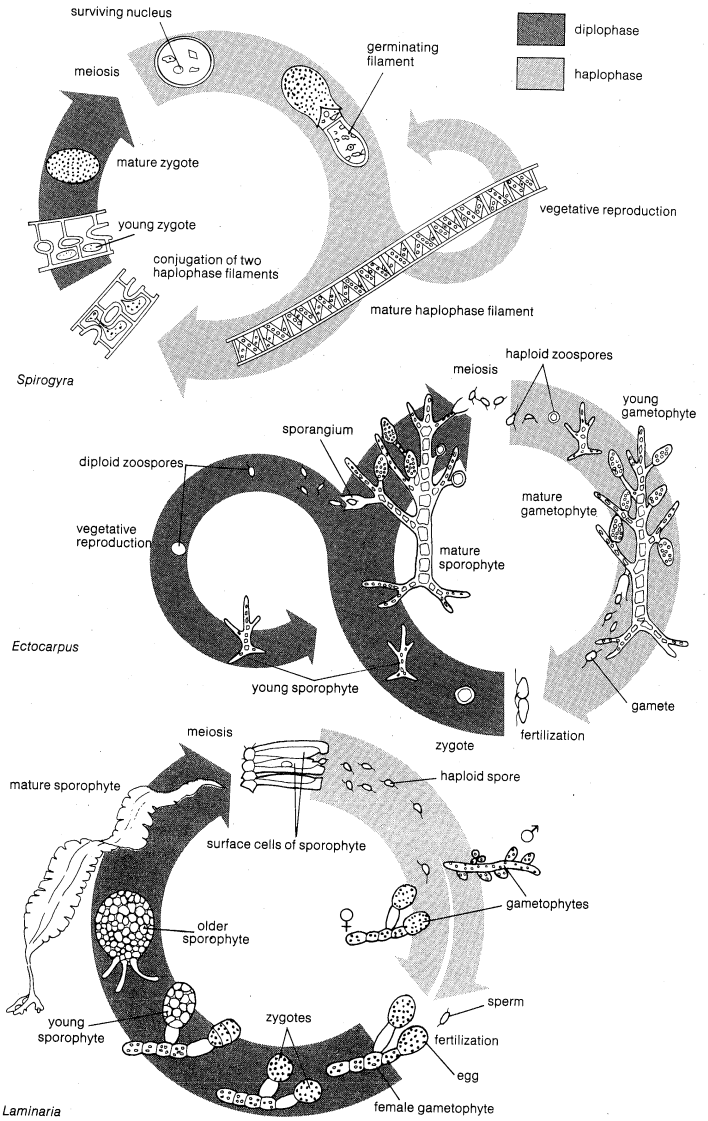


Figure 1: Reproductive cycles of *Spirogyra*, *Ectocarpus* and *Laminaria*.

Style from *Life: An Introduction to Biology*, 2nd ed., by George Gaylord Simpson and William S. Beck, copyright © 1957, 1965, by Harcourt Brace Jovanovich, Inc., and reproduced with their permission. Internal elements reprinted with permission of Macmillan Publishing Co., Inc., from *A Textbook of General Botany* by G. Smith *et al.* Copyright 1953 by Macmillan Publishing Co., Inc.

ECOLOGY

Habitats. On coastal areas the principal habitats occupied by marine algae are rocky shores, sand and sandy mud flats, salt marshes and mangrove swamps, estuaries and brackish water (bays or seas). Principal freshwater habitats of algae are dripping rocks and waterfalls; rivers and rapids; salty, cold, and hot springs; ponds, pools, ditches, and lakes. Algae also grow in nonparasitic associations with flowering plants or on nonliving matter (*e.g.*, rocks, fences). Diatoms, blue-green algae, and desmids can form mossy flora in wet habitats (*e.g.*, tropical rain forests). Distinct algal populations occur in the soil. In the sea and in freshwater, algal populations occur as phytoplankton, an important food source for fishes and the zooplankton. The distribution of marine phytoplankton is most abundant where upwelling cold water meets warm water. The nutrient cycles and temperature factors associated with phytoplankton development have been studied (see PLANKTON). When *Gonyaulax polyhedra* occurs in abundance in the sea, it creates the phenomenon known as red tides, and, in freshwater, blooms of *Microcystis aeruginosa* can be serious because they excrete into the water a toxin lethal to other forms of life. A specialized form of plankton, cryoplankton, includes those species that give a colour to snow fields (*e.g.*, green snow with *Chlamydomonas nivalis*).

Growth on rocky coasts and shores

Marine. On rocky coasts a succession of algal belts often can be recognized between the extreme high- and low-water marks of spring tides, a region called the littoral zone. Below the extreme low-water mark of spring tides is the sublittoral zone, where a few algal belts also can be recognized. Plant belts at the shore vary; they depend on exposure to wave action, latitude, and shore topography. Most of the shore falls in the eulittoral zone, which begins at the high-water mark, is subjected to tidal movement, and is divided into belts.

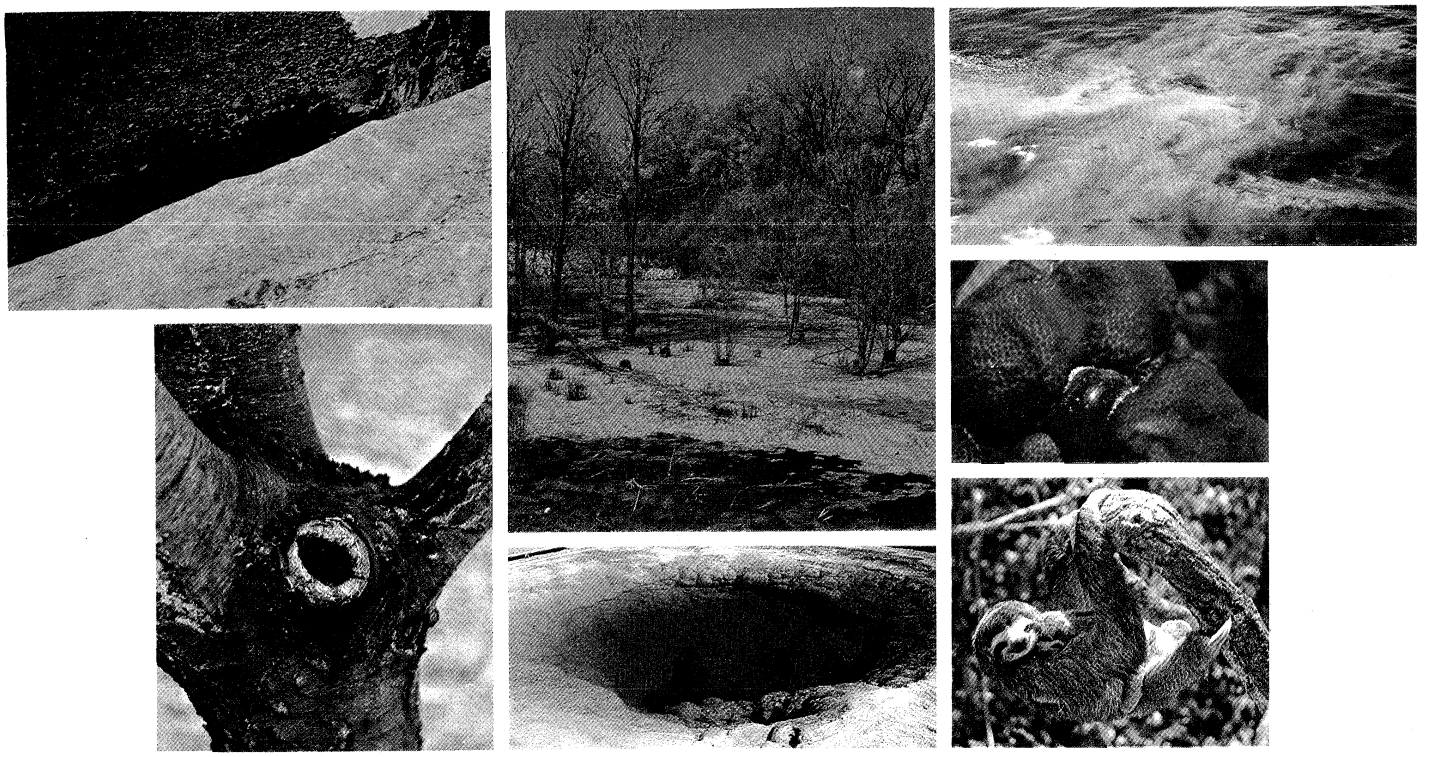
In many places a sublittoral fringe between the eulittoral and the sublittoral zones is distinguished by algae that can tolerate short periods of exposure but not deep submergence. In the cold Arctic and Antarctic waters, however, algae descend to considerable depths, and zonation of species occurs. In the Antarctic (Macquarie Island), for example, there is a sublittoral fringe belt that includes *Durvillea antarctica*, *Ascoseira*, and *Phaeurus*. Rhodophytes dominate the upper sublittoral zone, oarweeds (*Macrocystis* and *Lessonia*) the middle sublittoral zone, and rhodophytes (*Phycodrys* and *Phyllophora*) the lower sublittoral zone to the lower limit of algal growth.

In the Northern Hemisphere *Porphyra umbilicalis* and *Pelvetia canaliculata* are common in the supralittoral fringe (between the eulittoral and supralittoral zones), and the eulittoral zone sometimes is rich (especially on sheltered coasts) in fucoid species with which red algae are associated near low-water mark. Typically, the sublittoral fringe and sublittoral zone contain species of *Laminaria*. Fucoids, though present in North American waters, are not as important there as red algae. The laminarian belt is present in northeastern waters but disappears in southeastern ones; in the colder waters off the west coast of North America, the sublittoral laminarian forests extend as far south as Baja California. In Japanese waters these sublittoral laminarian forests are extensive and exploited commercially. Animals predominate in the eulittoral zone in southern temperate latitudes, and *Hormosira* and coralline red algae are the principal plants. In the warm-water flora of Carolina and Georgia, most species occur in the sublittoral zone. The lower eulittoral zone of the tropics commonly is characterized by growth of low turfy algae (*Cladophoropsis*, *Herposiphonia*, *Centroceras*) that survive during intertidal periods because water is retained within the turfy mass. In the tropical sublittoral zone, calcareous red algae (*Sargassum*)—important in coral reef growth—and members of the Codiaceae and Caulerpaceae occur.

On rocky coasts special algal habitats are provided by rock pools, gullies, and caves. Certain algae are restricted to rock pools, and many sublittoral species grow there. Gullies have peculiar light conditions and frequent tidal surges, and often contain species not found elsewhere. The reduced light intensity found in caves also promotes development of a characteristic flora; *e.g.*, the deepwater *Codium cranwelliae* in New Zealand caves.

Sandy beaches may contain a distinct flora of dinoflagellates; sand flats in river mouths often have a flora of euglenoids, *Ulva*, and *Enteromorpha*. In Europe *Fucus ceranoides* replaces other fucoids in brackish waters. Algal populations found in salt marshes and mangrove swamps throughout the world are more uniform than on rocky coasts, probably because environmental conditions in salt marshes and mangrove swamps are also more uniform. The principal algal population of mangrove swamps throughout the tropics is the *Bostrychia–Catenella–Caloglossa* red algal turf community that covers the pneumatophores and basal tree trunks of mangroves.

Freshwater. Among freshwater algal habitats are dripping rocks and waterfalls, often characterized by the rhodophyte *Lemanea* and a rich flora of Cyanophyta and Chlorophyta. Characteristic rhodophyte genera can be found in rapidly flowing rivers and rapids; sometimes *Compsopogon* is present in sufficient quantity to be a nuisance. In slower flowing rivers, *Entermorpha*, *Hydrodictyon*, *Oedogonium*, and other genera form floating



Diverse habitats of algae.
(Top left) Low temperature algae "red snow." (Top centre) Algal bloom on freshwater stagnant pond. (Top right) Marine algal "red tide." (Centre right) Unicellular marine algae colouring coral. (Bottom left) Algae covering the north side of a tree. (Bottom centre) High-temperature algae in hot springs at Yellowstone National Park. (Bottom right) Green algae living among the hairs of the three-toed sloth.
By courtesy of (bottom centre) Yellowstone National Park; photographs, (top left) Janet R. Stein, (top centre) E.R. Degginger, (top right and centre right) Photo Researchers, (top right) Carleton Ray, (centre right) Russ Kinne, (bottom left) Richard C. York, (bottom right) Carl W. Rettenmeyer

mats during the spring and summer. The distribution of algae often is used as an indicator of river water pollution. Highly polluted, oxygen-deficient waters contain only euglenoids, less polluted waters contain blue-green algae, and relatively unpolluted waters contain Chlorophyceae (a class of green algae). Some freshwater algae grow as epiphytes, using aquatic flowering plants as a physical base. An alga with an unusual habitat is *Basicladia*, which grows on turtles.

Algae in cold and hot springs

Cold springs, which usually contain alkaline ions, have an alkaliphilic flora of diatoms; saline springs have a rich flora since both marine and freshwater species thrive there. Hot springs possess a flora normally dominated by blue-green algae. In temperate regions the algal populations of bogs and swamps exhibit little periodicity, whereas those of ponds and pools show regular variations in flora. These changes in flora are associated with seasonal changes in light intensity, water temperature, and rainfall. Although they provide a more uniform aquatic environment than is found in bogs and swamps, lakes have a greater variety of habitats; these include shores, reeds, mud or boulder bottoms, and estuaries. Two main types of lake water are recognized, and each has a characteristic algal flora. Clear, clean, sparkling waters are characterized by many algal species present in small numbers. With gradual enrichment in nutrients resulting from plant decay and the introduction of soil from streams, the number of algal species decreases, but the number of individuals increases. A measure of the degree of enrichment is provided by the ratio Chlorococcales/desmids; the occurrence of blue-green algal blooms is an additional indication of enrichment. In lakes in which a seasonal fluctuation of water level occurs, distinct bands of algae can be found on the exposed littoral zone. Sometimes there is a distinct sublittoral zonation in lakes.

Soil. The soil algal flora consists of about 20 species of diatoms, 24 of Cyanophyta, and 20 of Chlorophyceae. Many of them have a worldwide distribution, but some are typical of certain soil types (*e.g.*, peat soils contain desmids, *Chroococcus, Gloeocapsa;* alkaline soils contain coccoid and filamentous Chlorophyceae). On the other hand, arid soils from Saudi Arabia, Mexico, Arizona, and Israel commonly have the same algae (*Chlorosarcinopsis, Chlorococcum diplobionticoideum*). There are two main groups of soil algae, the ephemerals (*Euglena*) and the perennials (diatoms, Cyanophyta). Many of the soil algae excrete polysaccharides and growth substances into the soil; many cyanophytes are nitrogen "fixers" that incorporate atmospheric nitrogen into chemical compounds (see BIOSPHERE).

Limiting ecological factors. *Wave action and tides.* One important ecological physical factor that limits the spread of certain algae is exposure to wave action, both on coasts and in lakes. This factor operates in two ways: by eliminating algae incapable of tolerating strong wave action, thus enabling other species that require surf to thrive; and by modifying growth belts, raising them as much as 40 feet (12 metres) on an exposed shore. At its upper limit this elevated area of growth is called the splash zone. The most important physical ecological factors on coasts, however, are tides and their secondary effects. The tidal range varies from place to place, affecting vertical extent of algal growth belts. Even in areas where the tidal range is small, characteristic narrow algal belts are found. In certain areas along shores, tidal currents may determine the prevailing species.

Effect of intertidal periods

The most significant effects of tides are intertidal periods and periods of continuous exposure to air; these occur during oceanic neap tides or during the summer months in lakes. Exposure of algae to air affects their rates of water loss, photosynthesis, and respiration; air exposure also causes changes in osmotic pressure of the cells and temperature changes in the thallus. Water loss, especially during periods of continuous exposure, apparently determines the upper survival limit for many algae on rocky coasts, salt marshes, and lake shores.

Some algae, which have evolved structures that aid in water retention, can survive longer periods of air exposure than others. In littoral fucoids, thick walls minimize water loss in some species, enabling them to grow higher on the shore, where periods of exposure are longer. The larger the water bladder of the fucoid *Hormosira*, the longer the organism can survive exposure to air. Decreased rates of photosynthesis during air exposure may affect upper survival limits. The kind of substrate (*e.g.*, rock, boulders, pebbles) determines the presence or absence of some species.

Biogeographic factors. In fucoids there appears to be a depth limitation since vesicles necessary for flotation are lost at critical pressures. Temperature operates primarily as a biogeographic factor; for example, Laminariales do not grow in warm waters. In some waters, another type of biogeographic change occurs; *i.e.*, the winter flora is northern in character, whereas the summer flora is subtropical. This kind of biogeographic change occurs east and west of the Cape of Good Hope in South Africa. In upper littoral rock pools water temperatures during summer exposure periods sometimes exceed the upper critical point for algae, and they die; algal growth is related directly to water temperature, and increases especially during spring and summer months in temperate latitudes are most important.

Effect of variations in light intensity

Light. In addition to tides, light is a major factor in determining the vertical distribution of algae. The effect of light is related to seasonal variation in light intensity at different latitudes, diurnal variations, and to the actual quantity and quality of light at different water depths. Silt deposits that occur on plants found at the mouths of estuaries, on salt marshes, and in mangrove swamps reduce the available light, and some algae cannot survive the resulting decrease in the rate of photosynthesis. Other algae achieve an optimum rate of photosynthesis at low light intensities; therefore, for a given species, the degree of wave action may determine whether maximum photosynthesis occurs at the water surface or below it. In addition, different phases in algal life histories may have different light demands; *e.g.*, gametophytes and sporelings of *Laminaria hyperborea* require less light than adult plants for maximum photosynthesis.

The pigments in marine algae may influence their optimum depths; for example, green algae occur in the littoral zone, brown algae in the upper sublittoral zone, and red algae in the lower sublittoral zone. The Phaeophyta (brown algae) contain fucoxanthinoid pigments that absorb light in the green spectral zone; as a result, many of these algae live at depths at which the light spectrum is predominantly blue and green. In the Rhodophyta (red algae) biliproteins absorb light in the green and yellow spectral regions found in deep waters. In some algae, however, the situation is not so simple; in *Tolypothrix* (a blue-green alga), for example, the proportion of pigments changes with available light. In other algae (*e.g.*, *Ulva*, a green alga, and *Porphyridium*, a red alga) there is intensity control; that is, an increasing amount of light causes decreases in available chlorophyll a or other pigments. Day length also is important in determining growth rate or development of reproductive organs for some algae.

Chemistry. Algae vary in their capacity to tolerate salinity changes. Many littoral algae can survive salt concentrations three times higher than that of seawater; sublittoral algae, however, will not tolerate a salt level higher than 1.4 times that of seawater. The algal flora of limestone shores is different in some ways from that of other shores. Acidity and dissolved oxygen changes generally are not significant for algal populations except in polluted waters. Nutrients are important in controlling the seasonal periodicity of plankton and in determining rates of enrichment of freshwaters.

Competition with other organisms. Competition between plants and animals sometimes is severe enough to eliminate algae: for example, *Macrocystis* (a laminarian) beds are decimated by sea urchins; establishment of *Fucus* species is prevented by limpets. On the other hand, barnacles must be established before *Pachymenia himantophora* can grow at a given site in New Zealand. Many specific host–parasite, host–epiphyte, and symbiotic relationships occur between algae and other organisms, and a relationship exists between flowering plant species and algae on salt marshes.

IV. Form and function

MORPHOLOGICAL CHARACTERISTICS

Blue-green algae. The cells of blue-green algae are unlike those of other algae. Blue-green algae (Cyanophyta) have a cell envelope composed of an outer mucilaginous sheath (sometimes pigmented as in *Lyngbya*), a middle pectin layer, and an inner wall of cellulose. In contrast to the cell walls of true algae, those of blue-green algae contain amino acids (*e.g.*, muramic, diaminopimelic) that occur also in bacterial walls. The cells of cyanophytes need not possess a cell membrane. In addition, blue-green algae contain no true vacuoles, although pseudovacuoles exist. Fine fibrils of deoxyribonucleic acid (DNA) and ribosome-like granules are scattered throughout the cytoplasm, and photosynthetic lamellae are arranged irregularly or in parallel stacks there. The structure of Cyanophyta indicates that they are related more closely to bacteria than to other algae.

The algal cell. *Cell wall.* Most algae have a two-layered cell wall; the inner wall consists of two or three layers of cellulose (xylan in Siphonales) and microfibrils (small fibrelike structures), while the outer one is amorphous pectin (a complex carbohydrate). Sometimes the wall contains silica (*e.g.*, diatoms), calcium (*e.g.*, *Halimeda*), or chitinous compounds (*e.g.*, *Oedogonium*). Next to the cell wall is the plasma membrane or pellicle; in some species it is rigid (*e.g.*, *Gymnodinium*), and in others (*e.g.*, *Euglena*) it is elastic enough to permit changes in shape. The cell wall can be surrounded by a shell or lorica (*e.g.*, *Dinobryon*), and in some forms it is covered with plates (*e.g.*, dinoflagellates), delicate scales (*e.g.*, Prasinophyceae), or mineralized scales (*e.g.*, Coccolithophoridaceae); in the diatoms the silica wall is penetrated by numerous pores. Pores exist also in other types of cells (*e.g.*, many members of the Rhodophyta, trumpet hyphal cells of *Laminaria*) and allow formation of intercellular cytoplasmic connections.

Cell-wall components

Cytoplasmic inclusions. The cytoplasm, surrounded by the plasma membrane, contains one or more vacuoles. In some algae (*e.g.*, *Valonia*) the vacuoles are so numerous and large that the cells are used in ion diffusion studies. Many of the simpler motile forms possess one or more contractile vacuoles (noncontractile in Pyrrophyta, the dinoflagellates) that probably function in excretion. Phaeophyta (brown algae) sometimes have refractive fucosan vesicles or physodes that are abundant, particularly in cells in which active metabolism or cell division is occurring. The cytoplasm contains mitochondria; the majority of algae contain xanthophylls that cause mitochondrial extensions. Other cellular inclusions are chondriosomes, a golgi-type apparatus (in desmids), and chloroplasts that are confined within a membrane and contain a number of lamellae. The number and shape of chloroplasts are variable in algal cells, and there is evidence that chloroplasts may rearrange themselves in response to light stimuli. Pyrenoids are organelles that generally lie in or on the surfaces of chloroplasts; they are proteinaceous granules that function in concentrating and accumulating starch. During zoospore formation pyrenoids may disappear from the cells and reappear in the swarmers. In the Rhodophyta (red algae) the cells of some species contain special bodies that cause iridescence.

Nucleus and flagella. Every algal cell contains at least one nucleus (and nucleolus); however, in some plants (*e.g.*, Siphonocladales, *Griffithsia*) individual cells contain many nuclei. Some algae have no inner cell walls except during the formation of reproductive bodies (coenocytes), and their numerous nuclei are scattered throughout the cytoplasm. The chromosomes, which commonly

number from 3 to 48, usually are small. Large chromosomes are found in the Cryptophyta in which the nucleus also is large. This phylum is interesting because spindle formation occurs at mitotic cell division even though centrosomes are absent (see CELL AND CELL DIVISION).

In *Chlamydomonas* and possibly other simple flagellated types, the centrosomes are united to small bodies called blepharoplasts found at the bases of flagella. This arrangement serves as a primitive form of neuromotor apparatus. Although the number may vary from one to four, algae usually have two flagella. Oedogoniales (a group of green algae) have motile bodies with a ring-shaped centrosome that gives rise to a ring of flagella. Flagella can be inserted into an alga apically (at one end) subapically, or laterally. Flagella used solely for locomotion are called isodynamic flagella. If one flagellum is used as a rudder or for attachment, however, the flagella are termed heterodynamic.

Motile algae and many reproductive bodies of non-motile algae contain a red eyespot, or stigma, that contains a pigment called hematochrome. Usually this apparently light-sensitive organelle is located near the anterior end of the cell. Some cells that lack the eyespot also may respond to light. The eyespot varies morphologically among algae; for example, Volvocales have layers of pigment chambers, while euglenoids have carotinoid granules that occur in a colourless matrix. Sometimes light is received and directed to the red eyespot by a lens-shaped plate.

Pigments. Pigments are important in algae and cyanophytes. An excess of one pigment (*e.g.*, fucoxanthin in brown algae, r-phycoerythin in red algae) is responsible for the overall colour of a group. Some colourless algae lacking pigments obtain nutrients either from dead organic matter (*i.e.*, saprophytic algae) or living organic matter (*i.e.*, parasitic algae); coloured algae manufacture their food by photosynthesis. Pigments found in algae include chlorophylls (a, b, and c), carotenes, xanthophylls, and biliproteins. Pigmentation in Chlorophyta (green algae) is similar to that in higher land plants (see COLORATION, BIOLOGICAL).

Trichocysts. Trichocysts are organelles that can be discharged as slime threads or club-shaped bodies. They may function as sensing or defensive devices. Trichocysts are distributed in the peripheral cytoplasm of Chloromonadophyta, line the gullet of Cryptophyceae, and originate in vesicles associated with the golgi-like bodies in dinoflagellates.

ALGAL METABOLISM

In most cases respiration follows normal metabolic pathways (see METABOLISM), but some algae (*e.g.*, *Iridaea, Polyneura, Cylindrospermum*) have other mechanisms. A few diatoms, some euglenoids, and *Gonium* (Volvocales) can use simple carbon compounds (*i.e.*, acetates) as nutrients. Some of the larger brown algae (*e.g.*, *Nereocystis, Pelagophycus*) produce carbon monoxide. Although about 20 cyanophyte species are capable of fixing nitrogen (see BIOSPHERE), normally they do so only when the chemical element molybdenum is present and other nitrogen compounds are not available readily. Maximum fixation occurs in winter and is minimal in summer. The rate of nitrogen fixation in *Tolypothrix* is equivalent to adding 2,900 pounds (1,315 kilograms) of nitrogen to an acre each year, and nitrogen fixation by Cyanophyta is essential to the growth of rice.

Accumulation of ions

Some algae can accumulate certain ions. Red and brown marine algae (*e.g.*, *Traillaella, Phyllophora, Fucus, Laminaria*) accumulate iodine from seawater. A definite influx "pump" (see MEMBRANE, BIOLOGICAL) for accumulation of potassium ions has been demonstrated in some algae (*e.g.*, *Hormosira*); others (*e.g.*, *Ulva, Hormosira, Valonia*) have an efflux "pump" that removes excess sodium. *Valonia* and *Halicystis* accumulate nitrate; the accumulation of silica by diatoms limits their distribution to areas where this element is available. Some vitamins (B_{12}, B, biotin,) essential for algal growth are synthesized by most algae; holozoic and saprophytic algae, however, obtain these compounds from an external source (see NUTRITION). Brown and red marine algae (*e.g.*, fucoids, *Rhodymenia*) excrete phenolic compounds into the sea. The effect of these compounds on the metabolic processes of the metabolism of other marine plants may be of ecological significance.

HORMONES

Growth hormones have been found in some algae, and a growth inhibitor (chlorellin) has been identified from *Chlorella.* A number of unicellular algae exhibit phototaxis, and, among large algae, *Bryopsis* and *Acetabularia* have a phototropic response to light. Algae exhibit, as do other plants, circadian rhythms in such phenomena as luminescence (dinoflagellates), photosynthesis (*Acetabularia*), and tidal rhythms in discharge of spores and gametes (*Halicystis, Dictyota;* see STEREOTYPED RESPONSES; PERIODICITY, BIOLOGICAL).

Sex hormones associated with the production of sexual gametes and their fusion have been recorded in a number of algae (*e.g.*, *Dasycladus, Oedogonium, Vaucheria*), though most have not been identified chemically. Despite their simple nature, some algae exhibit polarity (*e.g.*, *Enteromorpha, Bryopsis, Acetabularia*); for example, if a cut *Bryopsis* is reversed, the lower portion, formerly the apex, will produce rhizoids. An associated phenomenon is the production of two substances by the basal nucleus in *Acetabularia;* one initiates formation of certain features and the second, species specific, determines their shape. *Acetabularia* has been valuable in experimental morphogenetic studies. Nuclei of one species can be removed and replaced by those of the same or other species, or parts of plants can be removed and replaced by grafts of parts from other *Acetabularia.*

V. Evolution and phylogeny

FOSSIL ALGAE

Despite the fact that most algae have "soft" tissues and therefore cannot be preserved, skeletons of siliceous and calcareous algae have been found in the earliest fossil-bearing rock strata. In recent years Precambrian rocks have revealed very old fossil remains of algae. The earliest known algal fossils, more than 2,000,000,000 years old, are from the Gunflint chert of Ontario. They are well enough preserved to indicate affinities with the Cyanophyta (*e.g.*, *Animikiea septata* could have been a member of the Nostocales, *Entosphaeroides amplus* of the Chamaesiphonales). Other small septate filaments have been found in Precambrian rocks in Australia. All of these fossils, together with a group known as the Spongiostromata, probably are grouped best as the Protophyceae. The Spongiostromata, concretions (mineral deposits) associated with blue-green algae, are similar to existing aggregates found in the Bahamas. Columnar stromatolites (organic sedimentary structures) with which Cyanophyta are associated have been found in Precambrian rocks in Australia. The Porostromata (*Girvanella*) are another group with masses of entwined tubes that probably contained cyanophyte trichomes.

Earliest known fossil algae

Fossil dinoflagellates (*e.g.*, *Diconodinium*) have been recorded from the Jurassic (about 136,000,000 to 190,000,000 years ago). Another ancient (Cambrian, 500,000,000 to 570,000,000 years ago) flagellate group of uncertain relationships is the Hystrichosphaeridae.

Among the Chlorophyceae (green algae) the most ancient organisms, those of *Gunflintia,* are suggestive of present-day ulotrichalean filaments. The Siphonales are extensive from the Ordovician (430,000,000 to 500,000,000 years ago) onward; some genera (*e.g.*, *Abacella, Lancicula*) bear little relationship to present-day genera. In the Jurassic and Cretaceous (65,000,000 to 136,000,000 years ago) the Dasycladaceae were particularly abundant, and 48 out of the 58 known genera are extinct fossils (*e.g.*, *Cyclocrinus, Primicorallina, Paleodasycladus*), mostly from rocks of the Mesozoic (65,000,000 to 225,000,000 years ago). Charophytes (*Trochiliscus*) are known from the Devonian (345,000,000 to 395,000,000 years ago).

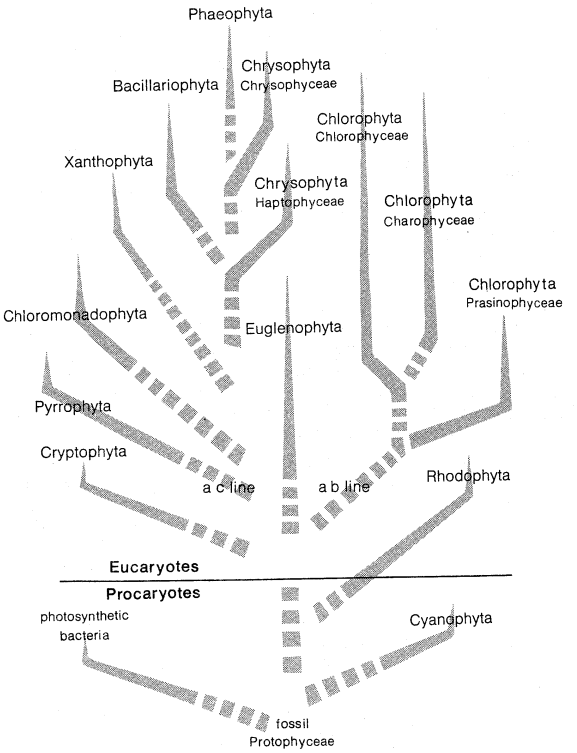


Figure 2: Family tree of algae.

There are some fossil algae from the Silurian (395,000,000 to 430,000,000 years ago) and Devonian that could have been phaeophytes; these include *Buthotrephis*, *Drydenia* and *Enfieldia*. Lime-encrusted nodular members (Solenoporaceae) of the Rhodophyta are known from the Ordovician and examples of *Melobesia* from the Cretaceous onward. Other ancient genera placed in the Rhodophyceae include the Silurian *Palmatophycus* and the Devonian *Thamnocladus*.

A fossil group of doubtful affinities, but perhaps of great significance in the migration of organisms to land, are the Devonian Nematophytales. These must be regarded as highly developed, almost terrestrial algae, possibly the most primitive of true land plants. The leaflike fossil *Nematothallus* was later found to be the photosynthetic organ of *Nematophyton*. Relationships have been suggested with both the Laminariales and the Chlorophyceae (Siphonales).

PRESENT PRIMITIVE FORMS

It is evident that evolution within the algae began very early and that many Mesozoic forms could be classified into present-day orders and families.

Antiquity of blue-green algae

The Cyanophyta have the most ancient history. Their development became practical after the evolution of pigments necessary for photosynthesis—chlorophyll a (possibly derived from simpler bacteriochlorophyll), and the development of a photosystem involving the biliproteins and secondary chlorophylls. After the evolution of photosynthesis, water became the electron donor and oxygen was produced. The presence of oxygen meant that organisms no longer were required to be anaerobes, and that carbon dioxide could be formed and used as a carbon source. The Cyanophyta contain both chlorophyll a and the biliproteins necessary for photosynthesis. The mechanism for the transition from these procaryotes to eucaryotes is not known, though it has been suggested that organisms like the thermal *Cyanidium caldarium* may represent a transition form.

Because of some of their morphological and biochemical features, the Rhodophyta and Cryptophyta are regarded as having evolved early in the geological time scale, probably from ancestral forms with some cyanophyte affinities. At the time of early cyanophyte evolution (from 2,000,000,000 to 2,500,000,000 years ago) the first wholly green algae (*Gunflintia*) were developing. With the evolution of these simple filaments, and even before, it is probable that single-celled green flagellates existed, providing an ancestral pool of primitive eucaryotes for subsequent evolution.

From this pool of eucaryotic ancestors two main lines have evolved (Figure 2). One main line of evolution is represented by organisms possessing chlorophylls a and b (the ab line) and starch-type food reserves (α - 1,4-glucoside linkages). This line contains the Prasinophyceae, Chlorophyceae, and Charophyceae. The Prasinophyceae and Charophyceae represent end lines in evolution; at the base of this line, the Euglenophyta probably split off as an independent end line. The ab line is essentially the Chlorophyta line from which land plants ultimately may have evolved, possibly from ancestral forms in the Chaetophorales. If the Nematophytales are algae-like and do represent migrant forms, it is probable that they were an abortive migrant attempt. There have been two major views about the origin of the Chlorophyceae. In one view the Chlamydomonadaceae are considered the most primitive; the other view regards the nonmotile colonial Palmellaceae as the most primitive. Evidence presently favours the former viewpoint. The logical sequence is from motile unicells to nonmotile or attached cells, then to some colonial (*e.g.*, *Hydrodictyon*) and siphonaceous forms (*e.g.*, *Protosiphon*). The attached unicell would develop readily into a simple filament (*Ulothrix* type) that might have evolved parenchymatous plate types (*Ulva*), filamentous partial coenocytes (Siphonocladales), total coenocytes (Siphonales), and finally branched heterotrichous types (Chaetophorales). Alternatively, the Siphonales could have arisen from a form such as *Protosiphon*.

The other main line of evolution is represented by those algae possessing chlorophylls a and c (the ac line) and laminarin type reserves (β - 1,3-glucoside linkages). Apparently the Pyrrophyta with their peculiar flagellar arrangement diverged early from this line, and their nearest relationships would be with the still more primitive Cryptophyta and Chloromonadophyta. Close similarities exist among the Xanthophyta, Bacillariophyta, and Chrysophyta, and, for many years, they were all included in one phylum. Now it is believed that they should be in separate phyla even though they probably evolved from what may have been a common ancestral plexus. Within the Chrysophyta, early development led to independent evolution of the Haptophyceae and Chrysophyceae. These evolutionary developments possibly occurred very early because the Phaeophyta represent the ultimate development of the ac line, and it seems likely that representatives of this group existed in the Cambrian or soon after. The Phaeophyceae may have arisen from a common stock with the Chrysophyceae or from an intermediate form within the Chrysophyceae such as *Phaeothamnion*. Within the Phaeophyta there appear to have been two main lines of evolution, one leading to the orders with isomorphic alternation of generations and the other to the orders with heteromorphic alternation of generations. The latter also developed along two different lines.

VI. Classification

DISTINGUISHING TAXONOMIC FEATURES

One principal criterion used by taxonomists to distinguish algae are differences in chlorophyll, carotenoid, and biliprotein pigments. Other important biochemical characteristics are the reserve food products and the assimilatory products of photosynthesis. The cell wall ultrastructure, its composition, whether of cellulose or other compounds, and the type and number of flagella are important. In certain cases differences in reproductive organs and life histories (especially in Phaeophyta and Rhodophyta) have been used. The Cyanophyta, which differ from all the algae in lacking nuclear, mitochondrial, and chloroplast membranes, usually are placed with bacteria in a group termed either Monera or Procaryota. Those who recognize the Monera include the algae with Protozoa in the Protista, the higher plants

forming the Metaphyta. Those who recognize the Procaryota include the algae and all other plants in the Eucaryota.

ANNOTATED CLASSIFICATION

The following classification is that used by V.J. and D.J. Chapman (1973). Those marked with an asterisk are classed also in the Protozoa.

PHYLUM CYANOPHYTA (blue-green algae)

Unicellular, colonial, and filamentous plants lacking distinct nuclei or mitochondria; nuclear material apparently localized; photosynthetic lamellae generally in outer cytoplasm; colour determined by proportion of chlorophyll a, carotenoids, and biliproteins; cell walls mainly of amino acids and amino sugars; product of photosynthesis cyanophycean starch; reproduction vegetative or asexual by simple division, fragmentation, spores, or nonmotile endospores.

Order Chroococcales. Unicellular free-living forms or colonies in a gelatinous matrix; reproduction by fission only; includes *Microcystis* (bloom former).

Order Chamaesiphonales. Attached, frequently elongate unicells, solitary or gregarious; reproduction by endospores.

Order Pleurocapsales. Cells arranged in prostrate clumps, expanded crusts or pseudofilaments; reproduction by endospores.

Order Nostocales. Plants multicellular filaments, with or without a sheath, simple or with false branching; filaments single or aggregated into a common sheath; in certain genera some cells differentiated to form heterocysts and akinetes; reproduction by fragmentation of short segments of filaments (hormogones).

Order Stigonematales. Filaments forming feathery tufts or mucilaginous clumps; true branching of uniseriate or multiseriate filaments; heterocysts known; reproduction by hormogones.

PHYLUM CHLOROPHYTA (green algae)

Unicellular, multicellular or colonial green algae; uninucleate, multinucleate or coenocytic; pigments characteristically those of higher plants; cell wall commonly of cellulose; food reserve starch.

***Class Prasinophyceae**

Flagellates characterized by 1 or 2 layers of platelike scales on flagella, sometimes also on the cell (*Micromonas*); flagella number 1 (*Micromonas squamata*), 2 (*Nephroselmis gilva*), or 4 (*Prasinocladus*); product of photosynthesis sometimes mannitol; reproduction by fission, aplanospores or zoospores.

Order Pyramimonadales. Free-living flagellates or unicells retained in a noncellulose wall or theca (*Prasinocladus*); cell division by fission; cyst formation also; *Platymonas convoluta* a symbiont of the wormlike *Convoluta roscoffensis.*

Order Halosphaerales. Large spherical uninucleate cells, reproducing by aplanospores, resting cysts, or typical Prasinophyceae swarmers.

Class Chlorophyceae

Motile or nonmotile single-celled, multicellular or colonial plants; uninucleate, multinucleate, or coenocytic; cell walls usually 2-layered, inner one cellulose (xylan in Siphonales), outer layer of pectin; some species with no rigid cell wall; chloroplasts, generally present, variable in number, size, shape, and normally contain pyrenoids; vegetative reproduction by fragmentation; asexual reproduction by zoospores; sexual reproduction by isogamy, anisogamy, or oogamy.

**Order Volvocales.* Small motile green flagellates (*Chlamydomonas*), regular colonial aggregates of 500 or more similar motile cells (*Volvox*), or irregular aggregates (*Tetraspora*) of nonmotile cells; in simpler forms all cells capable of simple fission or formation of zoospores, isogametes, anisogametes, or ova; in colonial forms differentiation into somatic and reproductive cells (*Pleodorina, Volvox*); some genera (*Polytoma*) without chloroplasts, others without a cellulose wall (*Dunaliella*).

Order Chlorococcales. Solitary, free-living or attached single cells or colonial aggregates (*Pediastrum*); cells without capacity for normal vegetative cell division; many genera form secondary carotenoids; reproduction by autospores; asexual reproduction by zoospores (rare); sexual reproduction by isogametes, anisogametes, and, in 1 case, oogamy; some forms (*Trebouxia*) an integral part of lichens, others (*Chlorococcum*) typical inhabitants of soil.

Order Ulotrichales. Unbranched filaments or derived parenchymatous types, attached to substrate by special basal cell(s) or free-living; heterotrichy in 1 genus (*Enteromorpha*); asexual reproduction by zoospores or meiospores (*Prasiola*); sexual reproduction isogamous (*Ulothrix*), anisogamous (*Enteromorpha*), or oogamous (*Cylindrocapsa, Sphaeroplea*); alternation of isomorphic (*Ulva*) or heteromorphic (*Monostroma*) generations.

Order Oedogoniales. Simple or branched filaments with modified attachment cell; cell division highly specialized; zoospores and gametes with ring of flagella; sexual reproduction is an advanced oogamy, some species of *Oedogonium* with dwarf male plants produced from special swarmers (androspores).

Order Chaetophorales. Fundamentally heterotrichous but sometimes only erect (*Draparnaldia*) or prostrate disc (*Ulvella*) present; erect portion usually a uniseriate, branched filament, or reduced (*Pleurococcus*); asexual reproduction by macrozoospores or microzoospores, aplanospores, or akinetes; sexual reproduction isogamous, anisogamous (*Aphanochaete*), or oogamous (*Coleochaete*); contains epiphytes (*Cephaleuros*) and endophytes.

Order Siphonocladales. Multinucleate either simple (*Lola*) or branched filaments (*Cladophora*), or vesicle-like (*Valonia*); vesicle-like forms with specialized cell division; chloroplast netlike; asexual reproduction by zoospores; sexual reproduction isogamous or anisogamous; alternation of isomorphic (*Microdictyon*) or heteromorphic generations (*Spongomorpha*), or haploid generation suppressed (*Valonia*).

Order Dasycladales. Acetabularia; thallus macroscopic, uninucleate, with whorled appendages, plastids disclike; reproduction of diploid generation by isogamete-producing cysts.

Order Derbesiales. Gametophytic plants, oval vesicles (*Halicystis*); sporophytic plants, branched siphonaceous thread (*Derbesia*); sexual reproduction anisogamous, asexual zoosporic.

Order Siphonales. Coenocytic filaments frequently arise from prostrate rhizoidal portion; cell membrane of mannan or xylan, characteristic pigment siphonoxanthin; vegetative reproduction by propagules (*Codium*) or rhizoidal (*Halimeda*); sexual reproduction mainly anisogamous.

Order Dichotomosiphonales. Filaments siphonaceous, lacking siphonoxanthin; sexual reproduction a distinct oogamy.

Order Zygnematales (or *Conjugales*). Unicellular (desmids) or filamentous (*Spirogyra*), free-floating; chloroplast large and varied; specialized sexual reproduction (conjugation) without motile gametes.

Class Charophyceae

Filamentous, branches in whorls; cells large; sexual reproduction only; highly specialized oogamy; 1 order, Charales.

***PHYLUM EUGLENOPHYTA** (euglenoids)

Unicellular flagellates and a few palmelloid forms with or without chlorophyll; pigments, when present, those of higher plants; storage reserve paramylum.

***Class Euglenophyceae**

Organisms with pellicle only; generally 2 flagella, 1st with row of hairs, 2nd greatly reduced; cells with reservoir, contractile vacuoles, sometimes a red stigma; reproduction by cell division; cysts known.

Order Rhabdomonadales. Colourless saprophytic flagellates; 2nd flagellum reduced; no euglenoid movement; no eyespot.

Order Heteronematales. Colourless phagotrophic flagellates; 2nd flagellum emergent; no eyespot; lack euglenoid movement.

Order Sphenomonadales. Colourless saprophytic or phagotrophic flagellates; no eyespot; 2nd flagellum emergent; no euglenoid movement.

Order Euglenamorphales. Saprophytic endosymbionts in digestive tract of tadpoles.

Order Eutreptiales. Flagellates with 2 emergent flagella; with euglenoid movement.

Order Euglenales. Coloured and colourless (*Astasia*) forms, single or forming dendroid or palmelloid colonies (*Colacium*).

PHYLUM XANTHOPHYTA (yellow-green algae)

Single-celled, colonial, filamentous or siphonaceous forms with chlorophyll a and specific carotenoids; motile bodies with 2 equal flagella, 1 short and smooth, 1 long and pleuronematic; food storage, oil or chrysolaminarin.

Class Xanthophyceae

In nonflagellated and filamentous forms cell wall of 2 overlapping halves; pyrenoids usually absent; asexual reproduction

by zoospores, aplanospores, or statospores, sexual reproduction rare, isogamous, or oogamous (*Vaucheria*).

**Order Chloramoebales* (or *Heterochloridales*). Naked, motile, flagellated single cells, sometimes amoeboid; reproduction by cell division or statospores.

**Order Rhizochloridales.* Vegetative phase permanently amoeboid; envelope or lorica may be present; reproduction by cell division.

Order Heterogloeales (or *Heterocapsales*). Cells aggregated into gelatinous thallus; individual cells can revert to motile phase; reproduction by cell division; thick-walled statocysts also known.

Order Mischococcales (or *Heterococcales*). Single celled or colonial (*Mischococcus*); cell division unknown; reproduction by zoospores or aplanospores.

Order Tribonematales (or *Heterotrichales*). Mostly unbranched filaments; reproduction asexual.

Order Vaucheriales (or *Heterosiphonales*). Filamentous or saccate; asexual reproduction by zoospores or compound zoospores; sexual reproduction isogamous (*Botrydium*) or oogamous (*Vaucheria*).

***PHYLUM CHLOROMONADOPHYTA**

Class Chloromonadophyceae

Naked, unicellular, biflagellate cells with trichocysts in peripheral cytoplasm; some colourless (*Thermatomastix*); pigments as in Xanthophyceae; food reserve oil; 1 order, Chloromonadales.

PHYLUM BACILLARIOPHYTA (diatoms)

Unicellular or colonial; cell wall in 2 halves built of silica on pectin foundation; highly ornamented by striae; major pigments chlorophylls a and c and fucoxanthin; principal photosynthetic product chrysolaminarin.

Class Bacillariophyceae

Radially or bilaterally symmetrical single cells; cell division produces 2 cells of slightly unequal size; sexual reproduction by oogamy or amoeboid isogamy.

Order Centrales. Cells with radial symmetry; sexual reproduction by oogamy.

Order Pennales. Cells variously shaped, bilaterally symmetrical, 1 or both halves with raphe, or pseudoraphe; sexual reproduction amoeboid isogamy.

PHYLUM CHRYSOPHYTA (golden algae)

Mostly single celled; many without cellulose wall, others with silicification or calcification; pigments, chlorophylls a and c, fucoxanthin; photosynthetic product chrysolaminarin; reproduction by cell division, zoospores, or statospores.

Class Chrysophyceae

Flagellates with 2 unequal flagella, 1 smooth, the other pleuronematic, or with single pleuronematic flagella.

**Order Chrysomonadales.* Motile cells with 1 or 2 flagella, some with external shell (lorica), others with rigid periplast; some colonial (*Synura*).

**Order Rhizochrysidales.* Amoeboid with temporary typical flagellate phase.

Order Chrysocapsales. Nonmotile cells aggregated into a mucilage; reproduction by fragmentation or motile cells.

Order Chrysosphaerales. Single celled or colonial, nonmotile; cell division internal; some forms cystlike.

Order Chrysotrichales (or *Phaeothamniales*). Branched or unbranched filament with modified basal cell; palmelloid stages known.

***Class Haptophyceae**

Flagellates with equal smooth flagella; scales, if present, not mineralized; usually a 3rd anchorage organ (haptonema).

Order Isochrysidales. Flagellate stage without haptonema.

Order Prymnesiales. Haptonema present; some flagellates with calcified scales (Coccolithophoridaceae).

PHYLUM PYRRHOPHYTA (dinoflagellates)

Mostly motile single cells, some short filaments; motile cells with 2 laterally or apically attached flagella, 1 encircles cell, the other trails behind; cells with longitudinal and transverse furrow, large nucleus; trichocysts present; pigments in coloured forms chlorophylls a and c with peridinin the specific xanthophyll; product of photosynthesis starch; reproduction by cell division, zoospores, or isogamy; fine-structure studies indicate there may be no basis for two classes.

Class Desmophyceae

Cells with periplast or cellulose envelope; flagella apically inserted; transverse furrow lacking.

**Order Prorocentrales.* 1 flagellum projecting forward, other at right angles; reproduction by cell division.

Class Dinophyceae

Cells with periplast or cellulose envelope with ("armoured") or without ("unarmoured") sculptured plates; flagella laterally attached; transverse and longitudinal furrows present; transverse flagellum with single row of fine hairs, longitudinal flagellum pleuronematic.

**Order Gymnodiniales* (unarmoured dinoflagellates). Most colourless, holozoic; reproduction by division of motile or encysted cells.

**Order Peridiniales* (armoured dinoflagellates). Most with chloroplasts; reproduction by division of motile or encysted cells, by isogametes (*Glenodinium*), or by amoeboid anisogametes (*Ceratium*); some species luminescent, some major components of "red tides" (*Gonyaulax*).

Order Dinophysidales. Cell wall with 2 large plates, some smaller ones; many species without chloroplasts.

Order Dinocapsales. Nonmotile cells enclosed in mucilage give rise to dinoflagellate zoospores.

Order Dinotrichales. Simple or sparsely branched, few-celled filaments; reproduction by zoospores.

**Order Dinamoebidiales* (or *Rhizodiniales*). Vegetative phase permanently amoeboid; reproduction by dinoflagellate zoospores.

Order Phytodiniales (or *Dinococcales*). Nonmotile single cells without plates; no cell division; reproduction by zoospores or autospores.

PHYLUM PHAEOPHYTA (brown algae)

Chlorophylls a and c masked by fucoxanthin.

Class Phaeophyceae

Plants from minute discs or branched filaments to erect, tufted, branched, uniseriate or corticated filaments to more complex thalli with uniaxial or multiaxial construction, to parenchymatous thalli of simple or highly complex structure, often heterotrichous or with anchoring organs and basal–distal distinction; principal products of photosynthesis mannitol and laminarin; cell walls include complex polysaccharides, motile reproductive cells with 2 unequal, laterally inserted flagella, the longer pleuronematic, the shorter smooth; asexual reproduction by zoospores, aplanospores, monospores, or tetraspores; sexual reproduction isogamous, anisogamous, or oogamous.

Order Ectocarpales. Branched uniseriate, heterotrichous filaments, microscopic or macroscopic with intercalary growth, aerial thallus sometimes lacking (*Phaeostroma*); reproduction isogamous (*Ectocarpus*) or anisogamous (*Giffordia*); alternation of generations isomorphic.

Order Tilopteridales. Branched, uniseriate (later multiseriate), filaments arise from basal disc; growth intercalary; asexual reproduction by monospores; sexual reproduction oogamous; life history alternation of unlike generations.

Order Sphacelariales. Branched uniseriate filaments show apical growth by tannin-containing apical cell, become multiseriate; heterotrichous; vegetative reproduction by propagules (*Sphacelaria*) or fragmentation (*Halopteris*), asexually by zoospores, sexually by isogametes or anisogametes; generally alternation of isomorphic generations.

Order Cutleriales. Branched, flat, ribbonlike or fan-shaped thalli with trichothallic growth; alternation of isomorphic generations (*Zanardinia*), heteromorphic generations (*Cutleria*) or asexual generation only (*Microzonia*); sexual reproduction anisogamous.

Order Dictyotales. Plants ribbonlike (*Dictyota*) or fan-shaped (*Padina*), dichotomously divided, parenchymatous, with medulla of large cells and cortex of small cells; growth by single apical cell (*Dictyota*) or marginal meristem (*Padina*); holdfast padlike; some genera (*Dictyopteris*) with thickened "midrib"; usually a regular alternation of isomorphic generations; asexual reproduction by tetraspores; sexual plants dioecious, sexual reproduction oogamous.

Order Chordariales. 1 or more axial branched filaments arising from prostrate basal portion; main filaments (medulla) produce interwoven lateral branches (subcortex) terminating in peripheral assimilatory filaments and hairs (cortex); growth commonly trichothallic, rarely apical (*Spermatochnus*); basic thallus can be modified (*Leathesia, Elachista*) or greatly reduced (*Myrionema*); generally macroscopic diploid sporophyte, microscopic gametophyte, alternation usually not regular; sometimes an asexual generation only (*Acrothrix*); asexual reproduction by zoospores; sexual reproduction generally isogamous, rarely anisogamous (*Notheia*).

Order Sporochnales. Thallus pseudoparenchymatous with intercalary trichothallic meristem at base of terminal tuft of hairs; sporangia grouped into swollen, oval, or elongate receptacles; sexual generation microscopic, probably reproduces oogamously.

Order Desmarestiales. Plants pseudoparenchymatous, bushy; axis with pinnately arranged branches terminating in branched uniseriate filaments; cortex of older parts of plants derived from basal cells of primary branches; growth trichothallic; lateral filaments deciduous (*Desmarestia, Arthrocladia*); regular alternation of macroscopic sporophyte and microscopic heteromorphic gametophytes; sexual reproduction oogamous.

Order Dictyosiphonales. Thallus parenchymatous, solid (*Myriotrichia, Scytothamnus*), tubular (*Scytosiphon*), saccate (*Colpomenia, Adenocystis*), or foliaceous (*Petalonia*); growth intercalary; large-celled medulla with cortex of smaller cells bearing assimilatory hairs and sporangia; alternation heteromorphic; gametophyte forming dwarf generation or a leafy thallus.

Order Laminariales. Sporophyte large, with distinct holdfast, stipe, and expanded lamina; growth zone intercalary; structure pseudoparenchymatous (*Chorda*) or parenchymatous, with central, sometimes complex, medulla, wide cortex, and outermost meristoderm; asexual reproduction by sporangia in sori on frond or on special leaflets (*Alaria*); gametophytes microscopic, dioecious; reproduction oogamous; includes largest algae (*Macrocystis, Nereocystis, Pelagophycus*).

Order Ascoseirales. Monotypic Antarctic order; plants with short stipe arising from padlike holdfast produce a single blade that splits; complex medulla like Laminariales, cortex and limiting layer like Fucales; reproduction by chains of sporangial cells borne in sunken regions.

Order Durvilleales. Thallus with large basal pad, short stipe, long dissected flattened frond; growing region marginal; sexual organs in sunken regions scattered over entire thallus; oogonia borne on branched filaments; embryo lacks apical hairs and groove.

Order Fucales. Thallus normally differentiated into discoid holdfast, stipe, and frond often with leaflike organs (*Sargassum*); growth in length by single apical cell, rarely by a group (*Hormosira*); external meristoderm, parenchymatous cortex, and hyphal medulla; growth in thickness by secondary hyphae and new cortical cells; surface of expanded part of frond with cavities (cryptostomata) containing tufts of hairs; reproduction oogamous with sex organs in localized sunken regions, oogonia borne directly on wall of receptacles; embryo with apical hairs and groove.

PHYLUM RHODOPHYTA (red algae)

Chlorophyll a masked by biliproteins.

Class Rhodophyceae

Basic structure of thallus a branched filament or multifilamentous axis; products of photosynthesis floridean starch, floridoside, and mannoglycerate; motile reproductive stages unknown; protoplasmic connections between most cells; sexual reproduction a complex oogamy, zygote gives rise to a parasitic generation, carposporophyte, whose spores in turn usually give rise to an independent asexual generation reproducing by tetraspores or yielding a new gametophyte vegetatively after meiosis in an apical cell (*Batrachospermum*).

Subclass Bangiophycideae

Plants single celled, filamentous or membranous; cells with single axile chloroplast; growth of thallus diffuse; sexual reproductive apparatus unspecialized; asexual reproduction by monospores; life cycle simple, gametophyte dominant.

Order Porphyridiales. Unicellular, often aggregated into gelatinous colonies; multiplication by cell division in all planes, by monospores or by endospores.

Order Goniotrichales. Uniseriate or multiseriate, minute, simple, or branched filaments; reproduction by monospores or akinetes.

Order Compsopogonales. Branched filaments with axis of large cells surrounded by cortex of small, highly pigmented cells; reproduction by monospores.

Order Rhodochaetales. Minute branched filaments; apical growth; intercellular pits; reproduction by monospores; sexual reproduction a primitive protogamy.

Order Bangiales. Attached, unbranched, uniseriate and later multiseriate filaments or expanded mono- or distromatic fronds; reproduction by monospores and sexually by simple spermatia and carpogonia; apparent alternation of heteromorphic haploid generations (*Porphyra*); dwarf form known as the "conchocelis" phase.

Subclass Florideophycideae

Plants of branched filaments, compact pseudoparenchymatous, or membranous thalli; many heterotrichous; growth from 1 or more apical cells; sexual plants dioecious; carpogonial branch complex, involved in transfer of fertilized nucleus to special auxiliary cell; alternation generally isomorphic, heteromorphic in simpler forms.

Order Nemalionales. Vegetative structure uniaxial or multiaxial, often not compact, many genera lack true auxiliary cells (*Batroschospermum*); alternation, where known, heteromorphic (*Batrachospermum, Lemanea catenata*), but isomorphic in *Acrochaetium* species, dwarf generation being diploid.

Order Bonnemaisoniales. Thallus construction uniaxial and multiaxial, compact; auxiliary cell develops from carpogonial branch cell, is largely nutritive; isomorphic (*Delisea*) or heteromorphic (*Asparagopsis*) alternation of generations, dwarf phase diploid; sometimes sexual generation only (*Naccaria wiggii*).

Order Gelidiales. Construction compact, uniaxial; auxiliary cells nutritive only; alternation of isomorphic generations.

Order Cryptonemiales. Encrusting or erect, uniaxial or multiaxial, 1 family (Corallinaceae) heavily lime-encrusted; carpogonial branches sometimes grouped; auxiliary cells formed on accessory branches before fertilization; isomorphic alternation of generations.

Order Gigartinales. Uniaxial or multiaxial, often much modified, crustlike (*Cruoria*), membranous (*Ahnfeltia*) or expanded (*Gigartina* species); auxiliary cell usually an ordinary vegetative cell set aside before fertilization; alternation isomorphic.

Order Rhodymeniales. Construction wholly multiaxial, central portion sometimes hollow (*Champia*); auxiliary cells cut off before fertilization from branch derived from support cell; alternation isomorphic.

Order Ceramiales. Construction uniaxial but much modified to give wide range of form, *e.g.*, branched filaments (*Callithamnion*), leafy expansions (*Polyneura*), nets (*Martensia*); cartilaginous forms (*Laurencia*), polysiphonous forms (*Polysiphonia*); auxiliary cell cut off from support cell after fertilization; alternation isomorphic.

PHYLUM CRYPTOPHYTA

Colourless (*Chilomonas*) or coloured organisms, the latter with chlorophylls a and c, specific xanthophylls and biliproteins; food storage a starchlike compound; reproduction solely by longitudinal fission.

***Class Cryptophyceae**

Compressed, nominally naked, asymmetrical cells with slightly unequal flagella inserted laterally or apically; 1 genus of few-celled filaments (*Bjornbergiella*); gullet of motile bodies lined with trichocysts.

Order Cryptomonadales. Cells motile flagellates with characteristic asymmetry (*Cryptomonas*) or kidney shaped (*Hemiselmis*).

Order Cryptococcales. Cells nonmotile or forming very short filaments.

CRITICAL APPRAISAL

The taxonomy of the algae is fluid; as new discoveries occur, especially concerning pigments, flagella, and life cycles, taxonomic adjustments are necessary. Numbers of subphyla, orders, and families depend upon specific taxonomic viewpoints; for example, among the Chlorophyta the number of orders depends upon the subdivision of the Ulotrichales. Chaetophorales, Siphonocladales, and Siphonales also can be subdivided in various ways. In phyla with flagellate members both electron microscopy and pigment analysis are valuable diagnostic tools; flagellate phyla are in a state of flux. Among the phylum Chrysophyta, evidence supports recognition of a class Haptophyceae, although it is not always recognized. In the Phaeophyta a major alternate classification divides the phylum into Cyclosporeae (diploid generation only), Isogeneratae (isomorphic alternation), and Heterogeneratae (heteromorphic alternation of generations). Recent studies support the separation of the Ascoseirales and Durvilleales from the Fucales where they formerly were included. In the Rhodophyta some workers include members of the Bonnemaisoniales in the Nemalionales. Intensive work on life histories and cytology of primitive red algae may result in major changes in some orders.

BIBLIOGRAPHY. General texts covering the algae include those by F.E. FRITSCH, *Structure and Reproduction of the Algae*, vol. 1 (1933) and 2 (1945); G.M. SMITH (ed.), *Manual of Phycology: An Introduction to the Algae and their Biology* (1951); F.E. ROUND, *Biology of the Algae* (1965), in which particular attention is paid to the biology of the flagellate groups and to freshwater ecology; E.Y. DAWSON, *Marine Botany* (1965); G.W. PRESCOTT, *The Algae: A Review* (1968), which covers the field generally; T. LEVRING, H.A. HOPPE, and O.J. SCHMID, *Marine Algae: A Survey of Research and Utilisation* (1969), with an extensive bibliography; and V.J. and D.J. CHAPMAN, *The Algae* (1971), which covers all fields except that of economic uses.

The economic uses of algae are dealt with in detail by V.J. CHAPMAN, *Seaweeds and Their Uses*, 2nd ed. (1970); and in general terms by D.F. JACKSON (ed.), *Algae, Man, and the Environment* (1968).

Reference works that deal specifically with the physiology of the algae are those by R.A. LEWIN (ed.), *Physiology and Biochemistry of Algae* (1962); and G.E. FOGG, *Metabolism of Algae* (1967); while M.B.E. GODWARD (ed.), in *Chromosomes of the Algae* (1966), describes what is known of this aspect of the algae.

G.E. FOGG, *Algal Cultures and Phytoplankton Ecology* (1965), covers two other specialized fields; and finally V.J. CHAPMAN, *Coastal Vegetation* (1964), deals with the marine algal ecology of the British Isles.

(V.J.C.)

Algebra, Elementary and Multivariate

Algebra may be described as a generalization and extension of arithmetic. Elementary arithmetic is concerned primarily with the effect of certain operations, such as addition or multiplication, on specified numbers, hence, for instance, the multiplication tables; elementary algebra is concerned with properties of arbitrary numbers. For instance, the fact that 2 added to 3 gives the same result as 3 added to 2 is one of arithmetic; the formula $a + b = b + a$ for all numbers a, b is one of algebra.

HISTORY

The particular operations of arithmetic that came to be extended and generalized to provide the materials of algebra emerged only slowly. The earliest writings on the subject dealt with many topics that are not now regarded as part of algebra. (For a treatment of the evolution of algebra as a well-defined branch of mathematics, see MATHEMATICS, HISTORY OF.) The earliest extant work with any claim to be regarded as a treatise on algebra is by the Greek philosopher Diophantus of Alexandria (*c.* AD 250). This work is devoted mainly to problems in the solution of equations. For this purpose a suitable notation had to be invented, and Diophantus gave rules for generating powers of a number and for the multiplication and division of simple quantities. Of great significance is his statement of the laws governing the use of the minus sign, which did not, however, imply any idea of negative quantities.

The work of Diophantus

During the 6th century the ideas of Diophantus were greatly improved on by various Hindu mathematicians, and many deficiencies in the Greek symbolism were remedied. The development of symbolic algebra by the use of general symbols to denote numbers is due to a 16th-century French mathematician, François Viète, a usage that led to the idea of algebra as generalized arithmetic. Sir Isaac Newton gave it the name Universal Arithmetic in 1707.

The main step in the modern development of algebra was the evolution of a correct understanding of negative quantities, contributed in 1629 by a French mathematician, Albert Girard, whose work was later overshadowed by that of his contemporary, René Descartes. While it is convenient to view Descartes's work as the starting point of modern algebra, for the sake of clarity notations and terminology will be used that were developed later.

Algebra is concerned with certain operations on numbers, and it is necessary to be precise about what a number is. The numbers dealt with are either the natural numbers, 0, 1, 2, 3, 4, · · · (some authors exclude 0); rational numbers, which have the form p/q, in which p and q are integers (natural numbers and their negatives) with $q \neq 0$; real numbers, which correspond to all the points on a line; or complex numbers, which are constructed from the real numbers together with a number i, the square of which is -1. The essential property that these numbers have is that they can be added and multiplied by well-established rules. Arithmetic becomes algebra when general rules are stated regarding these operations, as, for example, the commutative law of addition (see below). First the basic algebraic properties of such numbers are considered.

(1) $a+b=b+a$

(2) $(a+b)+c=a+(b+c)=a+b+c$

(3) $a+0=a$

(4) $a+(-a)=0$

(5) $ab=ba$

(6) $a\times(bc)=(ab)\times c=abc$

(7) $a\times(b+c)=ab+ac$

(8) $1\times a=a=a\times 1$

(9) $0a=0$, and if $ab=0$, either a or b is zero.

BASIC THEORY

Elementary operations. Addition has the following properties: The order of addition does not affect the result (see Box, equation 1)—that is, addition is commutative. If three numbers are being added, parentheses can be placed around either the first two or the last two, while the terms in the parentheses are added first, and their sum is added to the remaining term. The two results are, however, always equal (see 2). This second property is expressed by saying that addition is associative, and as a result either sum may be expressed as $a + b + c$. There is a number 0 (zero) such that the sum of any number a and zero is a (see 3). Corresponding to each number a there is a number $(-a)$ such that the sum of a and $(-a)$ is zero (see 4). From the second and fourth of these laws it follows that for any two numbers a, b, there is a unique number x such that $a + x = b$, namely, $x = b + (-a)$, more commonly written $x = b - a$. A set of numbers with the law of addition satisfying the above properties is called an additive group.

Laws of addition and multiplication

The result of multiplication of two numbers a, b is usually denoted by ab, but sometimes to avoid confusion $a \times b$ is written.

Multiplication of numbers has the following properties: The order of multiplication does not affect the result (commutativity; see 5) nor does the way the numbers are placed in parentheses (associativity; see 6). The product of a number and the sum of two numbers equals the sum of two appropriate products. This last is known as the distributive law (see 7). There exists a unique element 1 (unity) such that the product of any number a and 1 is a (see 8).

Any set of elements in mathematics that has two laws of composition, addition and multiplication, such that (1) with addition as the law of combination it forms an additive group, and (2) with multiplication is commutative and associative, and distributive over addition, is called a ring. It is because so much of modern algebra is concerned with groups and rings that it is important to realize that ordinary numbers have these properties. Because the numbers have unity the ring is a ring with unity. One other property of numbers that should be noted is that the product of 0 and a is 0 for all a, and if $ab = 0$, either a or b is 0 (see 9). With this last property the ring with unity becomes what is called an integral domain.

All accepted systems of numbers, integers, rational numbers, and real and complex numbers have the above properties (see ARITHMETIC; ANALYSIS, REAL). If the system has the property that when a is different from zero there is a number x such that $ax = 1$ (see 10), then the nonzero numbers of the system form a group in which multiplication is the group operation, and the integral domain is a field. The systems consisting of the rational, real, or complex numbers form a field, but the system of integers does not.

Complex numbers and root extraction. The operations so far considered—namely, addition, subtraction, multiplication, and division (in the case of fields)—are known as the elementary operations of algebra. In another section the extension of these operations to more complex systems, such as polynomials (see below), will be considered. One other operation on real or complex numbers must be considered here. A theorem of real variables states that: If a is any positive real number and n is any positive integer, there exists a unique positive real number x such that $x^n = a$, in which x^n is the product of n factors each equal to x.

It is important to note that if a is an integer (rational number, negative real number) there is, in general, no integer (rational number, real number) such that $x^n = a$. Thus this operation called extracting the nth root is not a satisfactory algebraic operation as it stands. To obtain a satisfactory operation, it is necessary to use complex numbers.

Definition of a complex number

A complex number, symbolized here by the Greek letter alpha, α, is any formal combination $\alpha = a + ib$ of two real numbers a, b, with addition and multiplication defined by certain appropriate rules (see 11). With addition and multiplication so defined, the complex numbers form a field. It should be noted that $i \times i = -1$.

Now, $\alpha = a + ib$ may be considered. Because $a^2 + b^2$ is a positive real number, there exists a unique positive real number r such that $r^2 = a^2 + b^2$, r being called the amplitude of α. If a, b are not both zero, there is an angle, symbolized by the Greek letter theta, θ, uniquely defined to within a multiple of 2π (Greek letter pi, the ratio of the circumference to the diameter of a circle), such that $a/r = \cos\theta$, $b/r = \sin\theta$, and hence α can be written in terms of r and θ (see 12). Let the Greek letter β stand for any other complex number (see 12). Following the rule for multiplying complex numbers, and again using elementary trigonometry, the product of two complex numbers is given in terms of the product of the amplitudes and the sum of the two angles (see 13).

$$(10) \qquad ax = 1, \qquad a \neq 0$$

$$(11) \qquad \begin{cases} (a+ib)+(c+id)=(a+c)+i(b+d) \\ (a+ib)\times(c+id)=(ac-bd)+i(bc+ad) \end{cases}$$

$$(12) \qquad \alpha = r(\cos\theta + i\sin\theta), \qquad \beta = \rho(\cos\phi + i\sin\phi)$$

$$(13) \qquad \alpha\beta = r\rho[\cos(\theta+\phi) + i\sin(\theta+\phi)]$$

It is possible to consider the problem of finding a complex number z such that $z^n = \alpha$. From the formula for the product, the amplitude s must satisfy $s^n = r$, and the angle ψ must satisfy $n\psi = \theta$ to within a multiple of 2π; that is, $n\psi - \theta = 2t\pi$ for some integer t. Because r is a positive real number, there must be a unique positive real number s satisfying $s^n = r$; ψ must have the form $\psi = (\theta/n) + (2t\pi/n)$, and there are n possible values of ψ (to within a multiple of 2π) satisfying this condition, corresponding to $t = 0, 1, 2, \cdots, n-1$, for example. Hence there are n possible numbers z such that $z^n = \alpha$.

Roots of unity

If z_1 is any one of these n numbers, the others can all be written $z_1\omega$, in which ω (omega) is an nth root of unity; that is, $\omega^n = 1$. Thus the problem of finding the nth roots of unity is basic in the theory of complex numbers.

It is clear that for any complex number α there are n solutions of $z^n = \alpha$, and this shows one of the advantages of complex numbers over the other kinds of numbers considered. In order to extract roots of integers, or rational numbers or real numbers, one notes that each of these number systems is a subsystem of the succeeding one, and that they are all subsystems of the complex numbers. To find z such that $z^n = \alpha$, when α is, say, a rational number, is to consider α as a complex number that just happens to belong to the subsystem of real numbers, and then to consider the problem for complex numbers. This process is usually referred to as extending the field of rational numbers to the complex field.

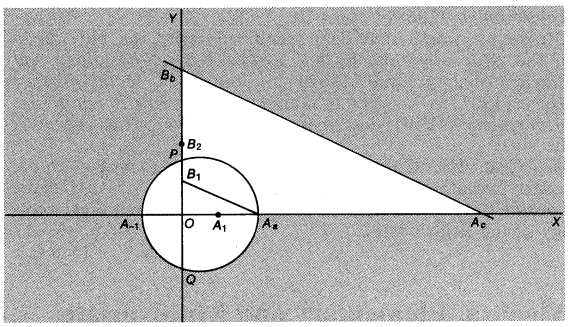


Figure 1: Geometric constructions necessary for the arithmetic operations of multiplication, division, and extraction of square roots (see text).

Geometric representation of numbers. Numbers may be represented geometrically, and the operations of algebra can be carried out by geometrical methods. The following representation of real numbers (which include the integers and rational numbers) was introduced by Descartes in the 17th century. On a straight line OX a point O is fixed, called the origin, and corresponding to each real number a the point A_a is taken, the distance of which from O is equal to a (if a is positive, $-a$ if it is negative), measuring distances to the right when a is positive, to the left when it is negative. One of the postulates of geometry is that there is a one-to-one correspondence between the real numbers and the points of the line established in this way. For convenience another line OY is considered, through O perpendicular to OX, and on it the point B_b is marked corresponding to the real number b, as before. The operations of addition, subtraction, multiplication, and division can then be obtained by simple geometrical constructions. The accompanying diagram (Figure 1) illustrates the geometric constructions necessary for multiplication, division, and the extraction of square roots.

Constructions involving ruler and compass

1. To add a and b the points A_a, A_b are marked. A segment A_aA_d, equal in length and direction to OA_b, is laid off on OX. Then $d = a + b$. By constructing the segment OA_{-b} equal in length to OA_b but in the opposite direction, $-b$ is found. In this way $a - b$ is constructed and geometrical representations of addition and subtraction are obtained.

2. To multiply a and b, A_a is marked on OX and OB_b is marked on OY. B_1 is marked on OY distant one unit from O. A_aB_1 is joined and, through B_b, B_bA_c is drawn parallel to B_1A_a. Then, by the geometry of similar triangles it may be shown that $c = ab$. Division by a nonzero number is obtained by a similar process. To construct c/b, when $b \neq 0$, B_b is joined to A_c, and the line through B_1 parallel to it. This meets OX in A_a, where $a = b/c$.

3. Let a be any positive real number, and on $A_{-1}A_a$ as diameter construct a circle. This meets OY in P, Q, and Euclidean geometry shows that OP and OQ are of equal length but opposite sign, and $OP \times OQ = OA_{-1} \times OA_a$. P represents a positive real number r on OY such that $r^2 = a$.

The geometrical operations described in 1, 2, 3 can all be carried out using only ruler and compass. It can be shown, on the other hand, that the operation of con-

structing the nth root of a, when n is not a power of 2, cannot in general be carried out solely by the use of ruler and compass.

The Argand diagram

Turning next to the representation of complex numbers in the Argand diagram (Figure 2) a number $\alpha = a + ib$ is represented by the point P constructed by constructing A_a on OX as above, B_b on OY similarly. P is then the intersection of the line A_aP through A_a parallel to OY, and the line B_bP through B_b parallel to OX. To construct $\alpha + \beta$, in which α is represented by P, β by Q, the parallelogram $OPRQ$ is completed. R represents $\alpha + \beta$ (Figure 2). If α has amplitude r and angle θ, r is the length of OP, and θ is the angle XOP. If Q represents β with amplitude ρ and angle ϕ, then ρ, ϕ can be interpreted in the same way. To construct the point representing $\alpha\beta$, the procedure adopted in Figure 1 is used to construct the product $r\rho$, and the angle $(\theta + \phi)$ can also be constructed by ruler and compass. It follows that the operations of addition, subtraction, multiplication, and division of complex numbers can be carried out geometrically by means of ruler and compass.

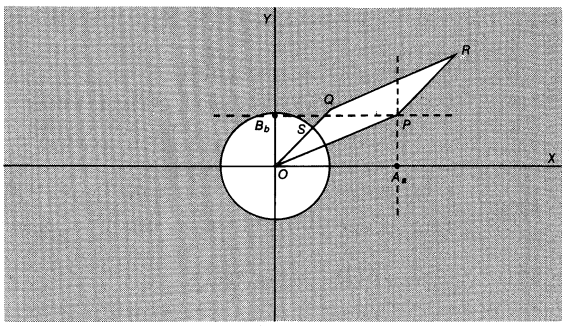


Figure 2: The representation of complex numbers and construction of the representation of the sum of two complex numbers (see text).

To construct a number z such that $z^n = \alpha$ it is necessary to find z with amplitude s and angle ψ such that $s^n = r$, $nx = \theta$ (modulo 2π). It is not possible to construct s by ruler and compass if n is not a power of 2. Nor can an angle be divided into n equal parts using ruler and compass if n is not a power of 2.

The consideration of problems that can be solved by the elementary operations of algebra and the extraction of roots, and by ruler and compass is of great importance in solving equations (see below *Solution of equations*).

POLYNOMIALS AND RATIONAL FUNCTIONS

Operations with coefficients and variables

Polynomials. The discussion of polynomials (defined below) begins with the coefficients, which, in elementary algebra, may be natural numbers, or real numbers, or complex numbers. Except when considering some special problems, the reasoning is the same in all three cases, and it is sufficient to speak of the coefficient field (or ring) without being more explicit.

Once the coefficient field is chosen, a new object is introduced, usually denoted by x, and called a variable or an indeterminate, which is not in the coefficient field. Laws are then specified by which to define addition and multiplication in the enlarged set consisting of the elements of the coefficient field and x. The powers of x, x^0, x^1, x^2, x^3, $\cdots$, are generated, together with products ax^n, in which a is in the coefficient field. Multiplication of two such products ax^n and bx^m is defined by $ax^n \times bx^m = abx^{n+m}$, and the sum of ax^n and bx^n is defined as $(a + b)x^n$; x^0 can be replaced by 1, and ax^0 by a. Moreover, $0x^n = 0$ for all n.

Sums of terms ax^n and bx^m, when $n \neq m$, are defined by introducing compound expressions such as $ax^n + bx^m$. It is finally postulated that multiplication of such expressions is distributive over addition, so that the product of x^n and such an expression is a similar expression with each term multiplied by x^n (see 14).

In this way a system is built up, any element of which can be written uniquely as the sum of terms of the form a_ix^{n-i}, for $i = 0, 1, \cdots, n$ (see 15), in which the coefficients a_i are in the coefficient field, and each expression only involves a finite number of terms a_rx^{n-r}. Here, and in what follows, $f(x)$ is an abbreviation for the expression that follows. If $a_0 \neq 0$, $f(x)$ is said to be a polynomial in x of degree n and a_rx^{n-r} is the term of degree $n - r$ in $f(x)$. If $g(x)$ is another polynomial of degree m (see 16), $f(x) + g(x)$ is obtained by adding the terms of like degree, and $f(x) \times g(x)$ is the polynomial of degree $n + m$ obtained by multiplication. With these laws of addition and multiplication, the polynomials in x form a ring. The ring has unity: the polynomial of degree zero written 1. Further, if neither $f(x)$ nor $g(x)$ is zero, their product $f(x) \times g(x)$ is not zero, because its leading term is $a_0b_0x^{n+m}$, and a_0b_0 is not zero unless one of a_0, b_0 is zero. Hence the ring of polynomials in x is an integral domain. It is easily seen, however, that it is not a field; for example, there is no polynomial $p(x)$ such that $p(x)(x + 1) = 1$.

Degree of polynomial expressions

It is now a straightforward matter to define polynomials in more than one variable. To the polynomials in x a new variable y is adjoined. Products are defined by $ax^py^q \times bx^ry^s = abx^{p+r}y^{q+s}$, $ax^py^q + bx^py^q$. is identified with $(a + b)x^py^q$, and x^0 and y^0 are identified with unity in the coefficient field. In this way the integral domain of polynomials in two variables is built up; a general element is written as a sum of these products (see 17). In a similar way the integral domain of polynomials in r variables may be constructed.

Elementary algebra is rich in identities concerning polynomials. Perhaps the most important is the binomial theorem, which expresses the sum of two terms raised to a power as a sum of products (see 18).

The ring of polynomials

Rational functions. The ring of polynomials in r variables is an integral domain, but not a field. It is possible, however, to embed the ring of polynomials in r variables in a field, the field of rational functions of r variables, just as the integers are embedded in the field of rational numbers. The case of one variable illustrates the general procedure. A rational function is one of the form $f(x)/g(x)$, in which $f(x)$ and $g(x)$ are polynomials; the following discussion makes this idea precise and rigorous. All pairs of polynomials $[f(x), g(x)]$ in which $g(x) \neq 0$, are considered and two pairs $[f(x), g(x)]$, $[F(x), G(x)]$ are defined to be equivalent if $f(x)G(x) - g(x)F(x) = 0$. This relation has all the properties of a true equivalence:

1. $[f(x), g(x)]$ is equivalent to itself (reflexive);
2. if $[f(x), g(x)]$ is equivalent to $[F(x), G(x)]$, then $[F(x), G(x)]$ is equivalent to $[f(x), g(x)]$ (symmetric); and
3. if $[f(x), g(x)]$ is equivalent to $[F(x), G(x)]$, and $[F(x), G(x)]$ is equivalent to $[P(x), Q(x)]$, then $[f(x), g(x)]$ is equivalent to $[P(x), Q(x)]$ (transitive).

As an illustration, this last property is proved. If $f(x)G(x) - g(x)F(x) = 0$, and $F(x)Q(x) - G(x)P(x) = 0$, then $Q(x)f(x)G(x) = Q(x)g(x)F(x) = g(x)G(x)P(x)$ and hence $G(x)[f(x)Q(x) - g(x)P(x)] = 0$. Because the ring of polynomials is an integral domain, and $G(x) \neq 0$, by hypothesis, $f(x)Q(x) - g(x)P(x) = 0$, which is the condition that $[f(x), g(x)]$ is equivalent to $[P(x), Q(x)]$.

$$(14) \qquad x^n(ax^p + bx^q) = ax^{p+n} + bx^{q+n}$$

$$(15) \qquad f(x) \equiv a_0x^n + a_1x^{n-1} + \cdots + a_n$$

$$(16) \qquad g(x) \equiv b_0x^m + b_1x^{m-1} + \cdots + b_m, \qquad b_0 \neq 0$$

$$(17) \qquad a_{mn}x^ny^m + a_{m-1,n}x^ny^{m-1} + \cdots + a_{rs}x^sy^r + \cdots + a_{00}$$

$$(18) \qquad \begin{cases} (a+bx)^n = a^n + c_1a^{n-1}bx + \cdots + \\ \qquad\qquad + c_ra^{n-r}b^rx^r + \cdots + c_nb^nx^n \\ c_r = \dfrac{1}{r!}n(n-1)\cdots(n-r+1), \qquad c_r = \dbinom{n}{r} \end{cases}$$

Addition and multiplication of pairs are defined by the appropriate formulas (see 19), which arise from the fact that the pair $[f(x), g(x)]$ is to represent the rational function $f(x)/g(x)$. By methods similar to that used to prove the transitivity of the equivalence relation, it can be shown that if $[f(x), g(x)]$ or $[p(x), q(x)]$ is replaced by an equivalent pair, their sum, or product, is replaced by an equivalent pair. Now consider a new set Σ of elements, one representing each class of equivalent pairs. Addition and multiplication can be defined in Σ and it can be verified that with those operations Σ is an integral domain. The unity of Σ is the class of pairs containing $[1, 1]$: and if $[f(x), g(x)]$ represents a class α that is not zero, $f(x)$ is not zero, and there is a class β represented by $[g(x), f(x)]$. Then $\alpha\beta$ is equal to the unity of Σ, and Σ is a field.

The set Σ contains a subring consisting of elements represented by pairs $[f(x), 1]$. If this class of pairs is identified with $f(x)$, it is seen that addition and multiplication in this subring are consistent with addition and multiplication in the ring of polynomials. Therefore Σ is regarded as an extension of the ring of polynomials. In this extension a class of pairs is represented by $f(x)/g(x)$ in which $[f(x), g(x)]$ is any pair representing the class of equivalent pairs. Thus the ring of polynomials in x has been extended to the field of rational functions of x. By a similar process the ring of polynomials in r variables can be extended to the field of rational functions of r variables.

Elementary properties of polynomials and rational functions. The remainder of this section deals with elementary properties of polynomials and rational functions.

Division of polynomials

The division algorithm (procedure) for polynomials in one variable is discussed first. If $f(x)$ and $g(x)$ are two polynomials of degree n and m, respectively (see 20), and if $m \leqslant n$, then $f(x) - (a_0/b_0)x^{n-m}g(x)$ is a polynomial $f_1(x)$ of degree n_1 ($n_1 < n$). If $n_1 \geqslant m$, in the same way, a number k in the coefficient field may be found such that a similar expression can be formed in terms of $f_1(x)$ and $g(x)$ (see 21). If $n_2 \geqslant m$, the process is repeated, and eventually $f(x)$ can be written as the sum of $a(x)g(x)$ and $b(x)$ (see 22), in which $a(x)$, $b(x)$ are polynomials [with coefficients in the coefficient field of $f(x)$ and $g(x)$] and the degree of $b(x)$ is less than m. This representation is unique, for if another such expression is formed and subtracted from the first (see 23), the right-hand side will be of degree less than m. Because $g(x)$ is of degree m, the left-hand side is of degree m at least unless $a(x) = a_1(x)$.

The equality therefore shows that $a(x) = a_1(x)$, $b(x) = b_1(x)$, and that the representation of $f(x)$ is unique. For use below (see *Solution of equations*) note that if $g(x) = x - \alpha$, $b(x)$ is a polynomial of degree zero equal to $f(\alpha) = a_0\alpha^n + a_1\alpha^{n-1} + \cdots + a_n$ (the remainder theorem).

Returning to the case in which $g(x)$ is a polynomial of degree m, suppose that $b(x) \neq 0$. The same process is applied to the polynomials $g(x)$, $b(x)$, leading to a similar formula expressing $g(x)$ as the sum of $a_1(x)b(x)$ and $b_1(x)$ (see 24), in which $a_1(x)$, $b_1(x)$ are polynomials the coefficients of which are in the coefficient field of $f(x)$, $g(x)$; and the degree of $b_1(x)$ is less than that of $b(x)$. If $b_1(x)$ is not zero, the process is repeated, until a stage is reached in which the remainder is zero. A set of equations of this form is obtained (see 25) ending when $b_{r+1}(x) = 0$.

From the first equation $b(x) = f(x) - a(x)g(x)$; substituting in the second, an expression $b_1(x) = p(x)f(x) + q(x)g(x)$ is obtained, and by successive substitution $b_r(x) = p_r(x)f(x) + q_r(x)g(x)$, in which $p_r(x)$, $q_r(x)$ are polynomials.

Next, starting from $f(x) = a(x)g(x) + b(x)$, one obtains by substitution two expressions for $f(x)$ involving the above quantities and the polynomials $c_j(x)$ and $d_j(x)$ (see 26).

Because $b_{r+1}(x)$ is zero, $f(x) = c_r(x)b_r(x)$. Starting from the second equation of (25), the same procedure gives $g(x) = e_r(x)b_r(x)$, in which $e_r(x)$ is a polynomial.

If $r(x)$ is a polynomial that is a factor of $f(x)$ and $g(x)$, so that $f(x) = l(x)r(x)$, $g(x) = m(x)r(x)$, then $b_r(x) = [p_r(x)l(x) + q_r(x)m(x)]r(x)$; that is, $r(x)$ is a factor of $b_r(x)$. Conversely, if $r(x)$ is a factor of $b_r(x)$, so that $b_r(x) = r(x)s(x)$, then $f(x) = c_r(x)r(x)s(x)$, $g(x) = e_r(x)r(x)s(x)$, so that $r(x)$ is a factor of $f(x)$ and $g(x)$. For this reason $b_r(x)$ is called the highest common factor of $f(x)$ and $g(x)$.

A polynomial $f(x)$ with coefficients in a field K is said to be reducible if it can be written in the form $f(x) = g(x)h(x)$, in which $g(x)$ and $h(x)$ are polynomials of degree greater than zero; otherwise it is irreducible. If $f(x)$ is reducible it can be written $g(x)h(x)$; if $g(x)$ or $h(x)$ is reducible it can be written, say, as $h(x) = l(x)m(x)$. In this way after a finite number of steps, one obtains $f(x) = p_1(x)p_2(x)\cdots p_r(x)$, in which each $p_i(x)$ is irreducible.

$$(19)\quad \begin{cases} [f(x), g(x)] + [p(x), q(x)] \\ \quad = [f(x)q(x) + p(x)g(x), g(x)q(x)] \\ [f(x), g(x)] \times [p(x), q(x)] = [f(x)p(x), g(x)q(x)] \end{cases}$$

$$(20)\quad \begin{cases} f(x) \equiv a_0x^n + a_1x^{n-1} + \cdots + a_n, & a_0 \neq 0 \\ g(x) \equiv b_0x^m + b_1x^{m-1} + \cdots + b_m, & b_0 \neq 0 \end{cases}$$

$$(21)\quad f_2(x) = f_1(x) - kx^{n_1-m}g(x)$$

$$(22)\quad f(x) = a(x)g(x) + b(x)$$

$$(23)\quad \begin{cases} f(x) = a_1(x)g(x) + b_1(x) \\ [a(x) - a_1(x)]g(x) = b_1(x) - b(x) \end{cases}$$

$$(24)\quad g(x) = a_1(x)b(x) + b_1(x)$$

$$(25)\quad \begin{cases} f(x) = a(x)g(x) + b(x) \\ g(x) = a_1(x)b(x) + b_1(x) \\ b(x) = a_2(x)b_1(x) + b_2(x) \\ \quad\vdots \qquad\qquad \vdots \\ b_{r-2}(x) = a_r(x)\,b_{r-1}(x) + b_r(x) \\ b_{r-1}(x) = a_{r+1}(x)b_r(x) \end{cases}$$

$$(26)\quad \begin{cases} f(x) = [a(x)a_1(x) + 1]b(x) + a(x)b_1(x), \\ f(x) = c_j(x)b_j(x) + d_j(x)b_{j+1}(x), \text{ in which } c_j(x) \text{ and} \\ d_j(x) \text{ are polynomials.} \end{cases}$$

The unique factorization theorem

This representation of $f(x)$ as a product of irreducible factors is unique, for suppose that $f(x) = q_1(x)q_2(x)\cdots q_s(x)$ is another such representation; $p_1(x)$ is a factor of $q_1(x)q_2(x)\cdots q_s(x)$, and so must be a factor of $q_i(x)$ for some i. This is proved by supposing $p_1(x)$ and $q_1(x)$ are not factors of one another; then their highest common factor is in the coefficient field, and from the above there exist polynomials $a(x)$, $b(x)$ such that $a(x)p_1(x) + b(x)q_1(x) = 1$; $b(x)q_1(x)q_2(x)\cdots q_s(x) = b(x)p_1(x)p_2(x)\cdots p_r(x) = [1 - a(x)p_1(x)]\,q_2(x)\cdots q_s(x)$.

Hence $p_1(x)\,[b(x)p_2(x)\cdots p_r(x) + a(x)q_2(x)\cdots q_s(x)] = q_2(x)\cdots q_s(x)$. Thus $p_1(x)$ is a factor of $q_2(x)\cdots q_s(x)$. The same argument can be repeated to show that if $p_1(x)$ is not a factor of $q_2(x)$, it must be a factor of $q_3(x)\cdots q_s(x)$, and so on, so that $p_1(x)$ must be a factor of $q_i(x)$ for some i. Since $p_1(x)$ and $q_i(x)$ are both irreducible, they can differ only by multiplication by an element of the coefficient field. In this way a pairing is obtained between factors $p_i(x)$ and $q_j(x)$, and the unique factorization theorem is proved.

It is easy to extend the theorem to polynomials in several variables: it is sufficient to demonstrate how it may be done for polynomials in two variables x and y. If $f(x, y)$ be a polynomial in x, y, it may be regarded as a polynomial in y, with coefficients in the field of rational functions of x. Because the unique factorization theorem

for one variable is proved by purely rational processes—that is, using only operations of addition and subtraction, and multiplication and division in the coefficients—it follows that $f(x, y) = r_1(x, y)\, r_2(x, y) \cdots r_k(x, y)$, in which $r_i(x, y)$ is an irreducible polynomial in y with coefficients that are rational functions of x, the k polynomials being uniquely determined save for order and factors that are rational functions of x. Let $b_i(x)$ be a common denominator of the coefficients of y in $r_i(x, y)$, so that each coefficient may be written $a(x)/b_i(x)$, in which $a(x)$ is a polynomial, and let $a_i(x)$ be the highest common factor of the coefficients of y in $b_i(x)\, r_i(x, y)$. Then $r_i(x, y) = [a_i(x)/b_i(x)]\, g_i(x, y)$, in which $g_i(x, y)$ is a polynomial in x and y, uniquely determined save for a factor in the field of coefficients. Hence $b_1(x)b_2(x) \cdots b_k(x)f(x, y) = a_1(x) \cdots a_k(x)g_1(x, y) \cdots g_k(x, y)$. Applying the unique factorization for polynomials in one variable to the coefficients of y, $b_i(x) \cdots b_k(x)$ is seen to be a factor of $a_i(x) \cdots a_k(x)$, and so $f(x, y) = c_1(x) \cdots c_s(x)g_1(x, y) \cdots g_k(x, y)$, in which the $c_i(x)$ and $g_j(x, y)$ are irreducible and uniquely determined to within a factor in the coefficient field.

There are many applications of the unique factorization theorem in elementary algebra.

Symmetric polynomials. Polynomials in r variables, $x_1, \cdots, x_r$, which are of importance in many parts of mathematics, for instance in the theory of equations, are the symmetric polynomials. A polynomial $f(x_1, \cdots, x_r)$ is a symmetric polynomial in its arguments $x_1, \cdots, x_r$ if the coefficients are rational numbers and if its value is unchanged whenever the subscripts in the arguments are replaced by any permutation of $(1, \cdots, r)$. For example, (2, 1, 3) is a permutation of (1, 2, 3). Important examples include the elementary symmetric polynomials (see 27), and sums of terms, each of the form x_i^k, for $i = 1, 2, \cdots, r$ (see 28).

Operations on symmetrical polynomials

In general (except for a plus or minus sign) the elementary symmetric polynomial γ_k in r variables $x_1, \cdots x_r$ is the coefficient of t^k in $(t - x_1)(t - x_2) \cdots (t - x_r)$ $(k = 1, 2, \cdots)$. The principal result concerning symmetric polynomials is that any such polynomial $F(x_1, \cdots, x_r)$ can be obtained by applying the operations of addition and multiplication to the elementary symmetric polynomials $\gamma_1, \gamma_2, \cdots, \gamma_r$, so that F is a polynomial in $\gamma_1, \cdots, \gamma_r$. This may be proved by the process of induction on r. There is, indeed, nothing to prove if $r = 1$. To begin the induction, it may be assumed that the theorem is true if applied to symmetric polynomials in $x_1, \cdots, x_{r-1}$. If $\gamma'_1, \gamma'_2, \cdots, \gamma'_{r-1}$ denote the elementary symmetric polynomials in $x_1, \cdots, x_{r-1}$, then $\gamma_1, \cdots, \gamma_r$ can be expressed in terms of $\gamma'_1, \cdots, \gamma'_r$ and x_r (see 29). These equations can be solved for $\gamma'_1 \cdots, \gamma'_r$ (see 30). Writing $F(x_1, \cdots, x_r)$ as a polynomial in x_r (see 31), with the coefficients α_i being polynomials in $x_1, \cdots, x_{r-1}$, one finds that because $F(x_1, \cdots, x_r)$ is unaltered by any permutation of $x_1, \cdots, x_r$, in particular by any permutation leaving x_r fixed, α_i is a symmetric polynomial in $x_1, \cdots, x_{r-1}$, and can therefore be represented as a polynomial in $\gamma'_1, \cdots, \gamma'_{r-1}$. Replacing γ'_1 by $\gamma_1 - x_r$, γ'_2 by $\gamma_2 - x_r\gamma_1 + x_r^2$, etc., it is possible to write $F(x_1, \cdots, x_r)$ as a polynomial $G(\gamma_1, \cdots, \gamma_{r-1}, x_r)$ in $\gamma_1, \cdots, \gamma_{r-1}, x_r$. Now if $f(z)$ is a product of terms of the form $(z - x_i)$, for $i = 1, 2, \cdots, r$ (see 32), then $f(x_r) = 0$. Using this, powers of x_r in $G(\gamma_1, \cdots, \gamma_{r-1}, x_r)$ of degree greater than $r - 1$ can be replaced by expressions involving $\gamma_1, \cdots, \gamma_r$ and lower powers of x_r. Finally an equation is reached for $F(x_1, \cdots, x_r)$ as a sum of terms $\beta_{i-1}x_r^{r-i}$ (see 33), in which $\beta_0 \cdots, \beta_{r-1}$ are polynomials in $\gamma_1, \cdots, \gamma_r$—and hence symmetric polynomials in $(x_1, \cdots, x_r)$. Because $F(x_1, \cdots, x_r)$ is symmetric, r similar equations result (see 34). Now consider the polynomial $H(z)$, formed by subtracting $F(x_1, \cdots, x_r)$ from a similar polynomial in the variable z (see 35), with coefficients in the field of rational functions of $x_1, \cdots, x_r$. $H(z)$ has r zeros, namely $z = \alpha_1, \cdots, \alpha_r$. But (see below *Solution of equations*) a polynomial of degree $r - 1$ or less that has r zeros is identically zero. Hence putting $z = 0$, $F(x_1, \cdots, x_r)$ equals β_{r-1} (see 36), in which β_{r-1} is a polynomial in the elementary symmetric polynomials $\gamma_1, \cdots, \gamma_r$. So if any polynomial in $r - 1$ variables may be expressed as a polynomial in the $r - 1$ elementary symmetric polynomials, it follows that any polynomial in r variables may be so expressed. Because the theorem is true for one variable, it is true for any number of variables.

The symmetric polynomials play an important role in the theory of equations (see below). A simple calculation enables expressions to be given for the sums s_r of the rth powers of $(x_1, \cdots, x_r)$ in terms of $\gamma_1, \cdots, \gamma_r$; from these relations it can be shown that $\gamma_1, \cdots, \gamma_r$ are equal to polynomials in $s_1, \cdots, s_r$; hence any symmetric polynomial can be expressed as a polynomial in $s_1 \cdots, s_r$.

(27) The *elementary symmetric polynomials*:
$$\gamma_1 = x_1 + x_2 + \cdots + x_r,$$
$$\gamma_2 = x_1x_2 + x_1x_3 + \cdots + x_2x_3 + \cdots + x_{r-1}x_r,$$
$$\gamma_r = x_1x_2 \cdots x_r.$$

(28) The sums of powers: $\rho_k = x_1^k + x_2^k + \cdots + x_r^k$, $k = 1, 2, \cdots$.

(29) $$\gamma_1 = \gamma'_1 + x_r, \qquad \gamma_2 = \gamma'_2 + x_r\gamma'_1, \quad \cdots$$
$$\gamma_r = x_r\gamma'_{r-1}$$

(30) $$\gamma'_1 = \gamma_1 - x_r, \qquad \gamma'_2 = \gamma_2 - x_2\gamma_1 + x_r^2, \quad \cdots$$

(31) $$F(x_1, \cdots, x_r) = \alpha_0 x_r^m + \alpha_1 x_r^{m-1} + \cdots + \alpha_m$$

(32) $$f(z) = (z - x_1)(z - x_2) \cdots (z - x_r)$$
$$= z^r - \gamma_1 z^{r-1} + \gamma_2 z^{r-2} - \cdots \pm \gamma_r$$

(33) $$F(x_1, \cdots, x_r) = \beta_0 x_r^{r-1} + \beta_1 x_r^{r-2} + \cdots + \beta_{r-1}$$

(34) $$F(x_1, \cdots, x_r) = \beta_0 x_i^{r-1} + \beta_1 x_i^{r-2} + \cdots + \beta_{r-1}$$
$$i = 1, \cdots, r$$

(35) $$H(z) = \beta_0 z^{r-1} + \beta_1 z^{r-2} + \cdots + \beta_{r-1} - F(x_1, \cdots, x_r)$$

(36) $$F(x_1, \cdots, x_r) = \beta_{r-1}$$

(37) $$F_i(x_1, x_2, \cdots, x_n) = 0, \qquad i = 1, \cdots, r$$

(38) $$F_i(\alpha_1, \alpha_2, \cdots, \alpha_n) = 0, \qquad i = 1, \cdots, r$$

SOLUTION OF EQUATIONS

One of the principal problems of elementary algebra is the solution of a set of equations in which r polynomials $F_i(x_1, x_2, \cdots, x_n)$, $i = 1, 2, \cdots, r$, with coefficients in a field (rational numbers, real numbers, or complex numbers), are set equal to zero (see 37). By a solution of these equations is meant a set of numbers $(\alpha_1, \alpha_2, \cdots, \alpha_n)$ in the coefficient field, or in some extension of the coefficient field, that, when substituted for $x_1, x_2, \cdots, x_n$, causes the polynomials all to equal zero (see 38).

Nature of the problem

The problem falls into three parts. (1) Does there exist a solution of the equations, and how can a given set of equations be tested for solvability? This question can be refined by asking whether there exist solutions in the field of the coefficients, or in some prescribed extension of this field. A particular case of this arises when the coefficients are natural numbers, and it is desired to find solutions $(\alpha_1, \alpha_2, \cdots, \alpha_n)$ in which α_i is rational. Such problems really belong to the theory of numbers, though methods of algebraic geometry (see below) play an important role in these investigations. (2) The criteria for determining whether a set of equations has a solution do not, as a rule, give any clue to finding a solution when one is known to exist. One asks whether there exists some algebraic process, involving the elementary operations of addition, subtraction, multiplication, and divi-

$$(39)\qquad y_0^{\pi_i} F_i\left(\frac{y_1}{y_0}, \frac{y_2}{y_0}, \cdots, \frac{y_n}{y_0}\right) \equiv G_i(y_0, y_1, \cdots, y_n)$$

$$(40)\qquad G_i(ty_0, ty_1, \cdots, ty_n) = t^{\pi_i} G_i(y_0, y_1, \cdots, y_n)$$

$$(41)\qquad G_i(y_0, y_1, \cdots, y_n) = 0, \qquad i = 1, \cdots, r$$

$$(42)\qquad f(x) \equiv a_0x^n + a_1x^{n-1} + \cdots + a_n = 0$$

$$(43)\qquad f(x) = f_1(x)(x-\alpha) + f(\alpha)$$

$$(44)\qquad f_1(x) = (x-\beta)f_2(x), \qquad f(x) = (x-\alpha)(x-\beta)f_2(x)$$

$$(45)\qquad f(x) = a_0(x-\alpha_1)(x-\alpha_2)\cdots(x-\alpha_n)$$

$$(46)\qquad \alpha_0(\beta-\alpha_1)(\beta-\alpha_2)\cdots(\beta-\alpha_n) = 0$$

$$(47)\qquad \begin{cases} a_0x^n + a_1x^{n-1} + \cdots + a_n \\ \quad = a_0(x-\alpha_1)(x-\alpha_2)\cdots(x-\alpha_n) \\ \quad = a_0x^n - a_0(\alpha_1+\alpha_2+\cdots+\alpha_n)x^{n-1} + \cdots \\ \quad = a_0x^n - a_0\delta_1x^{n-1} + a_0\delta_2x^{n-2} - \cdots \end{cases}$$

$$(48)\qquad x^2 + a_1x + a_2 = (x+\tfrac{1}{2}a_1)^2 + a_2 - \tfrac{1}{4}a_1^2 = (x+\tfrac{1}{2}a_1+\alpha)(x+\tfrac{1}{2}a_1-\alpha)$$

sion, together with the extraction of roots, which, when applied to the coefficients of the equation, will lead to a solution. More precisely, it is of considerable interest to ask whether the solution can be obtained by elementary operations only (the solutions are then said to be rational) or if they can be found by elementary operations together with the extraction of square roots so that the solutions can be constructed geometrically by ruler and compass only (see above). (3) Whether or not it is possible to find the solution of a set of equations explicitly, it may be possible to develop a theory of the aggregate of solutions of a set of equations. This leads to algebraic geometry. In elementary geometry of the plane coordinates (x, y) are introduced consisting of pairs of real numbers x and y, and a one-to-one correspondence is established between all pairs (x, y) and all points of the plane. This notion can be extended in several ways: First the plane may be replaced by the real affine space of n dimensions, in which there is a one-to-one correspondence between the points of the space and all possible sets $(\alpha_1, \alpha_2, \cdots, \alpha_n)$ of real numbers. Again, this notion may be extended to obtain complex affine space, in which the coordinates are sets of n complex numbers. Then each real (complex) solution of the equation defines a unique point in real (complex) affine space of n dimensions. The aggregate of the solutions defines a locus in the affine space, and the geometrical properties of this locus provide important information about the solution of the equations.

Solution as a unique point in affine space

It is familiar that, in the elementary geometry of the plane, simple theorems such as "two (distinct) lines have one point in common" have to be qualified by adding "or else are parallel." This difficulty is overcome by adding to the affine plane a "line at infinity" to make the plane into a projective plane, the points being given by sets of numbers (α, β, γ) not all zero, two sets of numbers which are proportional determining the same point so that (α, β, γ) and $(\lambda\alpha, \lambda\beta, \lambda\gamma)$ represent the same point. The points of the affine plane are identified with the points of the projective plane for which α is not zero. The affine coordinates of the point are $(\beta/\alpha, \gamma/\alpha)$. The same process can be used to extend the real (complex) affine space of n dimensions to a projective space of n dimensions. Now let $(\alpha_1, \alpha_2, \cdots, \alpha_n)$ be any solution of the equations being considered (see 38). This is represented in projective space of n dimensions by the point $(\beta_0, \beta_1, \cdots, \beta_n)$, in which $\beta_0\alpha_i = \beta_i$, $i = 1, \cdots, r$. In $F_i(x_1, x_2, \cdots, x_n)$ replace x_i by y_i/y_0 and multiply by the lowest power of y_0 that yields a polynomial G_i (see 39). This polynomial is homogeneous (see 40). Now consider the set of homogeneous equations (see 41) in which the G_i are set equal to zero. If $(\beta_0, \beta_1, \cdots, \beta_n)$ is a solution, so is $(\lambda\beta_0, \lambda\beta_1, \cdots, \lambda\beta_n)$ for any λ. If not all β_i are zero, the solution $(\lambda\beta_0, \lambda\beta_1, \cdots, \lambda\beta_n)$ defines a unique point in the projective space of n dimensions. If $\beta_0 \neq 0$, then $(\beta_1/\beta_0, \beta_2/\beta_0, \cdots, \beta_n/\beta_0)$ is a solution of the original equations (see 37). Thus the homogeneous equations (see 41) define a locus in projective space of n dimensions; and the points of this locus not in $y_0 = 0$ can be identified with the points of the affine space from which the projective space was constructed, which corresponds to the solutions of the original equations (see 37).

Just as the use of projective space allows certain difficulties arising in the geometry of affine space to be eliminated, so the restriction to homogeneous equations simplifies some difficulties in algebra, which arise when the solutions of nonhomogeneous equations are considered.

Existence of solutions. As a preliminary to the algebraic study of the equation in r variables, it is necessary to consider the solution of an n-degree polynomial equation in one variable (see 42), in some detail. The fundamental theorem of algebra states that if $n > 0$ and the coefficients are rational, or real, or complex numbers, there exists at least one complex number α such that $f(\alpha) = 0$. The number α is called a root of $f(x)$, or a solution of $f(x) = 0$. No elementary algebraic proof of this theorem exists, and the result is not proved here. There is an extension of the theorem which states that if the coefficients of $f(x)$ are in any field K, there exists an extension of K in which $f(x)$ has a root. In what follows, however, the coefficients of $f(x)$ are in the complex field: $f(x)$ then has a root in this field.

The fundamental theorem of algebra

The remainder theorem proved above (see *Elementary properties of polynomials and rational functions*) can be applied (see 43) to express $f(x)$ as the sum of $f(\alpha)$ and the product $f_1(x)(x-\alpha)$, in which $f_1(x)$ is a polynomial of degree $n-1$. If α is a root of $f(x)$, this implies that $(x-\alpha)$ is a factor of $f(x)$. If $(n-1) > 0$, $f_1(x)$ also has a root β, and the process can be continued to express $f(x)$ as a product (see 44 and 45) of terms of the form $(x-\alpha_i)$, for $i = 1, 2, \cdots, n$, in which $\alpha_1, \alpha_2, \cdots, \alpha_n$ are complex numbers. Because $f(\alpha_i) = 0$, each α_i is a root of $f(x)$. If β is any root of $f(x)$, then this expression equals zero (see 46), and because the complex numbers form a field, one term $(\beta - \alpha_i)$ must be zero. Hence $\alpha_1, \alpha_2, \cdots, \alpha_n$ are the only roots of $f(x)$. So now the polynomial can be written in terms of δ_i (see 47), the result of replacing $x_1, x_2, \cdots, x_n$ by $\alpha_1, \alpha_2, \cdots, \alpha_n$ in the elementary symmetric polynomial γ_i (see 27). A function δ_i is usually called an (elementary) symmetric function of the roots of $f(x)$, and a considerable body of algebra deals with properties of the symmetric functions of the roots of an equation. For example, note that $a_0\delta_i = -a_i$ if i is odd, and that $a_0\delta_i = a_i$ if i is even. Thus the symmetric functions of the roots lie in the coefficient field of $f(x)$, though the individual roots may not.

Extraction of the roots of a polynomial. Although the fundamental theorem of algebra enables one to show that a polynomial $f(x)$ of degree n has n roots, it gives no clue to finding the roots. In algebra, it is desired to give rules for applying a series of elementary operations, together with the extraction of rth roots (the standard operations of algebra) on the coefficients of $f(x)$ to find the roots. In considering this problem, note that the roots of $f(x)$ are the same as the roots of $bf(x)$ in which b is any nonzero number of the coefficient field; b can be chosen so that $ba_0 = 1$; this means that without loss of generality a_0 may be set equal to one in $f(x)$; by doing this the notation is greatly simplified.

Finding the roots

(I) $n = 1$. The equation is $x + a_1 = 0$, which has the unique solution $-a_1$ in the coefficient field.

(II) $n = 2$. Any quadratic expression can be put in the form $x^2 + a_1x + a_2$ and then can be factored, giving the solutions $-\frac{1}{2}a_1 + \alpha$, $-\frac{1}{2}a_1 - \alpha$ (see 48), in which α, $-\alpha$ are the square roots of $\frac{1}{4}a_1^2 - a_2$. The root α can be constructed by use of ruler and compass, and the roots

$-\frac{1}{2}a_1 + \alpha$, $-\frac{1}{2}a_1 - \alpha$ are obtained by the standard operations of algebra.

(III) $n = 3$. The variable in $f(x)$ is changed by writing $y = x + \frac{1}{3}a_1$ giving a cubic equation with no quadratic term (see 49), in which the constants p and q are obtained from a_1, a_2, a_3 by elementary operations. If a root β of $g(y)$ can be found, $\beta - \frac{1}{3}a_1$ is a root of $f(x)$. If $p = 0$, the roots of $g(y)$ are determined by extracting a cube root, and the solution of $f(x) = 0$ is complete.

Consider now $p \neq 0$ and make another change of variable determined by $y = z - p/3z$. Then express $g(y)$ in terms of z (see 50), and it is clear that if γ is any root of $h(z)$, $\gamma - p/3\gamma$ is a root of $g(y)$, and so a root of $f(x)$ is obtained. Defining two new quantities ξ and η and using the properties of homogeneous equations (see 40), one can factor $h(z)$ into a product (see 51), with terms of the product constructed by standard processes. By a standard process (but not by ruler and compass) a number α can be constructed such that $\alpha^3 = \xi$. If 1, ω_1, ω_2 are the cube roots of unity, the other roots of $z^3 - \xi$, one of the factors of $h(z)$, are $\alpha\omega_1$, $\alpha\omega_2$. It can be shown (see 52) that β is a cube root of η. The other cube roots are obtained, similarly, using $\alpha\omega_1$, $\alpha\omega_2$; they are in fact $\beta\omega_2$, $\beta\omega_1$, respectively.

If $\gamma_1 = \alpha - p/3\alpha$, γ_1 is a root of $g(y)$. Replacing α in this by $\alpha\omega_1$, $\alpha\omega_2$ or β, $\beta\omega_2$, $\beta\omega_1$ the other roots of $g(y)$ are obtained. On account of the relation $\beta = -p/3\alpha$, however, the pairs (α, β), $(\alpha\omega_1, \beta\omega_2)$, $(\alpha\omega_2, \beta\omega_1)$ give the same roots of $g(y)$. Thus three roots of $g(y)$ are obtained by the standard processes of algebra, and hence three roots of $f(x)$.

(IV) $n = 4$. It can also be proved that the fourth degree equation (see 53) can be solved by the standard processes of algebra. It is instructive to see how one is led to the solution by simple considerations of geometry in the complex affine plane. In this plane define two conics by equations in x and y (see 54 and 55). A conic in the real plane is formed by the intersection of the plane with a cone. A circle, ellipse, hyperbola, parabola, or pair of intersecting straight lines is a conic. The concept is generalized to the complex plane. If α is any root of the fourth-degree equation (see 53), (α, α^2) is a point common to the conics. Conversely, if (α, β) is any point common to the two conics, $\beta = \alpha^2$ from the second conic, α is a root of the fourth-degree equation (see 53). So it is necessary to find the four points common to the two conics. Define a function $F(x, y, \lambda)$, in which λ is any complex number, from these two conics (see 56). $F(x, y, \lambda) = 0$ is the equation of a conic in the plane. It is clear that if λ, μ are different complex numbers, the points common to $F(x, y, \lambda) = 0$ and $F(x, y, \mu) = 0$ are the points common to the conics (see 54 and 55). Now it is a theorem of geometry that $F(x, y, \lambda) = 0$ becomes a pair of lines if λ is a root of a certain cubic polynomial (see 57). By (III), this polynomial in λ has three roots λ_1, λ_2, λ_3, which can be found by the standard processes. Then, by solving quadratic equations, as in (II), $F(x, y, \lambda)$ can be written as a product of two terms, each linear in x and y (see 58). To find the four points common to the conics it is necessary to find the points common to $F(x, y, \lambda_1) = 0$ and $F(x, y, \lambda_2) = 0$. To find these, four pairs of linear equations (see 59 and 60) must be solved.

Because the four points so obtained satisfy the original equation and the second conic (see 37 and 55), the solution of the fourth-degree equation (see 53) is deduced, and can therefore be obtained by the standard processes of algebra.

Equations of degrees higher than four

In the 18th century, much effort went into attempts to find methods of finding the solution of an equation of degree higher than four by the standard processes of algebra. Early in the 19th century the Norwegian mathematician Niels Henrik Abel proved that the general equation of degree five could not be solved this way. Later, the French mathematician Évariste Galois extended Abel's result to equations of degree greater than five and developed a general theory of equations that threw much light on the problem (see ALGEBRAIC STRUCTURES).

Common roots of two polynomials. Now let $f(x)$ and $g(x)$ be polynomials of degree n and m, respectively, and let $f(x) = 0$ and $g(x) = 0$ (see 61 and 62). One may ask if these equations have any common solution, and if so how it is found. It was shown above how to construct the highest common factor of $f(x)$, $g(x)$. This is a polynomial $H(f, g)$, such that there exist polynomials $p(x)$, $q(x)$, $r(x)$, $s(x)$, in terms of which $f(x)$ and $g(x)$ can be written (see 63 and 64). From these equations it follows that any common root of $f(x)$ and $g(x)$ is a root of $H(f, g)$, and any root of $H(f, g)$ is a common root of $f(x)$, $g(x)$. Thus $f(x)$ and $g(x)$ have a common root if and only if the highest common factor is of degree greater than zero. The process described can easily be extended to find the common roots, if any, of any finite set of polynomials $f_i(x)$ $(i = 1, \cdots, r)$. The highest common factor $H(f_1, f_2)$ of $f_1(x)$, $f_2(x)$ is constructed; the highest common factor $H(f_1, f_2, f_3)$ of this and $f_3(x)$, and so on. The common roots of $f_i(x)$ are simply the roots of $H(f_1, f_2, \cdots, f_r)$.

Use of the highest common factor

If $f(x)$ and $g(x)$ have a common root, then let $\alpha_1, \cdots, \alpha_n$ be the roots of $f(x)$, $\beta_1, \cdots, \beta_m$ the roots of $g(x)$. Consider a product with terms of the form $(\alpha_i - \beta_j)$, for $i = 1, 2, \cdots, n$, and $j = 1, 2, \cdots, m$ (see 65). The first expression is a symmetric function of the roots of $f(x)$, and hence can be expressed in terms of the elementary symmetric functions of these roots, that is, in terms of $a_0, a_1 \cdots, a_n$. The expression is therefore a polynomial $R(f, g)$ in the coefficients $a_0, \cdots, a_n, b_0, \cdots, b_m$. If, for any given polynomials $f(x)$ and $g(x)$ in which

(49) $$f(x) = g(y) \equiv y^3 + py + q$$

(50) $$g(y) = \frac{z^6 + qz^3 - p^3/27}{z^3} = \frac{h(z)}{z^3}$$

(51) $$\begin{cases} h(z) = (z^3 - \xi)(z^3 - \eta) \\ \xi = -\frac{1}{2}q + \rho, \qquad \eta = -\frac{1}{2}q - \rho \end{cases}$$

(52) $$\begin{cases} \beta = \dfrac{-p}{3\alpha} \\ \beta^3 = \dfrac{-p^3}{27\xi} = \dfrac{-p^3}{27}\dfrac{1}{-\frac{1}{2}q + \rho} = \dfrac{-p^3}{27}\dfrac{-\frac{1}{2}q - \rho}{\frac{1}{4}q^2 - \rho^2} = -\frac{1}{2}q - \rho \end{cases}$$

(53) $$x^4 + a_1x^3 + a_2x^2 + a_3x + a_4 = 0$$

(54) $$y^2 + a_1xy + a_2y + a_3x + a_4 = 0$$

(55) $$y - x^2 = 0$$

(56) $$F(x, y, \lambda) \equiv \lambda(x^2 - y) + y^2 + a_1xy + a_2y + a_3x + a_4$$

(57) $$\lambda(\lambda - a_2)^2 + a_1a_3(\lambda - a_2) - 4a_4\lambda + a_3^2 + a_1^2a_4 = 0$$

(58) $$F(x, y, \lambda_i) = (a_ix + b_iy + c_i)(\alpha_ix + \beta_iy + \gamma_i) \quad i = 1, 2, 3$$

(59) $$a_1x + b_1y + c_1 = 0, \qquad a_2x + b_2y + c_2 = 0$$

(60) $$a_1x + b_1y + c_1 = 0, \qquad \alpha_2x + \beta_2y + \gamma_2 = 0$$

(61) $$f(x) \equiv a_0x^n + a_1x^{n-1} + \cdots + a_n = 0$$

(62) $$g(x) \equiv b_0x^m + b_1x^{m-1} + \cdots + b_m = 0$$

(63) $$H(f, g) = p(x)f(x) + q(x)g(x)$$

(64) $$f(x) = r(x)H(f, g), \qquad g(x) = s(x)H(f, g)$$

(65) $$\begin{aligned} &a_0^m b_0^n(\alpha_1 - \beta_1)(\alpha_1 - \beta_2)\cdots(\alpha_n - \beta_m) \\ &= a_0^m g(\alpha_1)g(\alpha_2)\cdots g(\alpha_n) \\ &= \pm b_0^n f(\beta_1)f(\beta_2)\cdots f(\beta_m) \end{aligned}$$

a_0 and b_0 are not zero, $R(f, g) = 0$, it follows that the equations have a common solution, and conversely if the equations have a common solution $R(f, g) = 0$.

In later developments, the condition of $a_0 \neq 0$, $b_0 \neq 0$ turns out to be an undesirable restriction. This restriction may be removed to some extent. Suppose that $a_0 \neq 0$, $b_0 = 0$. Because $R(f, g)$ can be written as a product of a_0^m and the n terms $g(\alpha_i)$ (see 66), in which $\alpha_1, \cdots, \alpha_n$ are roots of $f(x)$, $R(f, g) = 0$ implies $g(\alpha_i) = 0$ for some root α_i of $f(x)$. Hence this α_i is a common root of $f(x)$, $g(x)$. If $a_0 = 0 = b_0$, however, this argument breaks down. It can be verified that if a_0 and b_0 both vanish, $R(f, g) = 0$, whatever values are given to $a_1, \cdots, a_n, b_1, \cdots, b_m$, and nothing is learned about the roots of $f(x)$, $g(x)$.

To overcome this difficulty a new variable y is introduced in $f(x)$, $g(x)$ by replacing a_i by $A_i y^i$, b_j by $B_j y^j$; then three homogeneous polynomials $F(x, y)$, $G(x, y)$ and $R(f, g)$ are obtained (see 67). The polynomial $T(F, G)$ in $A_0, \cdots, A_n, B_0, \cdots, B_m$ is obtained by replacing $a_0, \cdots, a_n, b_0, \cdots, b_m$ by $A_0, \cdots, A_n, B_0, \cdots, B_m$ in $R(f, g)$.

If $F(x, y)$ and $G(x, y)$ both equal to zero (see 68) have a common solution (α, β) in which $\beta \neq 0$, α/β is a common root of $f(x)$, $g(x)$, and hence $T(F, G) = 0$. If the homogeneous equations have a common solution $(\alpha, 0)$ in which $\alpha = 0$, then A_0, B_0 are zero, and $T(F, G) = 0$. Conversely, if $T(F, G) = 0$ and A_0, B_0 are not both zero, $F(x, 1)$ and $G(y, 1)$ have a common root α, and $(\alpha, 1)$ is a common solution (see 68). If A_0, B_0 are zero the equations have the common solution $(1, 0)$. Hence $T(F, G) = 0$ is a necessary and sufficient condition for the homogeneous equations (68) to have a common solution.

The problem of solving equations in two variables (x, y) is now considered. If a simple equation $f(x, y) = 0$ is taken, y is given an arbitarary value η in the complex field, and the equation $f(x, \eta) = 0$ in one variable remains. If ξ is a root of $f(x, \eta)$, (ξ, η) is a solution for $f(x, y) = 0$.

Two simultaneous equations. Now consider two simultaneous nonhomogeneous polynomial equations (see 69 and 70) in which the coefficients a_i, b_j are now polynomials in y. $R(f, g)$ is then a polynomial in y; and it is clear that if (ξ, η) is a common solution of the two equations, then η satisfies $R(f, g) = 0$. If, on the other hand, η is a root of $R(f, g)$ and the polynomials a_0, b_0 do not both have η as a root, there exists a common root of $f(x, \eta)$, $g(x, \eta)$ and so a common solution of the two equations (see 69 and 70). If η is a common zero of a_0, b_0, however, the difficulty described above arises. To get over this difficulty, a new variable z is again introduced to make $f(x, y)$, $g(x, y)$ homogeneous, and $F(x, y, z)$ is defined as a sum of $N + 1$ terms of the form $A_i(y, z)x^{N-i}$, for $i = 0, 1, \cdots, N$, and $G(x, y, z)$ as a similar sum of terms of the form $B_j(y, z)x^{M-j}$, for $j = 0, 1, \cdots, M$ (see 71), where $A_i(y, z)$ is homogeneous of degree i and $\beta_j(y, z)$ homogeneous of degree j. Then ut is written for y, vt for z (see 72). Regarding these as equations in (x, t), $T(F, G)$ is constructed as before. It is homogeneous in (u, v). If A_0 and B_0 are both zero, the equation $F(x, y, z) = 0$, $G(x, y, z) = 0$ have a common solution $(1, 0, 0)$. If they are not both zero, $F(x, y, z) = 0$ and $G(x, y, z) = 0$ have a common solution (λ, μ, ν) if and only if (μ, ν) is a solution of $T(F, G) = 0$.

This again illustrates the advantages of studying homogeneous equations and the associated loci in projective space.

Equations in three variables

Linear systems. To conclude, there follows a brief consideration of the solution of a system of linear homogeneous equations in three variables x, y, z, with coefficients a_i, b_i, c_i, respectively, for $i = 1, 2, \cdots, r$ (see 73). If each a_i, b_i, c_i, is zero ($i = 1, \cdots, r$), the problem is trivial. Next, suppose, for instance, that a_1, b_1, c_1 are not all zero, and that a certain relation of the coefficients (see 74) holds. Then there is a nonzero number such that l_i equals the product of p_i and l_1 (see 75), and the problem reduces to finding the solution of a single such equation (see 76).

Next, suppose that $b_ic_j - b_jc_i$, $c_ia_j - c_ja_i$, $a_ib_j - a_jb_i$ are not all zero for a particular pair i, j, say $b_ic_j - b_jc_i \neq 0$. Then y can be eliminated yielding a simpler problem (see 77). This equation has a unique solution $[\rho(b_ic_j - b_jc_i), \rho(a_ib_j - a_jb_i)]$ in which $\rho \neq 0$, and substituting in the expressions for l_i, l_j, $l_i = 0 = l_j$ have the unique solution $(b_ic_j - b_jc_i, c_ia_j - c_ja_i, a_ib_j - a_jb_i)$. If equations being considered have any solution, it must be unique and this must be it. This solution, then, substituted into the original system, gives the condition for the existence of a solution (see 78). It must hold for all i, j, k. In the language of linear algebra the equations (73) have an infinity of solutions, a unique solution, or no solution according as the rank of the matrix of coefficients (see 79) is one, two, or three.

It should be noted that if the nonhomogeneous equations (see 80) had been considered, there would have been more possibilities. For instance, suppose that $a_ib_j - a_jb_i = 0$, $b_ic_j - b_jc_i \neq 0$, for given values of i, j. Then a system of equations (see 81) is reached that has no solution, and hence the nonhomogeneous equations have no solution. In this case, the geometrical explanation is that the lines $M_i = 0$, $M_j = 0$ are parallel, but not identical. To obtain an intersection, it is necessary to pass from the affine plane to the projective plane.

(66) $$R(f, g) = a_0^m g(\alpha_1) \cdots g(\alpha_n)$$

(67) $$\begin{cases} F(x, y) \equiv A_0x^n + A_1x^{n-1}y + \cdots + A_ny^n \\ G(x, y) \equiv B_0x^m + B_1x^{m-1}y + \cdots + B_my^m \\ R(f, g) = y^{mn}T(F, G) \end{cases}$$

(68) $$F(x, y) = 0, \qquad G(x, y) = 0$$

(69) $$f(x, y) \equiv a_0x^n + a_1x^{n-1} + \cdots + a_n = 0$$

(70) $$g(x, y) = b_0x^m + b_1x^{m-1} + \cdots + b_m = 0$$

(71) $$\begin{cases} F(x, y, z) \equiv A_0(y, z)x^N + A_1(y, z)x^{N-1} + \\ \quad + \cdots + A_N(y, z) = 0 \\ G(x, y, z) \equiv B_0(y, z)x^M + R_1(y, z)x^{M-1} + \\ \quad + \cdots + B_M(y, z) = 0 \end{cases}$$

(72) $$\begin{cases} F(x, y, z) = A_0(u, v)x^N + A_1(u, v)tx^{N-1} + \\ \quad + \cdots + A_N(u, v)t^N \\ G(x, y, z) \equiv B_0(u, v)x^M + B_1(u, v)tx^{M-1} + \\ \quad + \cdots + B_M(u, v)t^M \end{cases}$$

(73) $$l_i \equiv a_ix + b_iy + c_iz, \qquad i = 1, \cdots, r$$

(74) $$b_ic_j - b_jc_i = c_ia_j - c_ja_i = a_ib_j - a_jb_i = 0 \qquad i, j = 1, \cdots, r$$

(75) $$l_i = \rho_i l_1$$

(76) $$a_1x + b_1y + c_1z = 0$$

(77) $$b_il_j - b_jl_i = (b_ia_j - b_ja_i)x + (b_ic_j - b_jc_i)z = 0$$

(78) $$(b_ic_j - b_jc_i)a_k + (c_ia_j - c_ja_i)b_k + (a_ib_j - a_jb_i)c_k = 0$$

(79) $$\begin{pmatrix} a_1 & a_2 & \cdots & a_r \\ b_1 & b_2 & \cdots & b_r \\ c_1 & c_2 & \cdots & c_r \end{pmatrix}$$

(80) $$M_i \equiv a_ix + b_iy + c_i = 0, \qquad i = 1, \cdots, r$$

(81) $$b_iM_j - b_jM_i = 0x + b_ic_j - b_jc_i = 0$$

BIBLIOGRAPHY. For the history of algebra, see ERIC T. BELL, *The Development of Mathematics*, 2nd ed. (1945); and

DIRK J. STRUIK, *A Concise History of Mathematics*, 3rd rev. ed., 2 vol. (1967). There are many textbooks on elementary algebra, such as MORRIS S. KNEBELMAN and TRACY Y. THOMAS, *Principles of College Algebra* (1942). Somewhat more advanced are the books on the theory of equations: LEONARD E. DICKSON, *New First Course in the Theory of Equations* (1939); JAMES V. USPENSKY, *Theory of Equations* (1948); and LOIS W. GRIFFITHS, *Introduction to the Theory of Equations*, 2nd ed. (1947). Also more advanced are textbooks on abstract algebra, such as GARRETT BIRKHOFF and SAUNDERS MacLANE, *A Survey of Modern Algebra*, 3rd ed. (1965).

(Ed.)

Algebra, Linear and Multilinear

LINEAR ALGEBRA

Linear algebra is the branch of algebra that deals primarily with linear problems, that is to say, problems that depend for the most part on the solution of linear equations. An equation in two or more variables, or "unknowns," is linear if it contains no terms of the 2nd degree or greater; that is, if it contains no products or powers of the variables. The term linear derives from the fact that the graph of a linear equation in x and y is a straight line in the Cartesian xy-plane. Thus a linear equation represents a linear relationship between the variables x and y in a geometric sense. Similarly, a linear equation in three variables x, y, z represents a plane in three-dimensional space, and two such equations considered simultaneously represent the line of intersection of the two planes, provided they are not parallel. When the number of variables is greater than three, there is no longer a simple geometric interpretation because physical space is limited to three dimensions. Nevertheless, it is customary to continue the geometric analogy and to think of the solutions of a linear equation in four variables as constituting a "hyperplane" in a four-dimensional space, and similarly for any finite higher dimension.

The theory and solution of general systems of linear equations is facilitated by the introduction of entities called vectors and matrices. Vectors were originally introduced in order to interpret mathematically a physical quantity such as a velocity or force that has both a magnitude and an associated direction. A matrix is a rectangular array of numbers in a definite order, such as the array of coefficients of the unknowns in a system of linear equations. Rules of computation with vectors, which stem from their original physical interpretation, lead to closed systems of vectors called vector spaces, and matrices can be identified with special functions on these vector spaces called linear transformations. The theory of linear transformations of finite-dimensional vector spaces, which embraces the theory of matrices and that of systems of linear equations, constitutes the subject matter of linear algebra.

Historical and conceptual introduction. Although the term linear algebra is comparatively recent, the mathematical ideas that it embraces are found in the works of many mathematicians over the past two or three hundred years. The main development of the subject, however, has taken place since 1850. Linear algebra originated in the study of systems of linear equations in several unknowns and in the attempt to find general methods for their solution. This attempt led quite early to the notion of a determinant, in 1693 by the German mathematician Gottfried Wilhelm Leibniz. In 1750 the Swiss physicist Gabriel Cramer gave the rule bearing his name, for expressing the solution of n linear equations in n unknowns in terms of determinants. The French mathematician Joseph-Louis Lagrange noted in 1773 the connection between a third-order determinant and the volume of a tetrahedron. Lagrange and Germany's greatest mathematician, Carl Friedrich Gauss, both observed that if the quadratic form $F = ax^2 + 2bxy + cy^2$ is subjected to a linear transformation (see Box, equations 1), expressed in two linear equations, to obtain the new form $f = a'u^2 + 2b'uv + c'v^2$ in the variables u and v, then the effect (see 2) of the transformation is to multiply the discriminant of the form (a function expressed in terms of the coefficients) by a factor depending on the transformation and not the forming.

The term matrix was first used by the British-born U.S. mathematician James Joseph Sylvester in 1850 to designate a rectangular array of numbers from which determinants may be formed. In 1855 Sir Arthur Cayley pointed out that a linear transformation (see 1) can be written, using matrix notation (see 3), and that, if two successive linear transformations (see 3, 4) are written this way, the composite (see 5) of the two transformations involves a third matrix, constructed by a definite rule from the first two, and having certain algebraic properties of a product. This discovery led him to consider matrices as separate entities and to define a matrix product in the way indicated by the composition of the linear transformations (see 6). He also gave the formula, using determinants, for the inverse of a matrix and made his often quoted, prophetic remark, "There would be many things to say about this theory of matrices which should, it seems to me, precede the theory of determinants." Cayley applied these ideas to the transformation of quadratic forms and, in 1858, formulated a complete algebra of matrices considered as mathematical entities in their own right. In 1853 William Rowan Hamilton, the noted Irish mathematician, had also used what was essentially the algebra of matrices under the name of linear and vector functions. Extensive development of the theory of matrices continued, and in the 20th century mathematicians began to focus their attention on the linear transformations rather than on the matrices that represent them. Linear algebra thus became the study of linear transformations of a vector space.

The concept of a vector space grew out of the work of Hamilton on the algebra of vectors. An example of a vector space is the set, $\mathbb{R}_3$, of all ordered triples $[x, y, z]$ in which x, y, z are real numbers. These triples are called vectors with components x, y, z and are added according to the rule (see 7) that respective components of two vectors are added to yield a sum vector. Moreover, if $X = [x, y, z]$ is a vector and c is a real number, then the product cX, called a scalar multiple of X, is defined (see 8) to be the vector obtained by multiplying each component of x by c. In this context a real number is called a scalar, as opposed to a triple $[x, y, z]$ that is a vector. The vector $[0, 0, 0]$ is called the null vector or zero vector and is denoted by O.

Geometric interpretation of vectors

The vectors of $\mathbb{R}_3$ have a geometric interpretation. The vector $X = [x, y, z]$ is represented, relative to a Cartesian coordinate system with origin at O, by the directed line segment $\overrightarrow{OP}$ with initial point O and terminal point at the point P the coordinates of which are (x, y, z). It is

(1) $$\begin{cases} x = \alpha u + \beta v \\ y = \gamma u + \delta v \end{cases}$$

(2) $$a'c' - b'^2 = (\alpha\delta - \beta\gamma)^2(ac - b^2)$$

(3) $$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$$

(4) $$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \alpha' & \beta' \\ \gamma' & \delta' \end{bmatrix} \begin{bmatrix} w \\ z \end{bmatrix}$$

(5) $$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \alpha\alpha' + \beta\gamma' & \alpha\beta' + \beta\delta' \\ \gamma\alpha' + \delta\gamma' & \gamma\beta' + \delta\delta' \end{bmatrix} \begin{bmatrix} w \\ z \end{bmatrix}$$

(6) $$\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \begin{bmatrix} \alpha' & \beta' \\ \gamma' & \delta' \end{bmatrix} = \begin{bmatrix} \alpha\alpha' + \beta\gamma' & \alpha\beta' + \beta\delta' \\ \gamma\alpha' + \delta\gamma' & \gamma\beta' + \delta\delta' \end{bmatrix}$$

(7) $$[x_1, y_1, z_1] + [x_2, y_2, z_2] = [x_1 + x_2, y_1 + y_2, z_1 + z_2]$$

(8) $$cX = [cx, cy, cz]$$

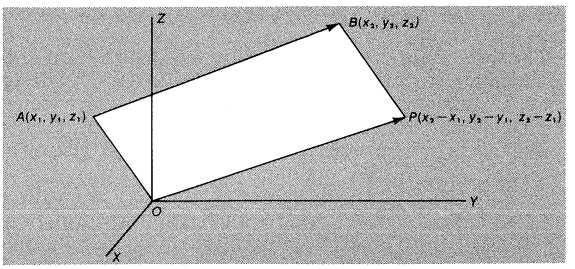


Figure 1: Geometric interpretation of the fact that vectors are invariant under translation. Specifically, the vector represented by the line segment $\overrightarrow{AB}$ is the same as the vector represented by $\overrightarrow{OP}$. The coordinates of each point are shown in parentheses (see text).

further stipulated that two directed line segments that have the same length and the same direction represent the same vector. Thus, vectors are invariant under translation and hence the vector represented by $\overrightarrow{AB}$ with initial point $A(x_1, y_1, z_1)$ and terminal point $B(x_2, y_2, z_2)$ is (see Figure 1) the same (see 9) as the vector with initial point at the origin and with terminus at the point with coordinates $(x_2 - x_1,\ y_2 - y_1,\ z_2 - z_1)$. If X_1 and X_2 are vectors with geometric representations $\overrightarrow{OP}$ and $\overrightarrow{OQ}$, respectively, then $X_1 + X_2$ has the geometric representation $\overrightarrow{OR}$ in which R is the point determined by $\overrightarrow{PR} = \overrightarrow{OQ}$. If O, P, and Q are not collinear, then OR is the diagonal of the parallelogram of which OP and OQ are adjacent edges (Figure 2). If c is any scalar, the geometric representation of cX is a directed segment the length of which is $|c|$ times the length of the segment representing X, and the direction of which is the same or opposite to that of X according as $c > 0$ or $c < 0$. If $c = O$ then $OX = O$.

The set $\mathfrak{U}_2$ of all ordered pairs $[x, y]$ of real numbers, with similar definitions of addition and multiplication by scalars, is another example of a vector space. The vectors of $\mathfrak{U}_2$ can be represented geometrically by directed line segments in a Cartesian plane.

A function or mapping τ from $\mathfrak{U}_3$ to $\mathfrak{U}_3$ can be defined by a system of three linear equations (for example, see 10) that give the components u, v, w of the vector $\tau[x, y, z]$ onto which $[x, y, z]$ is mapped by τ (see 11). The vector $[u, v, w]$ is called the image of $[x, y, z]$ under τ and images of particular vectors such as $[1, 2, 3]$ and $[1, 0, 0]$ can be computed from the given equations (see 12). Because the equations defining the mapping are linear, τ has the properties (see 13) that the image of the sum of two vectors is the sum of their images, and the image of a scalar multiple of a vector is the same scalar multiple of its image. A mapping with these properties is called a linear transformation, in this case from $\mathfrak{U}_3$ to $\mathfrak{U}_3$. Using Cayley's definition of matrix products, equations (10) defining the mapping τ can be written (see 14) $Y = AX$ in which A is the matrix (15) of coefficients of the equations, and X and Y are the vectors $[x, y, z]$ and $[u, v, w]$ written, for technical reasons, as columns. Thus the matrix A represents the linear transformation τ and operates on the vector X to give the vector Y. The linearity properties (13) of τ thus become two rules of matrix algebra (see 16). In the same way, a system (see 17) of two equations in three variables determines a linear transformation σ from $\mathfrak{U}_3$ into $\mathfrak{U}_2$ defined (see 18) by specifying the image in $\mathfrak{U}_2$ of each vector in $\mathfrak{U}_3$ under the mapping σ. The equations defining σ can also be written in matrix form (see 19). In general, both matrices and systems of linear equations may be considered as representations of linear transformations.

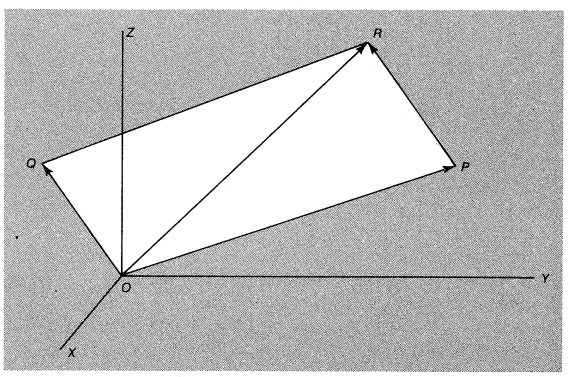


Figure 2: Geometric interpretation of the sum of two vectors. The vectors with representations $\overrightarrow{OP}$ and $\overrightarrow{OQ}$ add to yield the vector with representation $\overrightarrow{OR}$, which is the diagonal of a parallelogram (see text).

(9) $$\overrightarrow{AB} = [x_2 - x_1, y_2 - y_1, z_2 - z_1] = \overrightarrow{OP}$$

(10) $$\begin{cases} 2x - 3y + z = u \\ x + y - 2z = v \\ 3x + y + 5z = w \end{cases}$$

(11) $$\tau[x, y, z] = [u, v, w]$$

(12) $$\begin{cases} \tau[1, 2, 3] = [-1, -3, 20], \\ \tau[1, 1, 1] = [0, 0, 9] \end{cases} \qquad \tau[1, 0, 0] = [2, 1, 3]$$

(13) $$\begin{cases} \tau(X + Y) = \tau X + \tau Y \\ \tau(cX) = c(\tau X) \end{cases}$$

(14) $$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} 2 & -3 & 1 \\ 1 & 1 & -2 \\ 3 & 1 & 5 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

(15) $$A = \begin{bmatrix} 2 & -3 & 1 \\ 1 & 1 & -2 \\ 3 & 1 & 5 \end{bmatrix}$$

(16) $$A(X_1 + X_2) = AX_1 + AX_2, \qquad A(cX) = c(AX)$$

(17) $$\begin{cases} 5x - y + 2z = u \\ x + 3y - 4z = v \end{cases}$$

(18) $$\sigma[x, y, z] = [u, v]$$

(19) $$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 5 & -1 & 2 \\ 1 & 3 & -4 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

Subspaces

A subset $\mathfrak{S}$ of $\mathfrak{U}_3$ is called a subspace of $\mathfrak{U}_3$ if it is closed under addition of vectors and multiplication by scalars; that is, if, whenever X and Y belong to $\mathfrak{S}$, then also $X + Y$ and cX also belong to $\mathfrak{S}$, for all scalars c. The set of all scalar multiples of a fixed vector X in $\mathfrak{U}_3$ is a subspace of $\mathfrak{U}_3$ that consists, if $X \neq O$, of all vectors with geometric representations $\overrightarrow{OP}$ lying in a fixed line through the origin. Similarly, if X, Y are two vectors the geometric representations of which do not lie in one line (or are not parallel), then the set of all vectors of the form $c_1X + c_2Y$ is a subspace consisting of all vectors with geometric representations in the plane through the origin determined by the two vectors X and Y. Apart from the space $\mathfrak{U}_3$ itself (dimension 3), and the zero subspace consisting of the zero vector only (dimension 0), the subspaces of $\mathfrak{U}_3$ are essentially the straight lines (dimension 1) and the planes (dimension 2) that pass through the origin. Linear transformations always map subspaces onto subspaces,

but they may or may not preserve the dimensionality of all subspaces.

Vector spaces. In general, a nonempty set $\mathfrak{U}$, the elements of which will be called vectors, is called a vector space, or a linear space, over a field $\mathfrak{F}$ if addition of vectors and multiplication of vectors by elements of $\mathfrak{F}$ are defined and satisfy 10 postulates (see 20).

Basic concepts. The elements of $\mathfrak{F}$ are called scalars and $\mathfrak{F}$ is called the field of scalars. The field of scalars may be any field (see ALGEBRAIC STRUCTURES), but in most applications is either the field $\mathfrak{R}$ of real numbers or the field $\mathfrak{C}$ of complex numbers. The postulates just referred to imply that $0V = O$, and $(-1)V = -V$. An essential property of a vector space is that if $V_1, V_2, \cdots, V_r$ are vectors in $\mathfrak{U}$ and $c_1, c_2, \cdots, c_r$ are scalars, then the sum $c_1V_1 + c_2V_2 + \cdots + c_rV_r$, called a linear combination of $V_1, V_2, \cdots, V_r$, is a vector in $\mathfrak{U}$.

Examples of a vector space are the spaces $\mathfrak{U}_2$ and $\mathfrak{U}_3$ discussed in the introduction or, more generally, the vector space $\mathfrak{U}_n(\mathfrak{F})$ of all n-tuples $[x_1, x_2, \cdots, x_n]$ in which $x_1, x_2, \cdots, x_n$ belong to $\mathfrak{F}$, and addition and multiplication by scalars are defined as in $\mathfrak{U}_3$. Other examples are the set $\mathfrak{B}$ of all polynomials with coefficients in $\mathfrak{F}$, the set $\mathfrak{B}_n$ of all polynomials of degree $\leqslant n$ with coefficients in $\mathfrak{F}$. Similarly, the set of all continuous real-valued functions defined on the interval of real numbers between two fixed numbers a and b is a vector space over the field $\mathfrak{R}$ of real numbers. The set of all solutions of a linear differential equation (see 21) with real coefficients is also a vector space over $\mathfrak{R}$.

Subspaces. A subset $\mathfrak{S}$ of a vector space $\mathfrak{U}$ over $\mathfrak{F}$ is called a subspace of $\mathfrak{U}$ if $\mathfrak{S}$ itself is a vector space over $\mathfrak{F}$ under addition and scalar multiplication that it inherits from $\mathfrak{U}$; that is, if, whenever V_1, V_2 belong to $\mathfrak{S}$, $V_1 + V_2$ is in $\mathfrak{S}$ and cV_1 is in $\mathfrak{S}$, for every scalar c. The space $\mathfrak{U}$, itself, and the zero space consisting of the zero vector only are included among the subspaces of $\mathfrak{U}$. If $\mathfrak{H}$ is any subset of $\mathfrak{U}$, the set $\mathfrak{S}$ of all linear combinations of vectors in $\mathfrak{H}$ is a subspace of $\mathfrak{U}$, called the subspace spanned by the set $\mathfrak{H}$. It is the smallest subspace of $\mathfrak{U}$ containing all vectors in $\mathfrak{H}$. The set $\mathfrak{H}$ may be finite or infinite, but a linear combination (see 22) must have only a finite number of nonzero terms. If the subspace spanned by the set $\mathfrak{H}$ is $\mathfrak{U}$ itself, then $\mathfrak{H}$ is called a spanning set or generating system of $\mathfrak{U}$. If $\mathfrak{U}$ is spanned by a finite set of vectors, $\mathfrak{U}$ is said to be finitely generated or finite-dimensional. Linear algebra is concerned primarily with finite-dimensional vector spaces. Although many of the theorems of linear algebra can be generalized to infinite-dimensional spaces, in order to do so it is usually necessary to introduce a topology on the space and the results are generally analytic in nature rather than purely algebraic (see ANALYSIS, FUNCTIONAL).

Sums and intersections of subspaces

If $\{\mathfrak{S}_\alpha\}$ is any set of subspaces of $\mathfrak{U}$, then their intersection, consisting of all vectors of $\mathfrak{U}$ that belong to each of the subspaces $\mathfrak{S}_\alpha$, is a subspace of $\mathfrak{U}$. For example, the space spanned by the set $\mathfrak{H}$ is the intersection of all subspaces $\mathfrak{S}_\alpha$ that contain all vectors in $\mathfrak{H}$. The sum of the subspaces $\mathfrak{S}_\alpha$ is the space spanned by the set of all vectors V such that V belongs to at least one of the subspaces $\mathfrak{S}_\alpha$. The sum of a finite number of subspaces $\mathfrak{S}_1, \mathfrak{S}_2, \cdots, \mathfrak{S}_r$ is the set of all vectors $\sum_{i=1}^{r} S_i$, in which S_i belongs to $\mathfrak{S}_i$.

Linear dependence. The vectors $V_1, V_2, \cdots, V_r$ in a vector space $\mathfrak{U}$ over $\mathfrak{F}$ are linearly dependent, or constitute a linearly dependent set, if there exist scalars $c_1, c_2, \cdots, c_r$ in $\mathfrak{F}$, not all zero, such that $c_1V_1 + c_2V_2 + \cdots + c_rV_r = O$. If no such scalars exist, the vectors are linearly independent. For example, the vectors $X = [2, 5, -1]$, $Y = [1, 1, 4]$, $Z = [1, 3, -2]$ in $\mathfrak{U}_3(\mathfrak{R})$ are linearly dependent, because $2X - 1Y - 3Z = O$; but the vectors $V = [2, 5]$, $W = [4, 1]$ in $\mathfrak{U}_2(\mathfrak{R})$ are linearly independent because $c_1V + c_2W = O$ implies $2c_1 + 4c_2 = 0$, $5c_1 + c_2 = 0$, and the only solution of these two equations is $c_1 = c_2 = 0$. An infinite set of vectors is linearly dependent if it contains a finite subset that is linearly dependent. Linear dependence of two vectors means that one of them is a scalar multiple of the other. Linear dependence of three vectors in $\mathfrak{U}_3(\mathfrak{R})$ means that the three vectors have geometric representations that lie in one plane passing through the origin. In general, a set of vectors in $\mathfrak{U}$ is linearly dependent if, and only if, one of the vectors is in the subspace of $\mathfrak{U}$ spanned by the other vectors of the set.

A basis of vector space $\mathfrak{U}$ is a linearly independent set of vectors that spans $\mathfrak{U}$. Every vector space has a basis. If $\mathfrak{U}$ has a basis containing a finite number of vectors, then every basis of $\mathfrak{U}$ contains the same number of vectors. The number of vectors in a basis of $\mathfrak{U}$ is called the dimension of $\mathfrak{U}$. A finitely generated vector space has finite dimension, because every set of vectors that spans $\mathfrak{U}$ has a maximal linearly independent subset that forms a basis of $\mathfrak{U}$. The space $\mathfrak{U}_n(\mathfrak{F})$ is n-dimensional, because it is spanned by the n linearly independent vectors $E_1 = [1, 0, \cdots, 0]$, $E_2 = [0, 1, 0, \cdots, 0]$, $E_n = [0, 0, \cdots, 1]$, which therefore form a basis of $\mathfrak{U}_n(\mathfrak{F})$. If $\mathfrak{U}$ has dimension n, then any set of n linearly independent vectors is a basis of $\mathfrak{U}$, and any set of r linearly independent vectors, r less than n, is a subset of a basis.

(20) (1) The sum $V_1 + V_2$ of any two vectors V_1, V_2 is a uniquely defined vector in $\mathfrak{U}$. (2) Addition of vectors is commutative: $V_1 + V_2 = V_2 + V_1$. (3) Addition of vectors is associative: $(V_1 + V_2) + V_3 = V_1 + (V_2 + V_3)$. (4) There is a zero vector O such that $V + O = V$ for all V in $\mathfrak{U}$. (5) Every vector V has a unique negative $-V$ such that $V + (-V) = O$. (6) For any element c of $\mathfrak{F}$ and any vector V in $\mathfrak{U}$ the product cV is a uniquely defined vector in $\mathfrak{U}$. (7) $c(V_1 + V_2) = cV_1 + cV_2$. (8) $(c_1 + c_2)V = c_1V + c_2V$. (9) $c_1(c_2V) = (c_1c_2)V$. (10) $1V = V$.

(21) $$\frac{d^n y}{dx^n} + a_1 \frac{d^{n-1} y}{dx^{n-1}} + \cdots + a_n y = 0$$

(22) $$\Sigma c_i V_i$$

(23) $$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

Basis of a vector space

If $F_1, F_2, \cdots, F_n$, or, more briefly (F), is a basis of $\mathfrak{U}$, then every vector V in $\mathfrak{U}$ has a unique representation, namely, $V = x_1F_1 + x_2F_2 + \cdots + x_nF_n$, as a linear combination of the basis vectors. The vector $X = [x_1, x_2, \cdots, x_n]$ in $\mathfrak{U}_n(\mathfrak{F})$ is called the coordinate vector of V relative to the F-basis, and $x_1, x_2, \cdots, x_n$ are called the coordinates of V relative to this basis. If the coordinate vector X is denoted by $\alpha(V)$, then the mapping α maps the space $\mathfrak{U}$ onto the space $\mathfrak{U}_n(\mathfrak{F})$. This mapping is one-to-one and satisfies the rules that the image of the sum of vectors is the sum of the images and that the image of a scalar multiple of a vector is the same scalar multiple of its image. Such a mapping is called an isomorphism and the two spaces $\mathfrak{U}$ and $\mathfrak{U}_n(\mathfrak{F})$ are said to be isomorphic. In this way, every n-dimensional vector space over $\mathfrak{F}$ is isomorphic to $\mathfrak{U}_n(\mathfrak{F})$, and every basis of $\mathfrak{U}$ provides such an isomorphism.

Matrices. A matrix is a rectangular array (see 23) in which the individual entries a_{ij}, called the elements of the matrix, belong to some fixed field $\mathfrak{F}$ of scalars or possibly to a commutative ring (see ALGEBRAIC STRUCTURES) such as a ring of polynomials. If A has m rows and n columns, it is called an $m \times n$ matrix and is designated by $A = [a_{ij}]$ if m and n are assumed known. If $m = n$, A is a square matrix of order n and the elements $a_{11}, a_{22}, \cdots, a_{nn}$ are called the diagonal elements of A. If all elements of A other than the diagonal elements are 0,

A is a diagonal matrix. If all elements below (or above) the main diagonal are 0, A is a triangular matrix.

If $A = [a_{ij}]$ and $B = [b_{ij}]$ are $m \times n$ matrices the sum $A + B$ is the matrix $[a_{ij} + b_{ij}]$ obtained by adding corresponding elements. A matrix all of whose elements are 0 is a zero matrix or null matrix and denoted by O. If O is an $m \times n$ zero matrix then $A + O = A$. If c is any scalar, then cA is the matrix $[ca_{ij}]$. With these definitions, the set $[\mathfrak{F}]_{m,n}$ of all $m \times n$ matrices, with elements in the field $\mathfrak{F}$ is a vector space over $\mathfrak{F}$. A basis for this vector space is the set of mn matrices $\{E_{ij}\}$ in which E_{ij} is the $m \times n$ matrix with 1 in the ith row, jth column, and 0 in every other place. Thus the dimension of $[\mathfrak{F}]_{m,n}$ is mn.

If $A = [a_{ij}]$ is an $m \times n$ matrix, and $B = [b_{ij}]$ is an $n \times r$ matrix, then the product AB is defined by the product rule (see 24) to be an $m \times r$ matrix $D = [d_{ij}]$. The product AB is defined only when the number of columns in A is equal to the number of rows in B. If $C = [c_{ij}]$ is an $r \times s$ matrix, then $(AB)C = A(BC)$. Thus matrix multiplication is associative, but not commutative since in general $AB \neq BA$. The $n \times n$ diagonal matrix I with diagonal elements all equal to 1 is called the identity or unit matrix of order n, since $IA = AI = A$ for any $n \times n$ matrix A. The distributive laws hold; that is, $A(B + C) = AB + AC$, and $(B + C)A = BA + CA$, whenever the indicated products and sums are defined. The set $[\mathfrak{F}]_n$ of all $n \times n$ matrices over $\mathfrak{F}$ is therefore a vector space over $\mathfrak{F}$, on which an associative multiplication is defined that is distributive with respect to addition. Such a system is called a linear associative algebra over $\mathfrak{F}$, or briefly, an algebra over $\mathfrak{F}$.

The transpose of a matrix

If A is an $m \times n$ matrix, then the transpose of A is the $n \times m$ matrix A^T obtained by writing the rows of A as the columns of A^T. If B is an $n \times r$ matrix, $(AB)^T = B^T A^T$. A $1 \times n$ matrix is called a row vector and can be identified with a vector of $\mathfrak{U}_n(\mathfrak{F})$, because the vector space $[\mathfrak{F}]_{1,n}$ is isomorphic to $\mathfrak{U}_n(\mathfrak{F})$. Similarly, an $n \times 1$ matrix is called a column vector and its transpose is a vector of $\mathfrak{U}_n(\mathfrak{F})$. The row space of an $m \times n$ matrix A is the subspace of $\mathfrak{U}_n(\mathfrak{F})$ spanned by the row vectors of A. The column space of A is the subspace of $\mathfrak{U}_m(\mathfrak{F})$ spanned by the transposes of the column vectors of A. For any matrix, the row space and column space have the same dimension, called the rank of the matrix.

The solution of systems of linear equations and the determination of the rank of a matrix A are facilitated by the application to A of elementary row transformations. These are of three types: (1) interchange of two rows, (2) multiplication of one row vector by a nonzero scalar, (3) addition of a scalar multiple of one row vector to another. A matrix A is row-equivalent to B ($A \sim B$) if A can be transformed into B by a succession of elementary row transformations. Row-equivalence is reflexive ($A \sim A$), symmetric ($A \sim B$ implies $B \sim A$), and transitive ($A \sim B$ and $B \sim C$ imply $A \sim C$). Every matrix is row-equivalent to one and only one triangular matrix (see 25), characterized by four conditions: (1) all zero row vectors, if any, occur below all nonzero row vectors; (2) the first nonzero element in each row is 1; (3) if the first nonzero element in the ith row occurs in the j_ith column, every other element in the j_ith column is 0; and (4) $j_1 < j_2 < j_3 < \cdots < j_n$. The matrix R is called the reduced row-echelon form of A. Two matrices are row-equivalent if, and only if, they have the same reduced row-echelon form. Row equivalent matrices have the same row space and therefore the same rank. Because the nonzero row vectors of R are linearly independent, they constitute a basis of the row space of A, and the rank of A is equal to the number of nonzero rows in the reduced row-echelon form of A. An $n \times n$ matrix A is singular if its rank is less than n; nonsingular if its rank is equal to n. An $n \times n$ matrix is nonsingular if, and only if, its row vectors are linearly independent, and hence if, and only if, its reduced row-echelon form is the identity matrix.

An elementary matrix is a matrix obtained by applying an elementary row transformation to the identity matrix I. Every elementary row transformation of A can be achieved by multiplying A from the left by the elementary matrix obtained by applying the same elementary transformation to I. Every elementary matrix is nonsingular, and every nonsingular matrix is a product of elementary matrices. Thus, two $m \times n$ matrices A and B are row-equivalent if, and only if, there exists a nonsingular $m \times m$ matrix P such that $B = PA$. If A is a nonsingular $n \times n$ matrix, it is row-equivalent to I, and $PA = I$. In this case, the matrix P is unique and also satisfies $AP = I$. It is called the inverse of A, denoted by A^{-1}, and A is said to be invertible. An $n \times n$ matrix has an inverse if, and only if, it is nonsingular. The inverse, when it exists, is unique.

The matrix inverse

(24) $$d_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj}$$

(25) $$R = \begin{bmatrix} 0 \cdots 0 & 1 & c_{1j_1+1} \cdots 0 & \cdots & 0 \cdots c_{1n} \\ 0 \cdots 0 & 0 & 0 \cdots \cdots 1 & c_{2j_2+1} & 0 \cdots c_{2n} \\ & & & & 1 \cdots c_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 \cdots 0 & 0 & \cdots \cdots 0 & \cdots & 0 \cdots 0 \end{bmatrix}$$

(26) $$\det A = b_{i1}a_{i1} + b_{i2}a_{i2} + \cdots + b_{in}a_{in}$$

(27) $$A^{-1} = \frac{1}{\det A} C^T$$

in which $C = [c_{ij}]$ and $c_{ij} = (-1)^{i+j} \det A_{ij}$ in which A_{ij} is the matrix of order $n-1$ obtained from A by deleting the ith row and jth column.

Elementary column transformations and column-equivalence of two matrices are defined analogously. Every elementary column transformation of A can be achieved by multiplying A from the right by an elementary matrix, and B is column equivalent to A if, and only if, there exists a nonsingular $n \times n$ matrix Q such that $B = AQ$. The matrix B is equivalent to A if A can be transformed into B by a succession of elementary row and/or column transformations; hence B is equivalent to A if, and only if, there exist nonsingular matrices P and Q such that $B = PAQ$. Equivalence is reflexive, symmetric, and transitive; and every matrix A of rank r is equivalent to a diagonal matrix in which the first r diagonal elements are 1, and all other elements are 0. This is called the canonical form of A. Two matrices are equivalent if, and only if, they have the same canonical form and hence if, and only if, they have the same rank.

Associated with each $n \times n$ matrix A is a scalar called the determinant of A and denoted by det A. The determinant has these basic properties: (1) If $A = [a_{ij}]$, det A is, for each i, a linear homogeneous function (see 26) of the elements of the ith row in which the coefficients depend only on the a_{kj} with $k \neq i$. (2) The interchange of two rows of A changes the sign of det A. (3) det $I = 1$. These properties actually determine det A uniquely. Other important properties of the determinant are: det $A = 0$ if and only if A is singular; and for any two $n \times n$ matrices det $AB =$ det A det B. The determinant of a triangular matrix is equal to the product of the elements in its main diagonal. Elementary row transformations of type (3) leave the determinant of A unchanged, those of type (1) change the sign of det A; and those of type (2) multiply det A by the same nonzero scalar by which one row of A is multiplied. Thus for any $n \times n$ matrix A, det A can be found by reducing A to its reduced row-echelon form R, keeping track of the changes in det A that result from the row transformations used, and then evaluating det R by finding the product of its diagonal elements. If det $A \neq 0$, then A is invertible, and there is a formula (see 27) for the inverse of A in terms of determinants.

Conditions for solutions of systems of equations

Three distinct situations arise in attempting to solve general systems of linear equations, namely, a system may have a unique solution, infinitely many solutions, or no solution. These three possibilities can be illustrated by examples (see 28). A system of the first type is called an independent system; of the second type, a dependent system; and of the third type, an inconsistent system. Geometrically, in the case of two unknowns the first type represents two intersecting straight lines; the second type represents two coincident lines; and the third type represents two parallel lines.

A general system (see 29) of m equations with n unknowns is said to be consistent if solutions exist. This system can be written in matrix form, $AX = C$, in which $A = [a_{ij}]$ is called the coefficient matrix of the system, and X and C are the column vectors $[x_1, x_2, \cdots, x_n]^T$ and $[c_1, c_2, \cdots, c_m]^T$, respectively. A vector X that satisfies this equation is called a solution vector of the system. A solution exists if, and only if, the vector C of constant terms is in the space spanned by the column vectors of A or if and only if the augmented matrix of the system (see 30), constructed from A by adjoining the column vector C, has the same rank as the coefficient matrix A. The system is said to be homogeneous if $C = O$, nonhomogeneous if $C \neq O$. Every solution of a consistent nonhomogeneous system $AX = C$ has the form $S_0 + S$ in which S_0 is any particular solution of $AX = C$, and S is a solution of $AX = O$. Conversely, if $AS_0 = C$ and $AS = O$, then $A(S_0 + S) = AS_0 + AS = C$ and $S_0 + S$ is a solution of $AX = C$. If the number of unknowns is equal to the number of equations, then $AX = C$ has a unique solution if, and only if, A is non-singular, and this is given by $X = A^{-1}C$.

A homogeneous system $AX = O$ is always consistent because it has the solution $X = O$. The set of all solutions of this system is a subspace of $\mathfrak{U}_n(\mathfrak{F})$, called the solution space of the system or the nullspace of the matrix A. The dimension of the solution space is $n - r$, in which r is the rank of A and n is the number of unknowns. This means that $AX = O$ has $n - r$ linearly independent solution vectors, and every solution is a linear combination of these. Examples (see 31) illustrate these facts and are easy to construct.

The general solution of a homogeneous system $AX = O$ can be obtained by reducing the matrix A to row-echelon form. This process is based on the principle that if two matrices A and B are row-equivalent then the systems $AX = O$ and $BX = O$ have the same solution space. If A has rank r, and R is the reduced row-echelon form of A, then R has a 1 in the ith row, j_ith column ($i = 1, 2, \cdots, r$), and all other entries in the j_ith column are zero. Thus the system $RX = O$ can be solved immediately for the r unknowns $x_{j_1}, x_{j_2}, \cdots, x_{j_r}$ in terms of the remaining $n - r$ unknowns which may be assigned arbitrary values.

A nonhomogeneous system $AX = C$ can be solved in the same way. If R is the reduced row-echelon form of the augmented matrix, the first n columns of R provide the reduced row-echelon form R' of A. The system is consistent if, and only if, R and R' have the same number of nonzero rows. Again, provided the system is consistent its general solution can be read from the matrix R as in the homogeneous case. An example (see 32) illustrates these matters.

Linear transformations. If $\mathfrak{U}$ and $\mathfrak{W}$ are vector spaces over the same field $\mathfrak{F}$, a mapping σ that associates a unique vector σV in $\mathfrak{W}$ with each vector V of $\mathfrak{U}$ is a linear transformation if the image of the sum of any two vectors is the sum of their images and the image of a scalar multiple of any vector is the same scalar multiple of its image (see 33). A linear operator in $\mathfrak{U}$ is a linear transformation of $\mathfrak{U}$ into itself. If A is any $m \times n$ matrix over $\mathfrak{F}$, then the mapping σ from $\mathfrak{U}_n(\mathfrak{F})$ to $\mathfrak{U}_m(\mathfrak{F})$, defined by $\sigma X = AX$, is a linear transformation. If $\mathfrak{B}$ is the vector space, over the field $\mathfrak{R}$ of real numbers, of all polynomials in x with real coefficients, and if $\delta[p(x)]$ is the derivative of $p(x)$, then δ is a linear operator in $\mathfrak{B}$. Similarly, the integral (see 34) defines a linear operator τ in the space of continuous functions f from $\mathfrak{R}$ into $\mathfrak{R}$. If τ is any linear transformation that maps a vector space $\mathfrak{U}$ into a vector space $\mathfrak{W}$, then τ maps the zero vector of $\mathfrak{U}$ on the zero vector of $\mathfrak{W}$. Moreover, τ preserves linear combinations (see 35) and maps subspaces of $\mathfrak{U}$ onto subspaces of $\mathfrak{W}$. If $\mathfrak{U}$ is finite-dimensional with basis (E), and if $F_1, F_2, \cdots, F_n$ are arbitrary vectors of $\mathfrak{W}$, there exists a unique linear transformation τ from $\mathfrak{U}$ into $\mathfrak{W}$ such that $\tau E_i = F_i$, and hence (see 36) τ is completely determined by its effect on a basis of $\mathfrak{U}$.

Linear operators

The set $\mathfrak{L}(\mathfrak{U}, \mathfrak{W})$ of all linear transformations from $\mathfrak{U}$ to

(28)
$$\begin{cases} (a)\ x-2y=6 \\ \quad 2x+3y=5 \\ \quad \text{Solution: } x=4,\ y=-1 \\ (c)\ x-2y=6 \\ \quad 2x-4y=9 \\ \quad \text{Solutions: None} \end{cases} \qquad (b)\ \begin{array}{l} x-2y=6 \\ 2x-4y=12 \\ \text{Solutions: Infinitely many} \end{array}$$

(29)
$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = c_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = c_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = c_m \end{cases}$$

(30)
$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & c_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & c_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & c_m \end{bmatrix}$$

(31) $x - 5y + 2z = 0$ has independent solutions $S_1 = [5, 1, 0]$ and $S_2 = [-2, 0, 1]$ and every solution has the form $hS_1 + kS_2 = [5h - 2k, h, k]$, in which h and k are arbitrary. Here $n = 3$ and $r = 1$. The nonhomogeneous equation $x - 5y + 2z = 3$ has a solution $S_0 = [2, 1, 3]$ and hence every solution of this equation is obtained from the formula $S_0 + hS_1 + kS_2 = [2 + 5h - 2k, 1 + h, 3 + k]$.

(32) For example, the augmented matrix of the system
$x - 2y + 2z + w = 3$
$2x - 2y + 4z - 2w = 5$
$x - 4y + 2z + 5w = 4$
and its row-echelon form are
$$\begin{bmatrix} 1 & -2 & 2 & 1 & 3 \\ 2 & -2 & 4 & -2 & 5 \\ 1 & -4 & 2 & 5 & 4 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & 2 & -3 & 2 \\ 0 & 1 & 0 & -2 & -\frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$
and the general solution is therefore $x = -2z + 3w + 2$, $y = 2w - \frac{1}{2}$, in which z and w are arbitrary. In vector form the general solution is $[2, -\frac{1}{2}, 0, 0] + z[-2, 0, 1, 0] + w[3, 2, 0, 1]$, in which $z[-2, 0, 1, 0] + w[3, 2, 0, 1]$ is the general solution of the homogeneous system $AX = O$ and $[2, -\frac{1}{2}, 0, 0]$ is the particular solution obtained by putting $z = w = 0$.

(33) $\sigma(V_1 + V_2) = \sigma V_1 + \sigma V_2$ and $\sigma(cV) = c(\sigma V)$

(34) $(\tau f)(x) = \int_0^x f(t)\,dt$

(35) $\tau(O) = O$ and $\tau(c_1V_1 + c_2V_2 + \cdots + c_nV_n) = c_1(\tau V_1) + c_2(\tau V_2) + \cdots + c_n(\tau V_n)$

(36) $\tau(c_1E_1 + c_2E_2 + \cdots + c_nE_n) = c_1F_1 + c_2F_2 + \cdots + c_nF_n$

$\mathfrak{W}$ can be given the structure of a vector space. If σ, τ are such linear transformations, then sums and scalar multiples are defined by forming sums and scalar multiples of images (see 37). The zero element of $\mathfrak{L}(\mathfrak{U}, \mathfrak{W})$ is the zero mapping O, defined by $OV = O$ for all V, and $(-\sigma)V$ is defined to be $-(\sigma V)$. If $\mathfrak{U}$ and $\mathfrak{W}$ are finite-dimensional, and σ is a linear transformation that maps $\mathfrak{U}$ into $\mathfrak{W}$, then σ is uniquely determined by equations (38) or by the matrix $A_\sigma^{(E,F)}$ (see 39), called the matrix of σ relative to the E- and F-bases. If a vector V in $\mathfrak{U}$ has a coordinate vector X relative to the E-basis, and if σV has a coordinate vector Y relative to the F-basis, then $Y = A_\sigma^{(E,F)}X$. Because there is an isomorphism of the vector space $\mathfrak{L}(\mathfrak{U}, \mathfrak{W})$ onto the vector space of all $m \times n$ matrices over $\mathfrak{F}$ (see 40), the dimension of $\mathfrak{L}(\mathfrak{U}, \mathfrak{W})$ is mn.

If, $\mathfrak{U}$, $\mathfrak{W}$, and $\mathfrak{F}$ are vector spaces over $\mathfrak{F}$, and if σ and τ are linear transformations that map $\mathfrak{U}$ into $\mathfrak{W}$ and $\mathfrak{W}$ into $\mathfrak{F}$, respectively, then the composite mapping, or product, $\tau\sigma$ that maps $\mathfrak{U}$ into $\mathfrak{F}$ is defined by $(\tau\sigma)V = \tau(\sigma V)$ and is a linear transformation. Hence, by taking $\mathfrak{F} = \mathfrak{W} = \mathfrak{U}$, a multiplication is defined (see 41) in the vector space $\mathfrak{L}(\mathfrak{U}, \mathfrak{U})$ that is associative and distributive with respect to addition, and $\mathfrak{L}(\mathfrak{U}, \mathfrak{U})$ becomes a linear associative algebra over $\mathfrak{F}$. In the case of a linear operator σ that maps $\mathfrak{U}$ into $\mathfrak{U}$, the corresponding matrix $A_\sigma^{(E,E)}$ is denoted by $A_\sigma^{(E)}$ and is called the matrix of σ relative to the E-basis of $\mathfrak{U}$. The vector space isomorphism α, defined by $\alpha(\sigma) = A_\sigma^{(E)}$, is also an algebra isomorphism of the algebra of linear operators in $\mathfrak{U}$ onto the algebra of $n \times n$ matrices over $\mathfrak{F}$.

The image of a linear transformation

If τ is a linear transformation of $\mathfrak{U}$ into $\mathfrak{W}$, the image of τ, written im τ, is the subspace of $\mathfrak{W}$ consisting of all vectors τV where V belongs to $\mathfrak{U}$. The rank of τ is the dimension of im τ. The subspace $\mathfrak{N}$ of $\mathfrak{U}$, consisting of all vectors V such that $\tau V = O$, is called the nullspace or kernel of τ. The dimension of $\mathfrak{N}$, written dim $\mathfrak{N}$, is called the nullity of τ. If $\mathfrak{U}$ is finite-dimensional, then for every τ in $\mathfrak{L}(\mathfrak{U}, \mathfrak{W})$, a simple relationship holds between rank, nullity, and dimension, namely: rank τ + nullity τ = dim $\mathfrak{U}$. This relationship states, in effect, (see 42) that the solution space of the system of equations $AX = O$ has dimension $n - r$, in which n is the number of unknowns and r = rank A. A linear transformation is one-to-one (that is, $\tau V_1 = \tau V_2$ implies $V_1 = V_2$) if, and only if, its nullspace is O. If $\mathfrak{U}$ is finite-dimensional and τ belongs to $\mathfrak{L}(\mathfrak{U}, \mathfrak{U})$ then im $\tau = \mathfrak{U}$ only if the nullspace of τ is O. In this case, τ is called a nonsingular or invertible operator, and there exists an inverse mapping τ^{-1}, also a linear operator, such that $\tau^{-1}\tau = \tau\tau^{-1} = I$, in which I is the identity mapping on $\mathfrak{U}$. An operator τ is nonsingular, if and only if the matrix $A_\tau^{(E)}$ is nonsingular; then the matrix of the inverse of τ is the inverse of the matrix of τ relative to the E-basis.

If (E) and (F) are two different bases of $\mathfrak{U}$, and if X and Y are the coordinate vectors of vector V relative to the E- and F-bases, respectively, then $X = PY$ in which P is a nonsingular $n \times n$ matrix, depending only on the two bases, and is called the transition matrix from the E- to the F-basis. If τ is a linear operator in $\mathfrak{U}$, then the matrices of τ relative to the two bases are related by equation (43) in which P is the transition matrix from the E- to the F-basis. $A_\tau^{(F)}$ and $A_\tau^{(E)}$ have the same determinant. Thus det $A_\tau^{(E)}$ is independent of the basis (E) and is called the determinant of the linear operator τ and written det τ. If τ is a linear operator in $\mathfrak{U}_2(\mathfrak{R})$, then the area of the parallelogram with adjacent edges τV_1 and τV_2 is equal to det τ times the area of the parallelogram with edges V_1 and V_2. Similarly in $\mathfrak{U}_3(\mathfrak{R})$, the parallelepiped with edges V_1, V_2, V_3 is mapped by τ onto the parallelepiped with edges τV_1, τV_2, τV_3 whose volume is det τ times that of the original parallelepiped. If det $\tau = 0$, τ is singular and im τ is a proper subspace of $\mathfrak{U}$.

If A is an $n \times n$ matrix over $\mathfrak{F}$ and x is an indeterminate, the polynomial $f(x)$ (see 44) is called the characteristic polynomial of A. The roots in $\mathfrak{F}$ of the characteristic equation $f(x) = 0$ are called the characteristic values or eigenvalues of A. The Cayley-Hamilton theorem states that every $n \times n$ matrix satisfies its characteristic equation (see 45). It follows that there exists a polynomial $m(x)$ of minimum degree $r \leqslant n$ that $m(A) = O$. If $m(x)$ is monic—*i.e.*, has leading coefficient 1—then it is unique and is called the minimal polynomial of A. It is necessarily a divisor of the characteristic polynomial $f(x)$. Moreover, every irreducible factor of $f(x)$ is also a factor of $m(x)$, although it may occur with a smaller multiplicity. If the field of scalars is algebraically closed—*e.g.*, if $\mathfrak{F}$ is the field of the complex numbers—then $f(x)$ has exactly n roots in $\mathfrak{F}$ if a root of multiplicity r is counted r times. Thus, every $n \times n$ matrix over the complex field has exactly n eigenvalues. If A has n distinct eigenvalues, then $m(x) = f(x)$.

Similar matrices

If A and B are $n \times n$ matrices, B is said to be similar to A if there exists a nonsingular matrix P such that $B = P^{-1}AP$. Like equivalence, similarity is a reflexive, symmetric, and transitive relation. Moreover, for a fixed matrix P, the mapping α defined by $\alpha(A) = P^{-1}AP$ is an isomorphism of the algebra $[\mathfrak{F}]_n$ with itself. Thus, for any polynomial $g(x)$ with coefficients in $\mathfrak{F}$, $P^{-1}g(A)P = g(P^{-1}AP)$. Similar matrices have the same characteristic polynomials, the same minimum polynomials, the same eigenvalues, and the same determinants.

If τ is a linear operator on an n-dimensional vector space $\mathfrak{U}$, and if $A_\tau = A_\tau^{(E)}$ is the matrix of τ relative to a fixed basis (E), then the matrix of τ relative to any other basis is similar to A_τ. Thus, the characteristic polynomial of A_τ can also be called the characteristic polynomial of τ. The minimal polynomial of τ—that is, the unique monic polynomial $m(x)$ of least degree such that $m(\tau) = O$—is also the minimal polynomial of $A_\tau^{(E)}$ and is therefore independent of the basis (E). The eigenvalues of τ are those scalars c for which the operator

(37) $(\sigma + \tau)V = \sigma V + \tau V$ and $(c\sigma)V = c(\sigma V)$

(38)
$$\begin{cases} \text{For basis } E_1, E_2, \cdots, E_n \text{ of } \mathfrak{V} \\ \text{and basis } F_1, F_2, \cdots, F_m \text{ of } \mathfrak{W}: \\ \sigma E_1 = a_{11}F_1 + a_{21}F_2 + \cdots + a_{m1}F_m \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \sigma E_n = a_{1n}F_1 + a_{2n}F_2 + \cdots + a_{mn}F_m \end{cases}$$

(39)
$$A_\sigma^{(E,F)} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

(40)
$$\begin{cases} \text{Because } A_{\sigma+\tau}^{(E,F)} = A_\sigma^{(E,F)} + A_\tau^{(E,F)} \text{ and } A_{c\sigma}^{(E,F)} = cA_\sigma^{(E,F)} \\ \text{the mapping } \alpha: \mathfrak{L}(\mathfrak{U}, \mathfrak{W}) \to [\mathfrak{F}]_{m,n} \text{ defined by} \\ \alpha(\sigma) = A_\sigma^{(E,F)} \text{ is an isomorphism.} \end{cases}$$

(41)
$$\begin{cases} \text{Also, if } (E), (F) \text{ and } (G) \text{ are bases for the three spaces,} \\ \text{then } A_{\sigma\tau}^{(E,G)} = A_\sigma^{(E,F)} A_\tau^{(F,G)}. \text{ If } \sigma \text{ and } \tau \text{ are both linear} \\ \text{operators in } \mathfrak{U} \text{ then } \tau\sigma \text{ and } \sigma\tau \text{ are also linear operators} \\ \text{in } \mathfrak{U}. \end{cases}$$

(42)
$$\begin{cases} \text{The isomorphism } \alpha: \mathfrak{L}(\mathfrak{U}, \mathfrak{W}) \to [\mathfrak{F}]_{m,n} \text{ maps im } \tau \\ \text{onto the column space of } A_\tau^{(E,F)} \text{ and hence rank } \tau = \\ \text{rank } A_\tau^{(E,F)}. \text{ Also } \alpha \text{ maps the nullspace of } \tau \text{ onto the} \\ \text{nullspace of } A_\tau^{(E,F)}. \end{cases}$$

(43) $A_\tau^{(F)} = P^{-1}A_\tau^{(E)}P$

(44) $f(x) = (-1)^n \det(A - xI)$

(45)
$$\begin{cases} \text{If } f(x) = x^n + a_1x^{n-1} + \cdots + a_0 \text{ is the characteristic} \\ \text{polynomial of } A, \text{ then } f(A) = A^n + a_1A^{n-1} + \cdots + \\ + a_0I = O \end{cases}$$

$\tau - cI$ has nonzero nullspace. These are precisely the eigenvalues of A_τ because $A_{\tau-cI} = A_\tau - cI$. If c is an eigenvalue of τ, a nonzero vector V in the nullspace of $\tau - cI$ is called an eigenvector of τ belonging (or corresponding) to c. It has the property that $\tau V = cV$. A substance $\mathfrak{S}$ of $\mathfrak{U}$ is invariant under τ if τS is in $\mathfrak{S}$ whenever S is in $\mathfrak{S}$. Thus, an eigenvector of τ spans a one-dimensional subspace of $\mathfrak{U}$ that is invariant under τ. If V is an eigenvector of τ belonging to c, then the coordinate vector X of V relative to the basis $E_1, \cdots, E_n$ satisfies equation (46). Hence $A_\tau^{(E)}X$ is equal to cX and X is called an eigenvector of the matrix $A_\tau^{(E)}$ corresponding to c.

A vector space $\mathfrak{U}$ is a direct sum of subspaces $\mathfrak{S}_i$ (see 47) if $\mathfrak{S}$ is the sum of these subspaces and each of subspaces $\mathfrak{S}_i$ has zero intersection with the sum of the remaining subspaces. If $\mathfrak{U}$ is such a direct sum, then every vector V in $\mathfrak{U}$ has a unique representation as the sum (see 48) of vectors $\{S_i\}$, in which S_i belongs to $\mathfrak{S}_i$. If τ is a linear operator on a finite-dimensional space $\mathfrak{U}$ with minimal polynomial $m(x)$, and if the factorization (see 49) of $m(x)$ into irreducible factors over $\mathfrak{F}$ has the typical factor $[p_i(x)]^{e_i}$, then $\mathfrak{U}$ is the direct sum of the nullspaces $\{\mathfrak{S}_i\}$ of the operators (see 50) obtained from these factors by replacing x by τ. Moreover, each subspace $\mathfrak{S}_i$ is invariant under τ. In particular, if $\mathfrak{F}$ is algebraically closed, then $p_i(x) = x - c_i$ and $\mathfrak{S}_i$ is the nullspace of $(\tau - c_iI)^{e_i}$, in which $\{c_i\}$ are the distinct characteristic roots of τ, and e_i is the multiplicity with which c_i occurs as a root of $m(x)$. A basis (E) of $\mathfrak{U}$ can be chosen so that it is the ordered union of bases of the subspaces $\mathfrak{S}_i$. The matrix $A_\tau^{(E)}$ then takes the semi-diagonal form (see 51) in which, for each i, the matrix A_i on the diagonal of $A_\tau^{(E)}$ has order equal to the dimension of $\mathfrak{S}_i$ and has minimal polynomial $[p_i(x)]^{e_i}$.

The minimal polynomial

A linear operator τ on $\mathfrak{U}$ is said to be diagonalizable if its eigenvectors span the space $\mathfrak{U}$; that is, if it has n linearly independent eigenvectors, and in which $n = \dim \mathfrak{U}$. A necessary and sufficient condition that τ be diagonalizable is that the minimal polynomial of τ factors in $\mathfrak{F}$ into distinct linear factors (see 52). Diagonalizability of τ implies that the matrix $A_\tau = A_\tau^{(E)}$ of τ, relative to any basis (E), has n linearly independent eigenvectors $X_1, X_2, \cdots, X_n$. If X_i is the eigenvector corresponding to the eigenvalue c_i, then $A_\tau X_i = c_i X_i$, and hence, if P is the matrix with column vectors $X_1, X_2, \cdots, X_n$, then $A_\tau P = PD$ in which D is the diagonal matrix (see 53) with the eigenvalues c_i along the diagonal. Since $X_1, \cdots, X_n$ are linearly independent P is nonsingular, and hence $P^{-1}A_\tau P = D$, and A_τ is similar to a diagonal matrix. Conversely, the equation $P^{-1}A_\tau P = D$ implies that A_τ, and therefore τ, has n linearly independent eigenvalues.

Thus, diagonalizability of an operator means that its matrix, relative to any basis, is similar to a diagonal matrix, and that $\mathfrak{U}$ is the direct sum of n one-dimensional subspaces, each invariant under τ.

If τ is a linear operator in a nonzero n-dimensional space $\mathfrak{U}$ and V is a nonzero vector in $\mathfrak{U}$, then there is a unique monic polynomial $g(x)$ of minimum degree such that $g(\tau)V = O$, and $g(x)$ is called the τ-annihilator of V. If $g(x) = c_0 + c_1x + \cdots + c_{r-1}x^{s-1} + x^s$, then the vectors $V, \tau V, \tau^2 V, \cdots, \tau^{s-1}V$ are linearly independent and span a subspace $\mathfrak{C}(V, \tau)$, called the τ-cyclic subspace generated by V. The subspace $\mathfrak{C}(V, \tau)$ is invariant under τ and the linear operator induced by τ in $\mathfrak{C}(V, \tau)$ has a matrix relative to the basis $V, \tau V, \cdots, \tau^{s-1}V$ of a simple form (see 54) which is called the companion matrix of the polynomial $g(x)$. The minimum polynomial of S is $g(x)$. The rational decomposition theorem states that there exist r nonzero vectors $\{V_i\}$ in $\mathfrak{U}$ with τ-annihilators $\{g_i(x)\}$ such that $g_{i+1}(x)$ divides $g_i(x)$ for $i = 1, 2, \cdots, r-1$, and $\mathfrak{U}$ is the direct sum (see 55) of subspaces $\{\mathfrak{C}(V_i, \tau)\}$. Moreover, the number r and the τ-annihilators $\{g_i(x)\}$ are uniquely determined by these conditions and do not depend on the choice of the V_i. A basis (B) of $\mathfrak{U}$ can therefore be chosen (see 56) of a form such that relative to this basis the matrix of τ (see 57) will have the matrices $\{C_i\}$ along the diagonal and zeros elsewhere. Each C_i here is the companion matrix of $g_i(x)$. Such a matrix is said to be in rational canonical form. Thus every $n \times n$ matrix over $\mathfrak{F}$ is similar to a unique matrix in rational canonical form and two matrices are similar over $\mathfrak{F}$ if and only if they have the same rational canonical form.

The rational canonical form of a matrix

If the field $\mathfrak{F}$ is algebraically closed, the minimal polynomial (see 58) of τ is the product of factors of the type $(x - c_i)^{e_i}$, in which $\{c_i\}$ are the distinct eigenvalues of τ, and $\mathfrak{U}$ is the direct sum of the nullspaces $\mathfrak{S}_i$ of the operators $(\tau - c_iI)^{e_i}$. If τ_i is the operator induced in $\mathfrak{S}_i$ by τ, and if $\sigma_i = \tau_i - c_iI$, then $\tau_i = c_iI + \sigma_i$, in which σ_i has the minimal polynomial x^{e_i}. Thus, if a basis for $\mathfrak{S}_i$ is chosen in accordance with the rational decomposition of $\mathfrak{S}_i$ into cyclic subspaces relative to the operator σ_i, then the matrix J_i of τ_i relative to this basis has a form (see 59) with matrices J_k^i along the diagonal, zeros

(46) $(A_\tau^{(E)} - cI)X = O$

(47) $\mathfrak{U} = \mathfrak{S}_1 \oplus \mathfrak{S}_2 \oplus \cdots \oplus \mathfrak{S}_r$

(48) $V = S_1 + S_r + \cdots + S_r$

(49) $m(x) = [p_1(x)]^{e_1}[p_2(x)]^{e_2} \cdots [p_r(x)]^{e_r}$

(50) $[p_i(\tau)]^{e_i}$

(51) $$A_\tau^{(E)} = \begin{bmatrix} A_1 & O & \cdots & O \\ O & A_2 & & O \\ \cdots & \cdots & \cdots & \cdots \\ O & O & \cdots & A_r \end{bmatrix}$$

(52) $\begin{cases} m(x) = (x - c_1)(x - c_2) \cdots (x - c_r), \text{ in which} \\ c_1, c_2, \cdots, c_r \text{ are the distinct eigenvalues of } \tau. \end{cases}$

(53) $$\begin{bmatrix} c_1 & 0 & \cdots & 0 \\ 0 & c_2 & & \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & & c_n \end{bmatrix}$$

(54) $$C = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & -c_0 \\ 1 & 0 & 0 & \cdots & 0 & -c_1 \\ 0 & 1 & 0 & \cdots & 0 & -c_2 \\ & & & & & \\ 0 & 0 & 0 & \cdots & 1 & -c_{r-1} \end{bmatrix}$$

(55) $\mathfrak{U} = \mathfrak{C}(V_1, \tau) \oplus \mathfrak{C}(V_2, \tau) \oplus \cdots \oplus \mathfrak{C}(V_r, \tau)$

(56) $V_1, \tau V_1, \cdots, \tau^{s_1-1}V_1, \cdots, V_r, \tau V_r, \cdots, \tau^{s_r-1}V_r$

(57) $$A_\tau^{(B)} = \begin{bmatrix} C_1 & 0 & \cdots & 0 \\ 0 & C_2 & \cdots & 0 \\ 0 & 0 & \cdots & C_r \end{bmatrix}$$

(58) $m(x) = (x - c_1)^{e_1}(x - c_2)^{e_2} \cdots (x - c_r)^{e_r}$

(59) $$J_i = \begin{bmatrix} J_1^{(i)} & 0 & \cdots & 0 \\ 0 & J_2^{(i)} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & J_{n_i}^{(i)} \end{bmatrix}$$

elsewhere, and in which each matrix J_k^i has a simple triangular form (see 60) and the order of J_k^i is greater than or equal to that of J_{k+1}^i for each k. The union of these bases for the subspaces $\mathfrak{S}_i$ form a basis for $\mathfrak{U}$ relative to which the matrix of τ has the form (61) with the matrices J_i along the diagonal and zeros elsewhere. Such a matrix is said to be in Jordan canonical form. Every matrix A over an algebraically closed field is similar to a matrix in Jordan canonical form that is unique, if the diagonal blocks J_i are ordered so their eigenvalues increase with i. Two matrices are similar if and only if they have the same Jordan canonical form.

Linear functionals. If $\mathfrak{U}$ is a vector space over $\mathfrak{F}$, then a linear functional on $\mathfrak{U}$ is a mapping τ from $\mathfrak{U}$ into $\mathfrak{F}$ such that the linearity properties (see 62) are satisfied. If $\mathfrak{F}$ itself is considered as a vector space over $\mathfrak{F}$, with field addition as vector addition, and field multiplication as scalar multiplication, then a linear functional on $\mathfrak{U}$ is simply a linear transformation from $\mathfrak{U}$ into $\mathfrak{F}$. If $\mathfrak{U}$ has dimension n and a basis (E), examples (see 63) of linear functionals are easily constructed. In particular, the mapping of each vector V onto its ith coordinate, relative to the E-basis, is a linear functional. Other examples are the mapping σ that carries $\mathfrak{B}$ into $\mathfrak{F}$ defined by $\sigma f(x) = f(a)$ for a fixed scalar a, in which $\mathfrak{B}$ is the space of all polynomials in x with coefficients in $\mathfrak{F}$, and the mapping ϕ defined on the space of all continuous real-valued functions on the closed interval of real numbers between a and b in which ϕ is the integral (see 64) of f over that interval. The set of all linear functionals on $\mathfrak{U}$ can be given the algebraic structure of a vector space by identifying it with $\mathfrak{L}(\mathfrak{U}, \mathfrak{F})$, in which $\mathfrak{F}$ is viewed as a one-dimensional vector space over $\mathfrak{F}$. The vector space $\mathfrak{L}(\mathfrak{U}, \mathfrak{F})$ of linear functionals on $\mathfrak{U}$ is called the dual space of $\mathfrak{U}$ and is denoted by $\mathfrak{U}^*$. Since $\dim \mathfrak{F} = 1$, if $\dim \mathfrak{U} = n$, then $\mathfrak{U}^*$ also has dimension n. If (E) is a basis of $\mathfrak{U}$, since $F_1 = 1$ is a basis of $\mathfrak{F}$, there is a unique linear functional τ in $\mathfrak{U}^*$ such that, for each i, $\tau E_i = a_i$ in which the $\{a_i\}$ are n arbitrary scalars. The matrix (see 65) of τ relative to the E- and F-bases is the vector with components a_i. Thus every linear functional on a finite-dimensional space is defined by a fixed linear homogeneous function of the coordinates relative to the E-basis. Corresponding to an E-basis of $\mathfrak{U}$, there is a dual τ-basis of $\mathfrak{U}^*$, defined by $\tau_i E_i = 1$ and $\tau_i E_j = 0$ if $j \neq i$. Thus, τ_i is the functional that maps each vector V of $\mathfrak{U}$ on its ith coordinate relative to the E-basis. The matrix of τ_i, relative to the E-basis, is the vector with 1 in the ith place and 0 elsewhere, and every vector V of $\mathfrak{U}$ expressed in the E-basis has the ith coordinate $\tau_i(V)$ (see 66). The dual space $(\mathfrak{U}^*)^*$ of $\mathfrak{U}^*$ is denoted by $\mathfrak{U}^{**}$ and is called the double dual of $\mathfrak{U}$. There is a natural isomorphism α from $\mathfrak{U}$ onto $\mathfrak{U}^{**}$ defined by $\alpha V = \sigma_V$ in which σ_V is the linear functional on $\mathfrak{U}^*$, defined by $\sigma_V \tau = \tau V$ for all τ in $\mathfrak{U}^*$. The ismorphism α maps the E-basis of $\mathfrak{U}$ onto its double dual basis of $\mathfrak{U}^{**}$ (see 67).

Examples of linear functionals

If σ is a linear transformation of $\mathfrak{U}$ into $\mathfrak{W}$, and τ is a linear functional on $\mathfrak{W}$, then the composite $\tau\sigma$ is a linear functional on $\mathfrak{U}$. Thus, σ induces a linear transformation of σ^T of $\mathfrak{W}^*$ into $\mathfrak{U}^*$, defined by $\sigma^T\tau = \tau\sigma$, and called the transpose of σ. If (E) and (F) are bases of $\mathfrak{U}$ and $\mathfrak{W}$, and (α) and (β) are the dual bases of $\mathfrak{U}^*$ and $\mathfrak{W}^*$, then the matrix of σ^T, relative to the bases (β) and (α), is the transpose of the matrix $A_\sigma{}^{(E,F)}$.

Inner products. If X and Y are any two vectors in $\mathfrak{U}_3(\mathfrak{R})$, then their classical inner product or dot product (see 68) is a scalar-valued function of the pair of vectors X, Y. Its importance lies in the fact that the length of the vector X, denoted by $||X||$, and the angle θ between the directions of X and Y, can be defined in terms of inner products (see 69). Similarly, if X and Y are two vectors of $\mathfrak{U}_n(\mathfrak{R})$, their dot product or standard inner product is defined as a sum (see 70) of products of respective components. Because $(X, X) > 0$ if $X \neq O$, and because the Cauchy-Schwarz inequality ensures that $|(X, Y)| \leqslant \sqrt{(X, X)}\,\sqrt{(Y, Y)}$, the length of a vector and the angle between vectors can be defined by the same formulas as those used in $\mathfrak{U}_3$. In general, an inner product is used to introduce metric concepts in an abstract vector space over either the real or the complex field.

Inner products and metric concepts

An important property of the dot product in $\mathfrak{U}_n(\mathfrak{R})$ is that the mappings defined by $\sigma_X(Y) = (X, Y)$ for a fixed X, and $\sigma_Y(X) = (X, Y)$ for a fixed Y, are both linear functionals on $\mathfrak{U}_n(\mathfrak{R})$. The dot product is therefore an example of a bilinear functional. In general, if $\mathfrak{U} \times \mathfrak{U}$ denotes the set of all ordered pairs (V, W) in which V, W are vectors in a vector space $\mathfrak{U}$ over $\mathfrak{F}$, a bilinear functional on $\mathfrak{U}$ is a mapping ϕ from $\mathfrak{U} \times \mathfrak{U}$ into $\mathfrak{F}$ having linearity properties (see 71) with respect to both members of the pair. A bilinear functional ϕ is symmetric if $\phi(V, W) = \phi(W, V)$, and is positive if $\phi(V, V)$ is a real number for every vector V and $\phi(V, V) > 0$ whenever $V \neq O$.

If $\mathfrak{U}$ is any vector space over $\mathfrak{R}$, an inner product in $\mathfrak{U}$ is a positive symmetric bilinear functional on $\mathfrak{U}$. The dot product in $\mathfrak{U}_n(\mathfrak{R})$ is clearly an inner product under this definition. More generally, if (E) is any basis of $\mathfrak{U}$, an inner product in $\mathfrak{U}$ is obtained by defining (V, W) to be the dot product of the coordinate vectors of V and W, relative to the E-basis. A vector space over $\mathfrak{R}$ in which a fixed inner product is designated is called a real inner product space. A finite-dimensional real inner product

(60)
$$\begin{bmatrix} c_i & 0 & 0 & \cdots & 0 & 0 \\ 1 & c_i & 0 & \cdots & 0 & 0 \\ 0 & 1 & c_i & \cdots & 0 & 0 \\ & & & & & \\ 0 & 0 & 0 & \cdots & c_i & 0 \\ 0 & 0 & 0 & \cdots & 1 & c_i \end{bmatrix}$$

(61)
$$A_\tau = \begin{bmatrix} J_1 & 0 & 0 \\ 0 & J_2 & 0 \\ \cdots & \cdots & \cdots \\ 0 & 0 & J_r \end{bmatrix}$$

(62) $\tau(V_1 + V_2) = \tau V_1 + \tau V_2$ and $\tau(cV) = c(\tau V)$

(63) $\tau(x_1E_1 + x_2E_2 + \cdots + x_nE_n) = a_1x_1 + a_2x_2 + \cdots + a_nx_n$ in which $a_1, a_2, \cdots a_n$ are any fixed scalars

(64) $\phi f = \int_a^b f(x)\,dx$

(65) $A_\tau^{(E,F)} = [a_1, a_2, \cdots, a_n]$

(66) $V = \tau_1(V)E_1 + \tau_2(V)E_2 + \cdots + \tau_n(V)E_n$

(67) If $\tau_1, \cdots, \tau_n$ is the basis of $\mathfrak{U}^*$ dual to the basis $E_1, \cdots, E_n$ of $\mathfrak{U}$ then the basis of $\mathfrak{U}^{**}$ dual to $\tau_1, \cdots, \tau_n$ is $\alpha E_1, \cdots, \alpha E_n$

(68) $(X, Y) = x_1y_1 + x_2y_2 + x_3y_3$

(69) $\|X\| = \sqrt{(X, X)}$ $\qquad \cos\theta = \dfrac{(X, Y)}{\|X\|\,\|Y\|}$

(70) If $X = [x_1, \cdots, x_n]$ and $Y = [y_1, \cdots, y_n]$, $(X, Y) = x_1y_1 + x_2y_2 + \cdots + x_ny_n$

(71)
$$\begin{cases} \phi(V_1 + V_2, W) = \phi(V_1, W) + \phi(V_2, W) \\ \phi(V, W_1 + W_2) = \phi(V, W_1) + \phi(V, W_2) \\ \phi(cV, W) = \phi(V, cW) = c\phi(V, W) \end{cases}$$

space is called a Euclidean vector space. Inner products that lead to a suitable metric can also be defined in a vector space over the complex field $\mathfrak{C}$, but the linearity and symmetry conditions have to be modified. If $\mathfrak{U}$ is a vector space over $\mathfrak{C}$, an inner product on $\mathfrak{U}$ is a function that associates with each ordered pair of vectors V, W in $\mathfrak{U}$ a scalar (V, W) such that (see 72) (V, W) is linear in the first member of the pair, (V, W) is equal to the complex conjugate of (W, V), and (V, V) is real and is positive if $V \neq O$.

These postulates can be used to define an inner product on either a real or complex vector space. In the complex case σ_W, defined by $\sigma_W V = (V, W)$, is a linear functional on $\mathfrak{U}$ into $\mathfrak{C}$ and, in fact, every linear functional of $\mathfrak{U}$ has this form for some vector W. The behaviour of σ_V, defined by $\sigma_V W = (V, W)$, is, however, not linear because the functional value of σ_V at cW is c times the value at W. An example of an inner product (X, Y) in the complex case is the standard inner product of $\mathfrak{U}_n(\mathfrak{C})$, defined by the sum (see 73) of products of the type $x_i\bar{y}_i$. In general, if $\mathfrak{F}$ is either $\mathfrak{R}$ or $\mathfrak{C}$, and α is the isomorphism that maps each vector V on its coordinate vector αV in $\mathfrak{U}_n(\mathfrak{F})$, then $(V, W)' = (\alpha V, \alpha W)$, in which $(\ ,\ )$ denotes the standard inner product in $\mathfrak{U}_n(\mathfrak{F})$, defines an inner product $(V, W)'$ in $\mathfrak{U}$.

If $\mathfrak{F}$ denotes either the real field $\mathfrak{R}$ or the complex field $\mathfrak{C}$, and $\mathfrak{U}$ is a vector space over $\mathfrak{F}$ with basis (E), then an inner product (V, W) in $\mathfrak{U}$ is completely determined by the n^2 scalars (E_i, E_j). If these scalars are denoted by a_{ij}, then the matrix $A = [a_{ij}]$ has the property $A^* = A$, in which A^* is the transpose of the complex conjugate matrix $\bar{A} = [\bar{a}_{ij}]$. The matrix A^* is called the conjugate of A, and a matrix A for which $A^* = A$ is called a self-conjugate or Hermitian matrix, named for the French mathematician Charles Hermite. If A has real elements, then $A^* = A^T$, and hence a real Hermitian matrix is symmetric in the sense that $A^T = A$. An expression formed from a sum of terms of the type $a_{ij}x_i\bar{y}_j$ in which $a_{ij} = \bar{a}_{ji}$ naturally occurs when taking the inner product of two vectors V and W with coordinate vectors X and Y (see 74). Such an expression (see 75) is called a Hermitian bilinear form. It can be written in matrix notation as XAY^*, in which X and Y are the coordinate vectors of V and W, and $Y^* = \bar{Y}^T$. A Hermitian bilinear form has the property that XAX^* is real for all X. If in addition $XAX^* > 0$ for all nonzero X, then the Hermitian form and the Hermitian matrix A are said to be positive. Every inner product in $\mathfrak{U}$ has the form $(V, W) = XAY^*$ in which X, Y are the coordinate vectors of V, W relative to some basis, and A is a positive Hermitian matrix. Conversely, every positive Hermitian bilinear form in the coordinates of V and W defines an inner product on $\mathfrak{U}$.

Hermitian matrices

If $\mathfrak{U}$ is an inner product space over $\mathfrak{R}$ or $\mathfrak{C}$, then for any two vectors V, W the Cauchy-Schwarz inequality holds (see 76). The length of V is defined by $||V|| = (V, V)^{1/2}$ and has the three properties (see 77) that are normally associated with length. If $||V|| = 1$, then V is called a unit vector. Two vectors V, W are said to be orthogonal if $(V, W) = 0$. The zero vector is orthogonal to every vector in $\mathfrak{U}$ and is the only vector with this property. Every n-dimensional inner product space has an orthonormal basis; that is, a basis consisting of n mutually orthogonal unit vectors. If (E) is an orthonormal basis of $\mathfrak{U}$, a second basis (F) is orthonormal if, and only if, the transition matrix P from the E- to the F-basis has the property $PP^* = I$. Such a matrix P is called a unitary matrix. The columns (and the rows) of a unitary matrix constitute an orthonormal basis of $\mathfrak{U}_n(\mathfrak{C})$ relative to the standard inner product. If a unitary matrix P has real elements, then $PP^T = I$ and P is called an orthogonal matrix and its columns (and rows) form an orthonormal basis of $\mathfrak{U}_n(\mathfrak{R})$. If $\mathfrak{U}$ is an inner product space with orthonormal basis $E_1, E_2, \cdots, E_n$ and α is the isomorphism that maps each vector V on its coordinate vector αV relative to the E-basis, then the inner product (V, W) in $\mathfrak{U}$ is equal to the standard inner product $(\alpha V, \alpha W)$ of the coordinate vectors in $\mathfrak{U}_n(\mathfrak{F})$. Hence the standard inner product in $\mathfrak{U}_n(\mathfrak{F})$ is invariant under transformation from one orthonormal basis to another.

Linear operators in an inner product space. If $\mathfrak{U}$ is a finite-dimensional inner product space over $\mathfrak{F}$, in which $\mathfrak{F}$ is either $\mathfrak{R}$ or $\mathfrak{C}$, then for any linear operator τ on $\mathfrak{U}$ these exists a unique linear operator τ^* on $\mathfrak{U}$, called the adjoint of τ, such that $(\tau V, W) = (V, \tau^* W)$ for all vectors V, W in $\mathfrak{U}$. If A_τ is the matrix of τ relative to an orthonormal basis, then the matrix A_{τ^*} of τ^*, relative to the same basis, is the adjoint matrix $A_\tau{}^*$ of A_τ. If σ, τ are linear operators on $\mathfrak{U}$, and if c is any scalar, then $(\sigma + \tau)^* = \sigma^* + \tau^*$, $(c\sigma)^* = c\sigma^*$, $(\sigma\tau)^* = \tau^*\sigma^*$, and $(\sigma^*)^* = \sigma$. An operator τ is self-adjoint, or Hermitian, if $\tau^* = \tau$. If τ is Hermitian, then $(\tau V, V) = (V, \tau V) = \overline{(\tau V, V)}$, and hence $(\tau V, V)$ is real for all V. A Hermitian operator τ is positive if $(\tau V, V) > 0$ when $V \neq O$. If τ is a linear operator in an inner product space with inner product $(V, W)_1$, then the function $(V, W)_2 = (\tau V, W)_1$ is an inner product in $\mathfrak{U}$ if, and only if, τ is positive Hermitian. An operator is positive Hermitian if, and only if, its matrix, relative to an orthonormal basis, is positive Hermitian. A Hermitian operator on an n-dimensional inner product space $\mathfrak{U}$ has n real eigenvalues and n mutually orthogonal eigenvectors. Thus, if τ is Hermitian, it is diagonalizable, and $\mathfrak{U}$ is the direct sum of mutually orthogonal one-dimensional subspaces invariant under τ. Applied to the operator τ in $\mathfrak{U}_n(\mathfrak{C})$, defined by $\tau X = HX$ in which H is a Hermitian matrix, this implies that H has n mutually orthogonal unit eigenvectors in $\mathfrak{U}_n(\mathfrak{C})$ and, hence, that there exists a unitary matrix U such that $U^{-1}HU = U^*HU$ is a real diagonal matrix whose diagonal elements are the eigenvalues of H. Similarly, in the real case, if A is a real symmetric matrix, there exists a real orthogonal matrix P such that $P^{-1}AP$ is diagonal, and A has n mutually orthogonal eigenvectors in $\mathfrak{U}_n(\mathfrak{R})$. Using this diagonalization procedure, any Hermitian form XHX^* can be reduced by a unitary transformation $X^T = UY^T$ to a canonical form (see 78) in which the coefficients c_i are the eigenvalues of H. It follows that H is positive if, and only if, all its eigenvalues are positive, and the same is true of a positive Hermitian operator.

The adjoint of a linear operator

A unitary operator is a linear operator τ on $\mathfrak{U}$ that preserves lengths, that is, such that for every vector V in $\mathfrak{U}$, $||\tau V|| = ||V||$. A necessary and sufficient condition that τ preserves lengths is that for any two vectors V, W in $\mathfrak{U}$, $(\tau V, \tau W) = (V, W)$. Hence, unitary operators preserve orthogonality as well as lengths. Since $(\tau V, \tau W) =$

(72) (1) $(V_1 + V_2, W) = (V_1, W) + (V_2, W)$, (2) $(W, V) = \overline{(V, W)}$, in which $\bar{z}$ denotes the complex conjugate of z, (3) $(cV, W) = c(V, W)$, and (4) (V, V) is real and $(V, V) > 0$ if $V \neq 0$.

(73) $(X, Y) = x_1\bar{y}_1 + x_2\bar{y}_2 + \cdots + x_n\bar{y}_n$

(74) If $V = x_1E_1 + \cdots + x_nE_n$, $W = y_1E_1 + \cdots + y_nE_n$ the postulates require that
$$(V, W) = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i\bar{y}_j(E_i, E_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i\bar{y}_j,$$
in which $A = [a_{ij}]$ is Hermitian.

(75) $\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i\bar{y}_j$ in which $a_{ij} = \bar{a}_{ji}$

(76) $|(V, W)| \leq (V, V)^{1/2}(W, W)^{1/2}$

(77) (a) $\|V\| > 0$ if $V \neq 0$, (b) $|cV| = |c|\,\|V\|$
(c) $\|V + W\| \leq \|V\| + \|W\|$

(78) $c_1y_1\bar{y}_1 + c_2y_2\bar{y}_2 + \cdots + c_ny_n\bar{y}_n$

$(V, \tau^*\tau W)$ it follows that τ is unitary if, and only if, $\tau^*\tau = I$. Thus a unitary operator τ is invertible, with inverse $\tau^{-1} = \tau^*$, and τ^{-1} is also unitary. If σ and τ are unitary, the same is true also of $\sigma\tau$. Thus, the unitary operators on $\mathfrak{U}$ form a group called the unitary group. An operator τ is unitary if, and only if, its matrix A_τ, relative to an orthonormal basis, is unitary. The unitary matrices over $\mathfrak{C}$ also form a group under matrix multiplication, and the mapping of τ on A_τ is an isomorphism of the group of unitary operators onto the group of unitary matrices. A unitary operator on a real vector space is called an orthogonal operator. Its matrix relative to an orthonormal basis is orthogonal, and the mapping of τ on A_τ is an isomorphism of the (real) orthogonal group of orthogonal operators onto the group of real orthogonal matrices. Every nonsingular linear operator τ on $\mathfrak{U}$ has a unique polar factorization, $\tau = \rho\sigma$, into the product of a positive Hermitian operator ρ and a unitary operator σ. If τ is singular, it can still be factored in this way, but the Hermitian factor ρ is merely nonnegative (that is, has nonnegative eigenvalues) and the unitary factor σ is not uniquely determined.

Orthogonal operators.

An operator τ on $\mathfrak{U}$ is normal if $\tau\tau^* = \tau^*\tau$. Hermitian, unitary, and real orthogonal operators are all normal. In general, an operator is normal if and only if it has n mutually orthogonal eigenvectors with $n = \dim \mathfrak{U}$. Thus, a normal operator is diagonalizable; its matrix, relative to the orthonormal basis of $\mathfrak{U}$, consisting of n mutually orthogonal unit eigenvectors of τ, is a diagonal matrix with the eigenvalues of τ as diagonal elements. An $n \times n$ matrix A over $\mathfrak{C}$ is normal if $AA^* = A^*A$, and an operator is normal if and only if its matrix relative to an orthonormal basis is normal. Just as in the Hermitian case, if A is normal, then there exists a unitary matrix U such that $U^{-1}AU$ is a diagonal matrix. Conversely, every matrix that is unitarily similar to a diagonal matrix, is normal.

If $\mathfrak{U}$ is finite-dimensional, and if $\mathfrak{S}$ is any subspace of $\mathfrak{U}$, then the set $\mathfrak{S}^\perp$ of all vectors of T in $\mathfrak{U}$ such that $(T, S) = 0$ for all S in $\mathfrak{S}$ is a subspace of $\mathfrak{U}$ called the orthogonal complement of $\mathfrak{S}$. For any subspace $\mathfrak{S}$ of $\mathfrak{U}$, $(\mathfrak{S}^\perp)^\perp = \mathfrak{S}$, and $V = \mathfrak{S} \oplus \mathfrak{S}^\perp$. Thus every vector V has a unique representation $V = S + T$, in which S is in $\mathfrak{S}$, and T is in $\mathfrak{S}^\perp$. The mapping π from $\mathfrak{U}$ onto $\mathfrak{S}$ defined by $\pi V = S$ is a linear operator called the orthogonal projection of $\mathfrak{U}$ on $\mathfrak{S}$. Since $\pi V = 0$ if and only if V is in $\mathfrak{S}^\perp$, the nullspace of π is $\mathfrak{S}^\perp$.

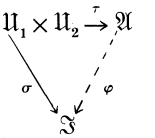
If τ is a normal operator on an n-dimensional inner product space $\mathfrak{U}$, $c_1, \cdots, c_r$ are the distinct eigenvalues of τ, and π_i is the orthogonal projection of $\mathfrak{U}$ onto the nullspace of $\tau - c_i I$, then a basic result (see 79) expresses τ as a linear combination of the orthogonal projection π. The result is called the spectral theorem for normal operators.

(79)
$$\begin{cases} \tau = c_1\pi_1 + c_2\pi_2 + \cdots + c_r\pi_r, \qquad \pi_i\pi_j = 0 \quad \text{if } i \neq j \\ \text{and} \quad \pi_1 + \pi_2 + \cdots + \pi_r = I \end{cases}$$

(80)
$$\begin{cases} \tau(V_1, \cdots, V_i + V_i', \cdots, V_r) \\ \quad = \tau(V_1, \cdots, V_i, \cdots V_r) + \tau(V_1, \cdots, V_i', \cdots, V_r) \\ \quad \tau(V_1, \cdots, cV_i, \cdots V_r) = c\tau(V_1, \cdots, V_i, \cdots, V_r). \end{cases}$$

(81)
$$\begin{cases} (x_1V_1 + x_2V_1', V_2) - x_1(V_1, V_2) - x_2(V_1', V_2) \\ \text{or} \quad (V_1, y_1V_2 + y_2V_2') - y_1(V_1, V_2) - y_2(V_1, V_2'). \end{cases}$$

(82)
$$\begin{cases} V_1 \otimes (V_2 + V_2') = V_1 \otimes V_2 + V_1 \otimes V_2' \\ (V_1 + V_1') \otimes V_2 = V_1 \otimes V_2 + V_1' \otimes V_2 \\ (cV_1) \otimes V_2 = V_1 \otimes (cV_2) = c(V_1 \otimes V_2) \end{cases}$$

(83)


MULTILINEAR ALGEBRA

The tensor product of vector spaces. If $\mathfrak{U}_1, \mathfrak{U}_2, \cdots, \mathfrak{U}_r$ are vector spaces over a field $\mathfrak{F}$, then the set of all r-tuples $(V_1, V_2, \cdots, V_r)$ in which V_i belongs to $\mathfrak{U}_i$, is called the Cartesian product of $\mathfrak{U}_1, \cdots, \mathfrak{U}_r$ and is denoted by $\mathfrak{U}_1 \times \mathfrak{U}_2 \times \cdots \times \mathfrak{U}_r$. The Cartesian product of r copies of a vector space $\mathfrak{U}$ is called the rth Cartesian power of $\mathfrak{U}$ and is denoted by $\times^r\mathfrak{U}$. If $\mathfrak{W}$ is also a vector space over $\mathfrak{F}$, a mapping τ from $\mathfrak{U}_1 \times \mathfrak{U}_2 \times \cdots \times \mathfrak{U}_r$ into $\mathfrak{W}$ is multilinear if for each value of i from 1 to r linearity holds (see 80) for the ith variable. A multilinear mapping from $\times^r\mathfrak{U}$ to $\mathfrak{W}$ is called an r-linear mapping. If $\mathfrak{W}$ is the scalar field $\mathfrak{F}$ considered as a vector space, then an r-linear mapping τ from $\times^r\mathfrak{U}$ into $\mathfrak{F}$ is called an r-linear functional on $\mathfrak{U}$.

If $\mathfrak{U}_1, \mathfrak{U}_2$ are vector spaces over $\mathfrak{F}$, and if $\mathfrak{M}$ is the free vector space over $\mathfrak{F}$ on the basis set $\mathfrak{U}_1 \times \mathfrak{U}_2$, then one may let $\mathfrak{N}$ be the subspace of $\mathfrak{M}$ spanned by the set of all elements having one of the two forms (81). The factor space $\mathfrak{M}/\mathfrak{N}$ consists of elements $Z + \mathfrak{N}$, Z in $\mathfrak{M}$, and is a vector space over $\mathfrak{F}$ when addition and scalar multiplication are defined by $(Z_1 + \mathfrak{N}) + (Z_2 + \mathfrak{N}) = (Z_1 + Z_2) + \mathfrak{N}$ and $c(Z + \mathfrak{N}) = cZ + \mathfrak{N}$, respectively. The mapping π from $\mathfrak{M}$ onto $\mathfrak{M}/\mathfrak{N}$ defined by $\pi Z = Z + \mathfrak{N}$, is a homomorphism called the canonical projection of $\mathfrak{M}$ onto $\mathfrak{M}/\mathfrak{N}$. The space $\mathfrak{M}/\mathfrak{N}$ is called the tensor product of $\mathfrak{U}_1$ and $\mathfrak{U}_2$ and is denoted by $\mathfrak{U}_1 \otimes \mathfrak{U}_2$. If π is the canonical projection, $\pi(V_1, V_2)$ is written $V_1 \otimes V_2$ (read V_1 tensor V_2). The elements $V_1 \otimes V_2$, with V_1 in $\mathfrak{U}_1$, and $\mathfrak{U}_2$, span $\mathfrak{U}_1 \otimes \mathfrak{U}_2$, and the operation $\otimes$ satisfies three basic relations (see 82). Thus the mapping τ, defined by $\tau(V_1, V_2) = V_1 \otimes V_2$, is bilinear.

The tensor product

The tensor product $\mathfrak{U}_1 \otimes \mathfrak{U}_2$, together with the bilinear mapping τ from $\mathfrak{U}_1 \times \mathfrak{U}_2$ into $\mathfrak{U}_1 \otimes \mathfrak{U}_2$, has the following "universal property." If σ is any bilinear mapping of $\mathfrak{U}_1 \times \mathfrak{U}_2$ into an arbitrary vector space $\mathfrak{F}$ over $\mathfrak{F}$, then there exists a unique linear transformation ϕ from $\mathfrak{U}_1 \otimes \mathfrak{U}_2$ into $\mathfrak{F}$, such that $\phi\tau = \sigma$. Most modern writers use this universal property to define the tensor product. From this point of view, and if $\mathfrak{A}$, $\mathfrak{U}_1$ and $\mathfrak{U}_2$ are vector spaces over $\mathfrak{F}$, and if τ is a bilinear mapping from $\mathfrak{U}_1 \times \mathfrak{U}_2$ into $\mathfrak{A}$, then the pair $(\mathfrak{A}, \tau)$ is a tensor product of the vector spaces $\mathfrak{U}_1$, $\mathfrak{U}_2$, if for every bilinear mapping σ from $\mathfrak{U}_1 \times \mathfrak{U}_2$ into $\mathfrak{F}$, there exists a unique linear transformation ϕ from $\mathfrak{A}$ into an arbitrary vector space $\mathfrak{F}$ over $\mathfrak{F}$, such that $\phi\tau = \sigma$ (see 83).

The existence of a tensor product, so defined, is proved by showing that $(\mathfrak{M}/\mathfrak{N}, \pi)$ has the required universal property. Moreover, the tensor product $(\mathfrak{A}, \tau)$ is unique in the sense that if $(\mathfrak{A}', \tau')$ is a second tensor product, then there exists an isomorphism α from $\mathfrak{A}$ onto $\mathfrak{A}'$ such that $\alpha\tau = \tau'$. Again, $\tau(V_1, V_2)$ is written $V_1 \otimes V_2$ and $(\mathfrak{A}, \tau)$ is written $\mathfrak{U}_1 \otimes \mathfrak{U}_2$. The tensor product $\mathfrak{U}_1 \otimes \mathfrak{U}_2$ is spanned by the elements $V_1 \otimes V_2$, with V_1 in $\mathfrak{U}_1$, and V_2 in $\mathfrak{U}_2$; and $V_1 \otimes V_2 = O$ if and only if $V_1 = O$ or $V_2 = O$. If $\mathfrak{U}_1 = \mathfrak{S}_1 \oplus \mathfrak{S}_2 \oplus \cdots \oplus \mathfrak{S}_n$ and $\mathfrak{U}_2 = \mathfrak{A}_1 \oplus \mathfrak{A}_2 \oplus \cdots \oplus \mathfrak{A}_m$, then $\mathfrak{U}_1 \otimes \mathfrak{U}_2$ is the direct sum of the nm spaces $\mathfrak{S}_i \otimes \mathfrak{A}_j$. If (E) and (F) are bases of $\mathfrak{U}_1$ and $\mathfrak{U}_2$, then the elements $\{E_i \otimes F_j\}$ form a basis of $\mathfrak{U}_1 \otimes \mathfrak{U}_2$. Thus $\dim(\mathfrak{U}_1 \otimes \mathfrak{U}_2) = (\dim \mathfrak{U}_1)(\dim \mathfrak{U}_2)$ if $\mathfrak{U}_1$ and $\mathfrak{U}_2$ are finite-dimensional. If τ and σ are linear transformations from the vector spaces $\mathfrak{U}_1$ and $\mathfrak{W}_1$ to vector spaces $\mathfrak{U}_2$, $\mathfrak{W}_2$, respectively, then the linear transformation ρ, defined by $\rho(V_1 \otimes W_1) = \tau V_1 \otimes \sigma W_1$, is called the tensor product of τ and σ and is denoted by $\tau \otimes \sigma$. If τ' and σ' are also linear transformations defined on the vector spaces $\mathfrak{U}_2$ and $\mathfrak{W}_2$, the rule for composition is $(\tau' \otimes \sigma')(\tau \otimes \sigma) = \tau'\tau \otimes \sigma'\sigma$, and $1 \otimes 1$ is the identity on $\mathfrak{U}_1 \otimes \mathfrak{W}_1$. The image of the tensor product of two linear transformations is the tensor product of the images. The kernel (nullspace) of a tensor product of linear transformation is given by $\ker(\tau \otimes \sigma) = \ker \tau \otimes \mathfrak{W}_1 + \mathfrak{U}_1 \otimes \ker \sigma$. For a fixed space $\mathfrak{W}$ a linear transformation σ from $\mathfrak{U}_1$ into $\mathfrak{U}_2$ induces a linear transformation σ^* from $\mathfrak{U}_1 \otimes \mathfrak{W}$ into $\mathfrak{U}_2 \otimes \mathfrak{W}$ defined by $\sigma^* = \sigma \otimes 1$; for fixed $\mathfrak{W}$ the map-

pings $f(\mathfrak{U}) = \mathfrak{U} \otimes \mathfrak{W}$, $f(\sigma) = \sigma^*$ define a covariant functor from the category of vector spaces over $\mathfrak{F}$ to itself. Similarly for a fixed space $\mathfrak{U}$, the mappings $\mathfrak{W} \to \mathfrak{U} \otimes \mathfrak{W}$ and $\tau \to 1 \otimes \tau$ define a covariant functor.

The tensor product of r vector spaces $\mathfrak{U}_1, \mathfrak{U}_2, \cdots, \mathfrak{U}_r$ over $\mathfrak{F}$ is defined similarly. If τ is a multilinear mapping of the set $\mathfrak{U}_1 \times \mathfrak{U}_2 \times \cdots \times \mathfrak{U}_r$ into a vector space $\mathfrak{A}$, then the pair $(\mathfrak{A}, \tau)$ is a tensor product of $\mathfrak{U}_1, \cdots, \mathfrak{U}_r$ if, for every multilinear mapping σ from $\mathfrak{U}_1 \times \mathfrak{U}_2 \times \cdots \times \mathfrak{U}_r$ into an arbitrary vector space $\mathfrak{T}$ over $\mathfrak{F}$, there exists a unique linear transformation ϕ of $\mathfrak{A}$ into $\mathfrak{T}$ such that $\phi\tau = \sigma$. As before $\tau(V_1, V_2, \cdots, V_r)$ is written as $V_1 \otimes V_2 \otimes \cdots \otimes V_r$ and $(\mathfrak{A}, \tau)$ is denoted by $\mathfrak{U}_1 \otimes \mathfrak{U}_2 \otimes \cdots \otimes \mathfrak{U}_r$. Existence and uniqueness of $\mathfrak{A}$ follow as in the case $r = 2$. There is an isomorphism β of $\mathfrak{U}_1 \otimes \mathfrak{U}_2 \otimes \mathfrak{U}_3$ onto $(\mathfrak{U}_1 \otimes \mathfrak{U}_2) \otimes \mathfrak{U}_3$, such that $\beta(V_1 \otimes V_2 \otimes V_3) = (V_1 \otimes V_2) \otimes V_3$ and similarly for $\mathfrak{U}_1 \otimes (\mathfrak{U}_2 \otimes \mathfrak{U}_3)$. This associativity principle extends to the tensor product of any finite number of vector spaces over $\mathfrak{F}$ (see 84).

The tensor algebra of a vector space. If $\mathfrak{U}$ is a vector space over $\mathfrak{F}$, then the tensor product $\mathfrak{U} \otimes \mathfrak{U} \otimes \cdots \otimes \mathfrak{U}$ of r copies of $\mathfrak{U}$ is called the rth tensorial power of $\mathfrak{U}$ and is denoted by $\otimes^r\mathfrak{U}$ or by $\mathfrak{A}_r$. Its elements are called tensors of order r. If $E_1, \cdots, E_n$ is a basis of $\mathfrak{U}$, then $\{E_{i_1} \otimes E_{i_2} \otimes \cdots \otimes E_{i_r}\}$ is a basis for $\otimes^r\mathfrak{U}$ and every tensor T of order r can be written in summation form (see 85). The n^r scalars $x_{i_1 \cdots i_r}$ that appear in this form are called the components of T relative to the E-basis of $\mathfrak{U}$.

Tensorial powers

As a notational convenience, $\mathfrak{A}_1$ is identified with $\mathfrak{U}$ and $\mathfrak{A}_0$ with the scalar field $\mathfrak{F}$. The direct sum $\mathfrak{A}(\mathfrak{U}) = \oplus \sum_{r=0}^{\infty} \mathfrak{A}_r$ is a vector space over $\mathfrak{F}$ and each tensorial power $\mathfrak{A}_r$ can be considered a subspace of $\mathfrak{A}(\mathfrak{U})$. Multiplication can be defined formally for the generating elements of $\mathfrak{A}(\mathfrak{U})$ (see 86) and can be extended linearly to arbitrary elements. With this definition of multiplication $\mathfrak{A}(\mathfrak{U})$ becomes an associative algebra over $\mathfrak{F}$, with unity element 1, and is called the tensor algebra of the vector space $\mathfrak{U}$.

The tensor algebra has the following universal mapping property: If τ is the canonical injection of $\mathfrak{U}$ into $\mathfrak{A}(\mathfrak{U})$, and if σ is a linear transformation of $\mathfrak{U}$ into an arbitrary associative algebra $\mathbf{A}$ over $\mathfrak{F}$ with unity, then there exists a unique algebra homomorphism ϕ of $\mathfrak{A}(\mathfrak{U})$ into $\mathbf{A}$ such that $\phi\tau = \sigma$. This property can be used as the definition of $\mathfrak{A}(\mathfrak{U})$, in which case existence is implied by the construction already given, and uniqueness follows in the same sense as the uniqueness of the tensor product. The universal property of $\mathfrak{A}(\mathfrak{U})$ also implies that every linear transformation σ of $\mathfrak{U}$ into a second vector space $\mathfrak{W}$ over $\mathfrak{F}$ induces a unique algebra homomorphism $T(\sigma)$ of $\mathfrak{A}(\mathfrak{U})$ into $\mathfrak{A}(\mathfrak{W})$ such that $T(\sigma)1 = 1$. Moreover, $T(I) = I$ and if ρ from $\mathfrak{W}$ into $\mathfrak{Y}$ is a second linear transformation $T(\phi\sigma) = T(\phi)T(\sigma)$. Thus the mappings $\mathfrak{U} \to \mathfrak{A}(\mathfrak{U})$ and $\sigma \to T(\sigma)$ define a covariant functor, from the category of vector spaces over $\mathfrak{F}$ to the category of associative algebras with unit over $\mathfrak{F}$.

If $\mathfrak{U}^*$ is the dual space of $\mathfrak{U}$, the tensor product $\mathfrak{U} \otimes \cdots \otimes \mathfrak{U}^* \otimes \cdots \otimes \mathfrak{U}^*$ of r copies of $\mathfrak{U}$ and s copies of $\mathfrak{U}^*$ is denoted by $\otimes_s{}^r(\mathfrak{U})$. If $s = 0$, $\otimes_0{}^r\mathfrak{U}$ is identified with $\otimes^r\mathfrak{U}$ and its elements are called contravariant tensors of order r. If $r = 0$, then $\otimes_s{}^0\mathfrak{U} = \otimes^s\mathfrak{U}^*$ and its elements are called covariant tensors of order s. Elements of $\otimes_s{}^r\mathfrak{U}$ are called mixed tensors of contravariant order r and covariant order s. Finally, $\otimes_0{}^1\mathfrak{U}$ is identified with $\mathfrak{U}$, $\otimes_1{}^0\mathfrak{U}$ with $\mathfrak{U}^*$ and $\otimes_0{}^0\mathfrak{U}$ with the scalar field $\mathfrak{F}$. If $E_1, E_2, \cdots, E_n$ is a basis of $\mathfrak{U}$ and $E^1, E^2, \cdots, E^n$ is the dual basis of $\mathfrak{U}^*$, then n^{r+s} elements (87) form a basis for $\otimes_s{}^r\mathfrak{U}$. Every tensor T of $\otimes_s{}^r\mathfrak{U}$ can be written in terms of these (see 88). The n^{r+s} scalars $x^{i_1 \cdots i_r}_{j_1 \cdots j_s}$ are called the components of T relative to the E-basis of $\mathfrak{U}$. By a process analogous to the construction of the tensor algebra $\mathfrak{A}(\mathfrak{U})$, the mixed tensors of all orders can be embedded in a mixed tensor algebra $\mathfrak{A}(\mathfrak{U}, \mathfrak{U}^*)$.

Mixed tensors

If $\mathfrak{U}$ is an inner product space with inner product $(\ ,\ )$, then an inner product on $\otimes^r\mathfrak{U}$ can be defined (see 89), and $\otimes^r\mathfrak{U}$ is then an inner product space. If (E) is an orthonormal basis for $\mathfrak{U}$, then $\{E_{i_1} \otimes E_{i_2} \otimes \cdots \otimes E_{i_n}\}$ is an orthonormal basis for $\otimes^r\mathfrak{U}$. Now, if $T = \Sigma T_r$, $U = \Sigma U_r$, T_r, U_r in $\otimes^r\mathfrak{U}$ are any two elements of the tensor algebra $\mathfrak{A}(\mathfrak{U})$, then an inner product (T, U) in $\mathfrak{A}(\mathfrak{U})$ (as a vector space) can also be defined (see 90). The tensor $G = \sum_{i=1}^{n} E_i \otimes E_i$ is independent of the choice of the orthonormal basis (E) and is called the contravariant metric tensor of the Euclidean space $\mathfrak{U}$; the covariant metric tensor G^* is defined similarly in terms of the dual basis (E^*) of $\mathfrak{U}^*$. In general if $X_1, \cdots, X_n$ and $X^1, \cdots, X^r$ are arbitrary dual bases of $\mathfrak{U}$ and $\mathfrak{U}^*$, respectively, then G^* and G can be expressed in terms of these bases (see 91).

The exterior algebra of a vector space. For the following discussion, three definitions are needed. (1) $\mathfrak{U}$ is a vector space over $\mathfrak{F}$; (2) $\mathfrak{A}_r(\mathfrak{U}) = \otimes^r\mathfrak{U}$ is the rth tensorial power of $\mathfrak{U}$; (3) $\mathfrak{N}_r(\mathfrak{U})$, $r \geqslant 2$, is the subspace of $\mathfrak{A}_r(\mathfrak{U})$ spanned by the elements $V_1 \times V_2 \times \cdots \times V_r$, in which $V_i = V_j$ for at least one pair of indices i, j with $i \neq j$. With these definitions, the factor space $\mathfrak{A}_r(\mathfrak{U})/\mathfrak{N}_r(\mathfrak{U})$ is called the rth exterior power of $\mathfrak{U}$ and is denoted by $\mathfrak{U} \wedge \mathfrak{U} \wedge \cdots \wedge \mathfrak{U}$ or $\wedge^r\mathfrak{U}$, and the element $V_1 \otimes \cdots \otimes V_n + \mathfrak{N}_r(\mathfrak{U})$ is denoted by $V_1 \wedge V_2 \wedge \cdots \wedge V_n$.

The rth exterior power can also be defined, like the rth tensorial power, by a universal mapping property. If $\mathfrak{W}$ is a second vector space over $\mathfrak{F}$, then a multilinear mapping τ from $\mathfrak{U} \times \mathfrak{U} \times \cdots \times \mathfrak{U}$ into $\mathfrak{W}$ is said to be skew symmetric if, for every permutation σ of the subscripts, $\tau(V_1, V_2, \cdots, V_r) = \varepsilon(\sigma)\tau(V_{\sigma(1)}, V_{\sigma(2)}, \cdots, V_{\sigma(r)})$, in which $\varepsilon(\sigma) = 1$ if σ is an even permutation and $\varepsilon(\sigma) = -1$ if σ is an odd permutation. A necessary and sufficient condition that τ be skew symmetric is that $\tau(V_1, V_2, \cdots, V_r) = 0$ whenever $V_i = V_j$ for at least one pair of indices i, j with $i \neq j$. A vector space $\mathfrak{B}_r$ over $\mathfrak{F}$, together with a skew symmetric mapping τ from $\times^r\mathfrak{U}$ into $\mathfrak{B}_r$, is called an rth exterior power of $\mathfrak{U}$ if for every skew symmetric multilinear mapping ρ, from $\times^r\mathfrak{U}$ into

The rth exterior power of a vector space

(84)
$$\begin{cases} \beta:(\mathfrak{U}_1 \otimes \cdots \otimes \mathfrak{U}_r) \otimes (\mathfrak{U}_{r+1} \otimes \cdots \otimes \mathfrak{U}_{r+s}) \\ \quad \to \mathfrak{U}_1 \otimes \cdots \otimes \mathfrak{U}_{r+s} \\ \beta[(\mathfrak{U}_1 \otimes \cdots \otimes \mathfrak{U}_r) \otimes (\mathfrak{U}_{r+1} \otimes \cdots \otimes \mathfrak{U}_{r+s})] \\ \quad = \mathfrak{U}_1 \otimes \cdots \otimes \mathfrak{U}_{r+s} \end{cases}$$

(85)
$$T = \sum_{(i)} x_{i_1 \cdots i_r} E_{i_1} \otimes \cdots \otimes E_{i_r}$$

(86)
$$\begin{cases} 1(V_1 \otimes V_2 \otimes \cdots \otimes V_n) = (V_1 \otimes V_2 \otimes \cdots \otimes V_n)1 \\ \quad = V_1 \otimes V_2 \otimes \cdots \otimes V_n \\ (V_1 \otimes V_2 \otimes \cdots \otimes V_n)(V_1' \otimes V_2' \otimes \cdots \otimes V_m') \\ \quad = V_1 \otimes V_2 \otimes \cdots \otimes V_n \otimes V_1' \otimes \cdots \otimes V_m' \end{cases}$$

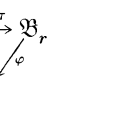
(87)
$$\begin{cases} E^{j_1 j_2 \cdots j_n}_{i_1 i_2 \cdots i_n} = E_{i_1} \otimes E_{i_2} \otimes \cdots \otimes E_{i_n} \otimes \\ \quad \otimes E^{j_1} \otimes E^{j_2} \otimes \cdots \otimes E^{j_n} \end{cases}$$

(88)
$$\begin{cases} T = \sum x^{i_1 i_2 \cdots i_n}_{j_1 j_2 \cdots j_n} E^{j_1 j_2 \cdots j_n}_{i_1 i_2 \cdots i_n}, \\ \text{the summation extending over each } i \text{ and each } j \end{cases}$$

(89)
$$(V_1 \otimes V_2 \otimes \cdots \otimes V_r, W_1 \otimes W_2 \otimes \cdots \otimes W_r) = (V_1, W_1)(W_2, W_2) \cdots (V_r, W_r)$$

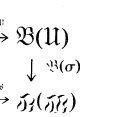
(90)
$$(T, U) = \Sigma(T_r, U_r)$$

(91)
$$G^* = \Sigma(X_i, X_j)X^i \otimes X^j \quad \text{and} \quad G = \Sigma(X^i, X^j)X_i \otimes X_j$$

(92)


(93) $$\mathfrak{B}(\mathfrak{U}) = \oplus \sum_{r=0}^{\infty} \mathfrak{B}_r(\mathfrak{U})$$

(94) $$\begin{cases} 1 \cdot V_1 \wedge V_2 \wedge \cdots \wedge V_r = V_1 \wedge V_2 \wedge \cdots \wedge V_r \cdot 1 \\ \quad = V_1 \wedge V_2 \cdots \wedge V_r \\ V_1 \wedge V_2 \wedge \cdots \wedge V_r \cdot V_{r+1} \wedge V_{r+2} \wedge \cdots \wedge V_{r+s} \\ \quad = V_1 \wedge V_2 \wedge \cdots \wedge V_{r+s} \end{cases}$$

(95)


(96) $$U = \Sigma x^{i_1 \cdots i_n} E_{i_1} \wedge \cdots \wedge E_{i_n}, \qquad i_1 < i_2 \cdots < i_n$$

(97) $$\dim \mathfrak{H} = \sum_{r=0}^{n} \binom{n}{r} = 2^n$$

(98) $$(V_1 \wedge V_2 \wedge \cdots \wedge V_r, W_1 \wedge W_2 \wedge \cdots \wedge W_r) = \det[(V_i, W_j)]$$

an arbitrary vector space $\mathfrak{F}$ over $\mathfrak{f}$, there exists a unique linear transformation ϕ from $\mathfrak{B}_r$ into $\mathfrak{F}$, such that $\phi\tau = \rho$; that is, the relevant diagram (see 92) is commutative. If $\tau(V_1, V_2, \cdots, V_r)$ is written as $V_1 \wedge V_2 \wedge \cdots \wedge V_r$, the definition implies that $\mathfrak{B}_r$ is spanned by the elements of the form $V_1 \wedge V_2 \wedge \cdots \wedge V_r$ and is unique in the sense that, if $(\mathfrak{B}_r, \tau)$ and $(\mathfrak{B}'_r, \tau')$ are two rth exterior powers, then there is an isomorphism α of $\mathfrak{B}_r$ to $\mathfrak{B}'_r$, such that $\alpha\tau = \tau'$. The existence of $\mathfrak{B}_r$ is proved by showing that $\mathfrak{B}_r = \mathfrak{A}_r(\mathfrak{U})/\mathfrak{N}_r(\mathfrak{U})$ has the universal property illustrated in the diagram.

The exterior algebra $\mathfrak{B}(\mathfrak{U})$ of the vector space is constructed from the rth exterior powers by a process analogous to that by which the tensor algebra is constructed from the rth tensorial powers. Defining $\mathfrak{B}_1(\mathfrak{U})$ to be $\mathfrak{U}$, $\mathfrak{B}_0(\mathfrak{U})$ to be $\mathfrak{f}$, the direct sum (see 93) of the $\mathfrak{B}_r(\mathfrak{U})$ can be formed. Each of the spaces $\mathfrak{B}_r(\mathfrak{U})$ is considered as a subspace of $\mathfrak{B}(\mathfrak{U})$. A unique associative multiplication can be defined in $\mathfrak{B}(\mathfrak{U})$ such equations (94) hold. Thus $\mathfrak{B}(\mathfrak{U})$ becomes a linear associative algebra with unity element. It is called the exterior algebra of $\mathfrak{U}$. The product of any two elements X, Y of $\mathfrak{B}(\mathfrak{U})$ can be written $X \wedge Y$ without conflict with the previous use of the symbol $\wedge$.

If $\mathfrak{N}(\mathfrak{U})$ is the direct sum $\oplus \sum_{r=2}^{\infty} \mathfrak{N}_r(\mathfrak{U})$, then $\mathfrak{N}(\mathfrak{U})$ is an ideal in the tensor algebra $\otimes\mathfrak{U}$ and $\mathfrak{B}(\mathfrak{U})$ is isomorphic, as an algebra, to $\otimes\mathfrak{U}/\mathfrak{N}(\mathfrak{U})$. Any elements X in $\wedge^r\mathfrak{U}$ and Y in $\wedge^s\mathfrak{U}$ satisfy the commutation rule $X \wedge Y = (-1)^{rs} Y \wedge X$. In particular ,for any vector V in $\mathfrak{U}$, $V \wedge V = 0$ or, more precisely, if τ is the canonical projection of $\mathfrak{U}$ into $\mathfrak{B}(\mathfrak{U})$, then $\tau V = 0$. This property can be used to formulate a universal mapping definition of the exterior algebra: $\mathfrak{U}$ is taken to be a vector space over $\mathfrak{f}$. An associative algebra $\mathfrak{B}$ over $\mathfrak{f}$, with unity element, together with a linear transformation τ, from $\mathfrak{U}$ into $\mathfrak{B}$, will be called an exterior algebra over $\mathfrak{U}$ if (1) for all V in $\mathfrak{U}$, $(\tau V)^2 = 0$; and (2) for every associative algebra $\mathfrak{A}$ over $\mathfrak{f}$, with unity, and for every linear transformation σ from $\mathfrak{U}$ into $\mathfrak{A}$ such that $(\sigma V)^2 = 0$ for all V in $\mathfrak{U}$, there exists a unique algebra homomorphism ϕ from $\mathfrak{B}$ into $\mathfrak{A}$ such that $\phi\tau = \sigma$. It then follows that $\mathfrak{B}(\mathfrak{U})$, as previously defined, has this universal property and that if $(\mathfrak{B}, \tau)$ and $(\mathfrak{B}', \tau')$ are two exterior algebras over $\mathfrak{U}$, then there is an isomorphism α from $\mathfrak{B}$ to $\mathfrak{B}'$ such that $\alpha\tau = \tau'$

The universal mapping definition of the exterior algebra

The universal property for $\mathfrak{B}(\mathfrak{U})$ implies, as in the case of the tensor algebra, that a linear transformation σ from $\mathfrak{U}$ into $\mathfrak{W}$ induces a unique algebra homomorphism $\mathfrak{B}(\sigma)$ from $\mathfrak{B}(\mathfrak{U})$ into $\mathfrak{B}(\mathfrak{W})$ such that the relevant diagram (see 95) is commutative, $\mathfrak{B}(I) = I$, and $\mathfrak{B}(\sigma\rho) = \mathfrak{B}(\sigma)\mathfrak{B}(\rho)$ and hence the mapping of $\mathfrak{U}$ on $\mathfrak{B}(\mathfrak{U})$ and σ on $\mathfrak{B}(\sigma)$ is a functor on the category of vector spaces over $\mathfrak{f}$ to the category of associative algebras over $\mathfrak{f}$.

If $\mathfrak{U}$ is finite-dimensional with basis (E), then a basis for the rth exterior power $\wedge^r\mathfrak{U}$ is the set of exterior products $\{E_{i_1} \wedge \cdots \wedge E_{i_n}\}$ in which $i_1 < i_2 < \cdots < i_n$. Thus every vector U in $\wedge^r\mathfrak{U}$ can be expressed in the form (96) in which scalars $x^{i_1} \cdots {}^{i_n}$ are called the components of U with respect to the E-basis of $\mathfrak{U}$. If $\dim \mathfrak{U} = n$, then $\dim \wedge^r V = \binom{n}{r}$ if $r \leqslant n$. If $r > n$, then $\wedge^r V = 0$, because in that case $\otimes^r\mathfrak{U}$ is a subset of $\mathfrak{N}_r(\mathfrak{U})$. Thus, if $\mathfrak{B} = \wedge\mathfrak{U}$ is the exterior algebra of an n-dimensional space $\mathfrak{U}$, $\dim \mathfrak{B}$ is determined by (97). Finally, if $\mathfrak{U}$ is an inner product space with inner product $(\ ,\ )$ then an inner product can be defined in $\mathfrak{B}$ (see 98). If (E) is an orthonormal basis of $\mathfrak{U}$, then $\{E_{i_1} \wedge \cdots \wedge E_{i_r}\}$, with $i_1 < i_2 < \cdots < i_r$, is an orthonormal basis of $\mathfrak{B}$.

BIBLIOGRAPHY. RICHARD BELLMAN, *Introduction to Matrix Analysis* (1960), a comprehensive treatment of matrix theory and its applications to differential equations, probability theory and mathematical economics; SIR ARTHUR CAYLEY, *The Collected Mathematical Papers* (1897), mainly of historical interest; F.R. GANTMACHER, *Theory of Matrices*, vol. 1 and 2 trans. from Russian by K.A. HIRSCH (1959), a standard textbook, advanced level; W.H. GREUB, *Linear Algebra*, 3rd ed. (1967), a standard textbook, advanced level; and *Multilinear Algebra* (1967), a standard modern treatment of multilinear algebra that fully exploits universal mapping properties; P.R. HALMOS, *Finite-Dimensional Vector Spaces*, 2nd ed. (1958), a standard textbook, advanced level, which emphasizes methods that can be carried over to infinite-dimensional spaces, particularly to Hilbert space; SIR WILLIAM ROWAN HAMILTON, *Lectures on Quaternions* (1853), mainly of historical interest; K. HOFFMAN and RAY KUNZE, *Linear Algebra*, 2nd ed. (1971), a standard textbook, intermediate level; N. JACOBSON, *Lectures in Abstract Algebra*, vol. 2, *Linear Algebra* (1953), a standard textbook, advanced level; C.C. MACDUFFEE, *The Theory of Matrices*, 2nd ed. (1946), a concise but comprehensive account of matrix theory up to 1932 including copious references to original sources; and *Vectors and Matrices* (1943), a comparatively elementary discussion of similarity and the derivation of the rational and Jordan canonical forms of a matrix; MARVIN MARCUS and HENDRYK MINC, *Introduction to Linear Algebra* (1965), a standard textbook, elementary level; and *A Survey of Matrix Theory and Matrix Inequalities* (1964), a very concise account, omitting standard proofs, of the main facts of matrix algebra, with many references to original sources; DAVID C. MURDOCH, *Linear Algebra* (1970), a standard textbook, elementary level; THOMAS MUIR, *Theory of Determinants in the Historical Order of Development*, 2nd ed., 4 vol. (1906–23), an exhaustive history of the development of determinants to 1900; S. PERLIS, *Theory of Matrices* (1952), a standard textbook, intermediate level; O. SCHREIER and E. SPERNER, *Einführung in die analytische Geometrie und Algebra*, vol. 1 (1959; Eng. trans of vol. 1, *Introduction to Modern Algebra and Matrix Theory*, 2nd ed., 1959), a standard textbook, intermediate level; ROBERT C. THOMPSON and ADIL YAQUB, *Introduction to Linear Algebra* (1970), a standard textbook, at the elementary level; J.H.M. WEDDERBURN, *Lectures on Matrices* (1934, reprinted 1964), a comprehensive account of matrix theory, including linear associative algebras, up to 1933—this book includes brief historical notes and also a bibliography from 1853 to 1933, which is arranged chronologically.

(Ed.)

Algebraic Structures

Out of the developments of elementary algebra evolved the abstract algebra used today and the idea of an algebraic structure. Elementary algebra was originally concerned with a set of elements, the numbers used in arithmetic. Together with these elements there were two operations, addition and multiplication (subtraction and division being the inverse of these). It became recognized that there was also a collection of basic rules that constitute an axiomatic structure. The axiomatic structure describes even today the assumptions made in elementary algebra and arithmetic. Certain entities, however, do

not follow these rules. Abstract algebra is concerned with the formulation and properties of quite general axiomatic abstract systems of this type. These systems are sets of elements with general operations and with a number of axioms. Just as new self-consistent geometries can be based on axioms other than those of Euclid, new algebras can be based on axioms that differ from those of elementary algebra. These new algebras may describe mathematical objects other than numbers. This article is concerned with algebraic structures of major importance as shown in the outline below.

I. Lattice theory
- Definitions and examples
- Boolean algebra
- Distributive lattices
- Modular and geometric lattices
- Other lattice-ordered structures

II. Groups
- Definition
- Examples of groups
- Elementary group structure
- Subgroups
- Homomorphisms; factor groups
- Direct products; Abelian groups
- Automorphisms of groups; semi-direct products
- Permutation groups
- Group representation
- The classical groups; continuous groups
- Free groups; the word problem; the Burnside problem
- Simple groups

III. Fields
- Three familiar fields
- Elementary theory of fields
- Some general concepts
- Field extensions
- Applications to coding
- Galois theory
- Ordered fields
- Valuations

IV. Rings
- Definition and examples
- Algebras
- Matrices; representations of rings
- Rings as modules
- Special classes of rings

V. Categories
- The notion of a category
- Functors
- Natural transformations
- Universal constructions
- The duality principle
- Conclusions

VI. Homological algebra
- Homology groups of groups
- Derived functors
- Homology of Lie algebras
- Applications to algebra
- Other aspects of homological algebra

VII. Universal algebra
- Basic notions of universal algebra
- The connection with mathematical logic
- Varieties of algebras
- Universal algebra and category theory

I. Lattice theory

Lattice theory is the branch of mathematics that deals in precise mathematical language with the relation of different parts of the same whole to each other. The fundamental concept is that one part x may include (or contain) another part y, a relation written symbolically as $x \geq y$. A lattice, defined in more detail below, is a set of elements related by an inclusion relation such that, for any x and y in the set, there is a largest element, denoted $x \wedge y$, that is contained in both x and y and a smallest element, denoted $x \vee y$, that contains both x and y.

In general, in lattice theory the concepts of order and structure are analyzed from a mathematical standpoint, much as group theory (described later) analyzes the concept of symmetry. Thus, lattice theory considers families of subsets $X, Y, \cdots$ of a given set U under the operation of intersection of sets, denoted $X \cap Y$, in much the same spirit that group theory considers families of symmetry transformations (symbolized by Greek letters alpha, beta, $\cdots$) $\alpha, \beta, \cdots$ of a given configuration when combined by performing them in succession (*i.e.*, taking their composite $\alpha \circ \beta$). As a generalization of set theory, it forms a fundamental part of higher algebra, geometry, and mathematical logic.

DEFINITIONS AND EXAMPLES

Partial orderings. In detail, given a selected family S of different parts of the same whole, these parts can be represented symbolically by letters $x, y, \cdots, x \geq y$ signifying that x contains y. The resulting inclusion relation is characterized by the reflexive law, which states that $x \geq x$ for all x; the transitive law, which states that if $x \geq y$ and $y \geq z$, then $x \geq z$; and the antisymmetric law, which states that if $x \geq y$ and $y \geq x$, then $x = y$. In general, any relation having the above properties is called a partial ordering; and a partially ordered set, or poset for short, is just a set S together with a partial ordering of its elements.

The reflexive law

The inclusion relation is basic for the study of the subsets of a set (class, or aggregate). The relation of implication in logic has the same formal properties. In logic, $p \supset q$ signifies that the proposition (or statement) p implies the proposition q: that "if p, then q." The reflexive law $p \supset p$ of logic states that any proposition p implies itself; the transitive law, that if p implies q and q implies r, then p implies r; the antisymmetric law, that if p implies q and q implies p, then p and q are logically equivalent. Thus the propositions of any deductive system (*e.g.*, lattice theory or group theory) form a poset when partially ordered by the implication relation.

Mathematics is concerned with many other familiar partial orderings, such as the relation $x \leq y$ of inequality between real numbers, and the relation $a \mid b$ of divisibility (a is said to divide b when $b = na$ for some integer n) between positive integers. That is, if $\Re$ denotes the set of all real numbers and inequality has its usual meaning, then $[\Re, \leq]$ is a partially ordered set or poset. Likewise, if P denotes the set of all positive integers and $a \mid b$ means that $b = na$ for some positive integer n, then $[P, \mid]$ is a poset.

Universal bounds. In a poset P, the identical inequalities relating an element x with an element 0, for all x, and relating an element x with an element I, for all x (see Box, inequalities 1), can be satisfied by at most one

(1) $\quad 0 \leq x$ for all x, $\qquad x \leq I$ for all x

element 0 and by at most one I, respectively. When they exist, 0 is called the null element and I the universal element of the poset, which is said to have universal bounds. In the poset of the subsets $X, Y, \cdots$ of a set U, for example, the empty set, represented by the symbol $\emptyset$, is the null element and U the universal element, because $\emptyset \subset X \subset U$ for any subset X of U. In the poset $[P, \mid]$, 1 is the null element because 1 is a divisor of every positive integer n (in symbols, $1/n$ for all n in P). The poset $[P, \mid]$ has no universal element.

Formal definition. A formal definition of a lattice can now be given. It is in general any abstract system L of elements $x, y, z, \cdots$ having a reflexive, transitive, and antisymmetric relation that satisfies the following lattice condition. For any two elements x and y of L there is a least element including both x and y, denoted by $x \vee y$, and a greatest element included in both x and y, denoted by $x \wedge y$. More precisely, it is required that $x \leq x \vee y$, that $y \leq x \vee y$, and that $x \leq z$ and $y \leq z$ together imply $x \vee y \leq z$. Similarly, it is required that $x \wedge y \leq x$, that $x \wedge y \leq y$, and that $w \leq x$ and $w \leq y$ together imply $w \leq x \wedge y$. Such elements $x \vee y$ and $x \wedge y$ are called the least upper bound (l.u.b.) and greatest lower bound (g.l.b.) of x and y, respectively; when they exist, they are unique.

The preceding examples of posets satisfy this lattice condition. In the case of the subsets $X, Y, \cdots$ of an aggregate U, $X \vee Y$ is the set-union (or sum $X \cup Y$) of

X and Y, and $X \wedge Y$ the set of elements common to X and Y (their intersection $X \cap Y$). In logic, $p \wedge q$ is the statement "p and q"; $p \vee q$ means "p or q" (or both p and q); with real numbers, $x \vee y$ is simply the greater and $x \wedge y$ the lesser of x and y. In the example of the positive integers, $m \vee n$ is the least common multiple m and n and $m \wedge n$ is the greatest common divisor of m and n.

Examples of lattices. Lattices of various kinds arise in all sorts of mathematical situations. Some of these have very few elements and can be conveniently described by diagrams: any graph with sloping lines defines a poset, the elements of which are the nodes of the graph (the points at which two or more lines meet), and in which $a \geq b$ means that one can go from a to b along a sequence of line segments that slope downward. Figures

Posets

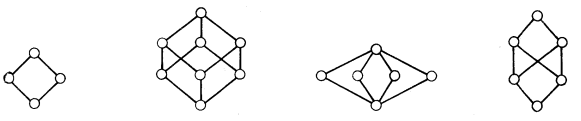


Figure 1: Diagrams of lattices and one non-lattice (see text).

1A–1D describe lattices, of which Figure 1A represents the lattice of all subsets of a set of two elements and Figure 1B that of all subsets of a set of three elements. Figure 1C represents the lattice of all subgroups of the group G of all six symmetries of an equilateral triangle. The bottom element represents the (trivial) subgroup consisting of the group identity; the first three elements (nodes) above it the subgroups generated by the reflections in the three altitudes; the fourth the subgroup of rotations of the triangle into itself; and the top element G itself.

Figure 1D represents a poset that is not a lattice: the two elements immediately above the bottom element (the universal lower bound 0) do not have a least upper bound. All four posets of Figures 1A–1D have universal bounds: their bottom and top elements.

Figure 2 represents the partition lattice π_4 of all partitions of a set of four elements, in which by a partition of

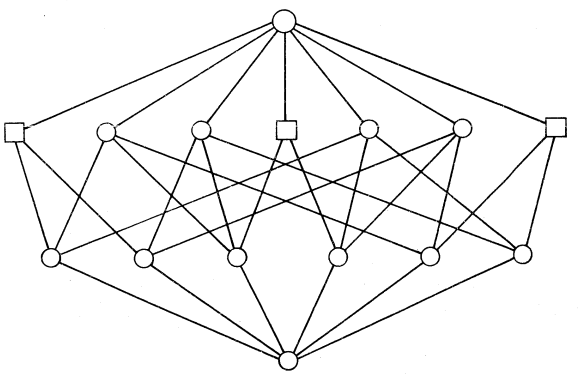


Figure 2: The partition lattice π_4 (see text).

a set is meant a division of its elements into nonoverlapping subsets. In Figure 2, the square nodes represent the three partitions of the set $\{a, b, c, d\}$ into two subsets of two elements, namely: $(ab)(cd)$, $(ac)(bd)$, and $(ad)(bc)$. The circles on the same level represent the four partitions (123)(4), (124)(3), (134)(2), (234)(1) into one subset of three elements and a singleton; and so on.

Algebraic definition. Lattices can also be defined in terms of the algebraic properties of the two operations $\wedge$ and $\vee$ just defined, without reference to the concept of order. In a general poset P, the l.u.b. $x \vee y$ of two elements x and y (when it exists) is called the join of x and y, and the g.l.b. $x \wedge y$ (when it exists) is called their meet. The relatively colourless words "join" and "meet" are intended to suggest "union" and "intersection" but to refer only to the underlying order abstractly. In any lattice, the join and meet operations are commutative and associative, so that $x \vee y = y \vee x$, $x \wedge y = y \wedge x$, $x \vee (y \vee z) = (x \vee y) \vee z$, and also $x \wedge (y \wedge z) = (x \wedge y) \wedge z$ for all x, y, z. They are idempotent, which means that $x \vee x = x \wedge x = x$ for all x; and they satisfy the law of contraction: $x \wedge (x \vee y) = x \vee (x \wedge y) = x$ for all x, y.

The law of contraction

It is easy to verify the preceding laws in the examples described above; the essential fact is that they can also be deduced formally from the reflexive, transitive, and antisymmetric laws, considered as postulates used to define the abstract concept of a poset.

Conversely, a lattice can be defined as an algebraic system, or algebra, having two commutative, associative, idempotent, and contractive operations. This system can be designated symbolically as $[S, \wedge, \vee]$, in which S stands for the set consisting of all the elements of the lattice, and $\wedge$, $\vee$ for its two operations. The inclusion relation may be defined in terms of either lattice operation, because $x \leq y$ is equivalent either to $x \wedge y = x$ or to $x \vee y = y$. In particular, O is characterized in any lattice by either of the laws $O \wedge x = O$ or $O \vee x = x$, valid for all x; while I is characterized by $x \wedge I = x$ or by $x \vee I = I$.

Thus, the lattice of Figure 1A is defined by means of the tables for combining elements with the operations $\wedge$ and $\vee$ (see 2). From either of these tables, it can be de-

(2)

$\wedge$	0	a	b	I
0	0	0	0	0
a	0	a	0	a
b	0	0	b	b
I	0	a	b	I

$\vee$	0	a	b	I
0	0	a	b	I
a	a	a	I	I
b	b	I	b	I
I	I	I	I	I

duced that $0 < a, b < I$, and that these are the only proper inclusion relations.

More generally, $L = [S, \wedge, \vee]$ is assumed to be any algebraic system having two operations $\vee$, $\wedge$ that are idempotent, commutative, associative, and satisfy the contraction laws. $x \leq y$ is defined in L to mean that $x \vee y = y$ (or equivalently, that $x \wedge y = x$). Then the relational system $[S, \leq]$ defined thereby is a lattice in the sense of the definition above, in which $x \wedge y$ is the meet (g.l.b.) and $x \vee y$ the join (l.u.b.) of x and y. Hence the idempotent, commutative, associative, and contraction laws can also be used as postulates with which to define the concept of a lattice.

Other examples. There are many kinds of lattices. Thus, a wide class of lattices arises from the study of groups. The subgroups of any group G form a lattice, in which G itself and the identity element 1 act as universal bounds. Figure 3E shows the lattice of all subgroups of the octic group of symmetries of a square. A few other small lattices are depicted in Figures 3A–3D. Thus Figure 3A depicts a chain; *i.e.*, a poset in which for all x, y, either $x \leq y$ or $y \leq x$. Figure 3C depicts the smallest lattice in which not all chains from 0 to I have the same length.

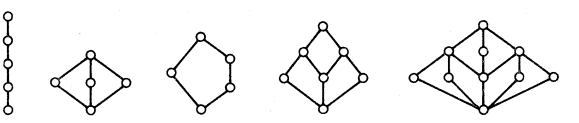


Figure 3: Lattices with special properties (see text).

Duality. The concepts of a poset and of a lattice have important duality properties. If $y \leq x$ is defined to mean that $x \geq y$ (by conversion), it is evident that the relation $\leq$ is reflexive, transitive, and antisymmetric if and only if $\geq$ is. Furthermore, if $\geq$ is replaced by $\leq$, $\vee$ and $\wedge$ are simply interchanged in the statement of the lattice condition without affecting the set of postulates as a

whole. In other words, the concepts of a poset and of a lattice are self-dual under the duality between greater and less (container and the contained).

In the particular lattice $[P, |]$, for instance, the dualization effected by the preceding construction changes greatest common divisors (g.c.d.) into least common multiples (l.c.m.), and vice versa. More generally, it turns the diagram of any finite lattice upside-down.

Isotonicity. Lattices have many general properties. Thus, any lattice with a finite number of elements $x_1, \cdots, x_n$ has the universal bounds $x_1 \wedge \cdots \wedge x_n = 0$ and $x_1 \vee \cdots \vee x_n = I$. More important is the following isotonicity principle, valid in any lattice. This principle is: If for any elements u, u^*, v, and v^*, in a lattice L, $u \geq u^*$ and $v \geq v^*$, then $u \vee v \leq u^* \vee v^*$ and $u \wedge v \geq u^* \wedge v^*$: the lattice operations preserve order.

Power set lattices. The power set lattice $[P(U), \subset]$, defined by the inclusion relation on the power set $P(U)$ of all subsets of a set U, has important special properties not shared by lattices in general. One of these is distributivity (see 3) that holds for any three subsets X, Y, Z of U. In other words, defining a distributive lattice as a lattice in which this property holds, as expressed with the operations $\wedge$ and $\vee$ (see 4), for all x, y, z, $[P(U), \subset]$ is always a distributive lattice.

Second, every power set lattice $[P(U), \subset]$ has the universal bounds property: the empty set $\emptyset$ and U itself are universal bounds (in the sense of (1)).

Finally, in the lattice $[P(U), \subset]$ each set X has a complement X', consisting of all elements of U that are not in X. Clearly, two specific properties hold with respect to the complement of a set and the operations $\wedge$ and $\vee$; these are to be expressed in more general form (see 5) in what follows. A poset P may now be defined to be complemented when: (A) it has universal bounds 0 and I satisfying (1), and (B) any x in P has a complement x' in P such that two properties hold (see 5). Summarizing the above, then, it can be said that any power set lattice $[P(U), \subset]$ is a complemented distributive lattice.

BOOLEAN ALGEBRA

Boolean polynomials. Complemented distributive lattices arise in many other contexts besides set theory. They were first studied systematically by the English logician George Boole around 1850, because of their applications to logic, and are called Boolean algebras.

One of the most important concepts of Boolean algebra is that of a Boolean polynomial, by which is meant any symbolic expression built up by repeated application of the two binary lattice operations $\wedge$, $\vee$, and the unary operation $'$, and the special symbols 0 and I. The algebraic manipulations of Boolean polynomials are governed by the laws given below: the idempotent, commutative, associative, contractive, and distributive laws, plus the universal bounds and complementation laws (see 4 and 5). From these many others can be deduced, among which de Morgan's laws (see 6), the law $(x')' = x$, and the curious identity involving three elements combined in pairs with the operations $\wedge$ and $\vee$ (see 7) are especially noteworthy. Also, the inclusion relation $x \leq y$ is equivalent to either $x \wedge y' = 0$ or $x' \vee y = I$ as well as to a pair of inequalities that do not involve 0 and I (see 8).

Laws governing manipulation of Boolean polynomials

Extensive research into the foundations of Boolean algebra has shown that the list of laws given in the preceding paragraph is highly redundant, and that various small subsets taken from this list imply the others. For example, if $[L, \vee, ']$ is an algebraic system that is assumed to have one binary operation $\vee$ and one unary operation $'$, then the following was proved by the U.S. mathematician Edward V. Huntington in 1933.

Huntington's theorem. In an algebraic system $[L, \vee, ']$ define $a \wedge b = (a' \vee b')'$. Suppose $a \vee b = b \vee a$, $a \vee (b \vee c) = (a \vee b) \vee c$, and $(a \wedge b) \vee (a \wedge b') = a$ for all elements a, b, c in L, then $[L, \vee, ']$ is a Boolean algebra.

Venn's diagram. The study of the laws of Boolean algebra, including Huntington's theorem, was motivated by the desire to reduce all of logic and set theory to a purely symbolic calculus. This study is helped by using

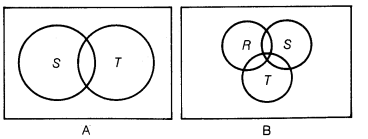


Figure 4: Venn diagrams (see text).

so-called Venn diagrams, of which two are shown in Figure 4. In them, U has been taken to be a rectangle. Figure 4A depicts two discs S and T; they are to be imagined as typical subsets of U. Clearly the intersections $S \cap T$ $S \cap T'$, $S' \cap T$ and $S' \cap T'$ divide U into four nonoverlapping parts. If boundaries are neglected, the unions of these parts give precisely $2^4 = 16$ different Boolean combinations of the original sets S, T.

Likewise, John Venn, the British logician for whom the diagram was named, considered three discs R, S, and T as typical subsets of U. The intersections of these discs and their complements divide U into $8 = 2^3$ nonoverlapping regions, the unions of which give $2^8 = 256 = 2^{2^3}$ $= 16^2$ different Boolean combinations of the original sets S, T.

Free Boolean algebras. These examples suggest the conjecture that, in general, there are exactly 2^{2^n} different Boolean polynomials in n symbols $x_1, \cdots, x_n$, such that any given polynomial can be reduced to one, and only one, of these by repeated applications of the basic laws listed above. The truth of this conjecture is part of the following basic theorem, essentially due (in another terminology) to Boole himself.

Boole's theorem. A Boolean polynomial $(x_1, \cdots, x_n)$ can be reduced to one and only one join $\vee_S p_k$ of meets $p_k = z_1 \wedge \cdots \wedge z_n$, in which each z_i is x_i or x'_i (there being 2^n possibilities), depending on k, and S is any subset of the 2^n p_k.

The preceding form, $\vee_S p_k$, is called the disjunctive normal form of the polynomial and provides a systematic way of testing for the equivalence of two Boolean polynomials. Thus it solves the so-called word problem for Boolean algebra in an elementary way.

Isomorphism; direct product. Two general algebraic concepts, which are applicable to groups and rings (for example) as well as to posets, lattices, and Boolean algebras, are those of isomorphism and direct product. They will now be defined in the general context of posets.

An isomorphism between two posets P and Q is, by definition, a pairing of the elements of P with the elements of Q (a so-called bijection β: $P \leftrightarrow Q$, the symbol $\leftrightarrow$ denoting bijection) such that $x \leq y$ in P if and only if $\beta(x) \leq \beta(y)$ in Q. If P is a lattice, then the preceding condition implies that Q is also a lattice and that $\beta(x \wedge y) = \beta(x) \wedge \beta(y)$ and $\beta(x \vee y) = \beta(x) \vee \beta(y)$. If P is a Boolean algebra, then so is Q, and in addition $\beta(x') = [\beta(x)]'$: isomorphisms of posets carry complements into complements.

Again, the direct product $P \times Q$ of two posets P and Q is the set of all ordered pairs (x, y) of elements, x from P and y from Q, the inclusion relation $(x, y) \leq (x_1, y_1)$ in

(3) $$X \cap (Y \cup Z) = (X \cap Y) \cup (X \cap Z)$$
$$X \cup (Y \cap Z) = (X \cup Y) \cap (X \cup Z)$$

(4) $$\begin{cases} x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z) \\ x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z) \end{cases}$$

(5) $$x \wedge x' = 0, \qquad x \vee x' = I$$

(6) $$(x \wedge y)' = x' \vee y', \qquad (x \vee y)' = x' \wedge y'$$

(7) $$(x \wedge y) \vee (y \wedge z) \vee (z \wedge x) = (x \vee y) \wedge (y \vee z) \wedge (z \vee x)$$

(8) $$x \wedge y = x, \qquad x \vee y = y$$

(9) $(x, y) \wedge (x_1, y_1) = (x \wedge x_1, y \wedge y_1)$

(10)

$\vee$	0	1
0	0	1
1	1	1

$\wedge$	0	1
0	0	0
1	0	1

a	a'
0	1
1	0

(11) $\xi(a_i) = \begin{cases} 1 & \text{if } x \geq a_i \\ 0 & \text{otherwise} \end{cases}$

$P \times Q$ being defined to mean that $x \leq x_1$ in P and $y \leq y_1$ in Q. If P and Q are both lattices or both Boolean algebras, then the same is true of $P \times Q$; moreover, the combination of pairs is achieved by the operation $\wedge$ (see 9) in $P \times Q$, and dually by the operation $\vee$.

It is easy to show that the direct product of two posets is always a poset and that the formation of direct products is a commutative and associative operation on posets, relative to isomorphism: that $P \times Q \simeq Q \times P$ and $P \times (Q \times R) \simeq (P \times Q) \times R$. For instance, the bijection that pairs $(x, (y, z))$ with $((x, y), z)$ for any x in P, y in Q, and z in R is an isomorphism of $P \times (Q \times R)$ with $(P \times Q) \times R$.

It can therefore be written $P \times P = P^2$, $(P \times P) \times P = P^3, \cdots$, with the understanding that $P^m \times P^n = P^{m+n}$ for any positive integers m, n; *i.e.*, that the usual laws for powers hold.

Binary strings. The simplest nontrivial Boolean algebra has two elements, 0 and 1, on which its three operations act as is prescribed by certain tables (see 10). Boolean algebra has been called the algebra of 0 and 1, because every Boolean algebra can be built up by combining copies of this simple Boolean algebra, which will be designated by the symbol 2.

Thus, if the set U has n elements, then the power set lattice (Boolean algebra) $P(U)$ is isomorphic with $2^n = 2 \times \cdots \times 2$ (n factors). To establish the isomorphism $P(U) = 2^n$, the elements of U must be arranged in some order, as a sequence $p_1, p_2, \cdots, p_n$. Each subset A of U is then associated with an n-tuple $a = (a_1, \cdots, a_n) = \beta(A)$ of 0's and 1's, by setting $a_j = 1$ when p_j is in A and $a_j = 0$ when it is not. Dropping commas and parentheses, this associates with each A a binary string $a_1 a_2 \cdots a_n$.

Logic networks. Boolean algebra is used for designing so-called logic networks for digital computers. Such a network typically has a finite number n of input leads, each of which goes into one of two states when pulsed, which can be designated as 0 and 1. The state of the set of all n leads can therefore be described by a binary string of length n. The problem is to design a circuit that, when it is interposed between these input leads and a specified set of output leads, will put each of the latter in a specified state, 0 or 1, determined by the states of the input leads.

The problem is thus to implement electronically a given Boolean function $f\colon 2^n \to 2$ from the set of all binary strings of length n to the set $2 = \{0, 1\}$; but these functions constitute precisely the free Boolean algebra with n generators, namely, the functions $\delta_i\colon (x_1, \cdots, x_n) \to x_i$ that associate with each string its ith member. Boole showed that every function $f\colon 2^n \to 2$ can be represented as a Boolean polynomial in the δ_i (in disjunctive normal form).

The problem of optimal logic network design is to find the Boolean polynomial that does this most economically. Roughly, this corresponds to finding the shortest Boolean polynomial that has a given disjunctive normal form. This problem still lacks a general solution.

DISTRIBUTIVE LATTICES

Switching circuits. The three basic Boolean operations $\wedge$, $\vee$, $'$ (corresponding to "and," "or," and "not" in logic) can be realized electronically by three kinds of simple circuits, called AND-gates, OR-gates, and inverters, respectively. The first general purpose digital computers used, instead of electronic circuits, ordinary relay networks the connections of which could be open or closed. As the U.S. engineer and mathematician Claude E. Shannon showed in 1938, the Boolean operations of "and" and "or" are realized by putting relay networks in series and parallel, respectively. This raises the problem of constructing Boolean functions $f\colon 2^n \to 2$ from $\wedge$ and $\vee$ alone without using "not."

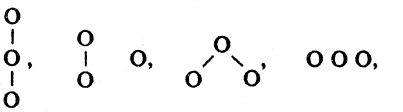
Algebraically, this amounts to determining the most general distributive lattice—*i.e.*, the lattice satisfying the distributive laws (see 4)—with n generators. This is also the set of all lattice polynomials in n symbols $\delta_1, \cdots, \delta_n$, two polynomials being considered equal when equality follows from the idempotent, commutative, associative, contractive, and distributive laws. This free distributive lattice had been considered by the German mathematician Richard Dedekind in 1900, and in 1912 the Norwegian mathematician Thoralf Albert Skolem showed that it contained precisely those Boolean functions $f\colon 2^n \to 2$ that were order-preserving in the following sense. The domain 2^n of the function f may be considered as itself a Boolean algebra, by defining $x \leq y$ for two binary strings $x = (x_1, \cdots, x_n)$ and $y = (y_1, \cdots, y_n)$ to mean that $x_i \leq y_i$ for all $i = 1, 2, \cdots, n$. Then f preserves order (is isotone) when $x \leq y$ implies $f(x) \leq f(y)$.

Finite distributive lattices. The most general finite distributive lattice may be constructed by a simple generalization of Skolem's theorem. Namely, L is assumed to be any finite distributive lattice, and A is assumed to be the set of all properly join-irreducible elements of L; that is, the set of all elements $a > 0$ in L with the property that $x \wedge y = a$ implies $x = a$ or $y = a$. The elements of a may be imagined as listed in some order, $a_1, \cdots, a_n$. Then each x in L can be associated with the subset $S(x)$ of A, consisting of exactly those elements a_i in A such that $x \geq a_i$. This is clearly order-preserving: $x \geq y$ implies $S(x) \supset S(y)$. Alternatively, each x in L can be associated with the function $\xi\colon A \to 2$ defined by assigning the value 1 to $\xi(a_i)$ when $a_i \leq x$, and 0 otherwise (see 11); this function is isotone.

Generalization of Skolem's theorem

The fundamental theorem on finite distributive lattices asserts that the converse is also true: that every isotone function $f\colon A \to 2$ is associated in this way with some x in L.

Some examples. Figures 5A–5B exhibit the diagrams of four distributive lattices, illustrating the preceding statements. The posets of join-irreducible elements, corresponding to the black circles, have the diagrams



respectively.

In particular, Figure 5C exhibits similarly the free distributive lattice with two generators: the poset of all isotone functions $f\colon 2^n \to 2$; it should be noted that it has four properly join-irreducible elements, which form a poset having the diagram



of 2^2, as stated in Skolem's theorem. It should also be noted that this distributive lattice has length three, which means that one can ascend from the universal lower bound 0 to the universal upper bound I in four steps. This illustrates the fact that, in general, a distributive

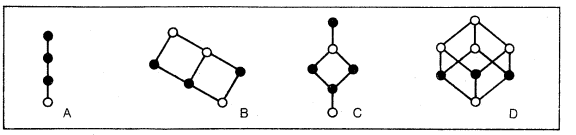


Figure 5: Some distributive lattices (see text).

lattice of finite length n contains exactly n properly join-irreducible elements, a result that can also be verified (with $n = 3$) in the other examples of Figure 5.

MODULAR AND GEOMETRIC LATTICES

Modular lattices. Although distributive lattices (including especially Boolean algebras) are very important, they are not typical. Most lattices are not distributive and cannot be characterized so simply. The rest of this section will describe a few basic properties of the most notable kinds of nondistributive lattices.

The five-element lattice

The simplest nondistributive lattice is the five-element lattice of Figure 3B. Its elements satisfy the modular law (see 12). This is clearly implied by the distributive law $x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$, because $x \wedge z = z$ when $x \geq z$. It is also self-dual, in the sense of being equivalent to its dual (see above; see 13).

(12) If $x \geq z$, then $x \wedge (y \vee z) = (x \wedge y) \vee z$.

(13) If $z \leq x$, then $z \vee (y \wedge x) = (z \vee y) \wedge x$.

Modular lattices arise in many parts of algebra. Thus the subgroups of any Abelian group form a modular lattice. More generally, the normal subgroups of any group form a modular lattice, in which, moreover, the quotient-group $M \vee N/N$ is isomorphic to $M/M \wedge N$. In 1937 the Norwegian-born U.S. mathematician Oystein Ore showed that most of the structural properties of groups can be derived from these properties alone.

Covering properties. In any poset P, the relation "a covers b" is defined to mean that $a > b$, but $a > x > b$ for no x. Graphically, "a covers b" means that a is just above b in the diagram of P.

In 1900, Dedekind proved that a lattice of finite length was modular if and only if its covering relation has the following two dual properties:

1. If b and c both cover a, then $b \vee c$ covers both b and c.
2. If a covers both b and c, then b and c both cover $b \wedge c$. (It is understood here that $b \neq c$.)

Projective geometries. If F is any field (see below *Fields*) or division ring, and if $V = V_{n+1}(F)$ is the vector space (see ALGEBRA, LINEAR AND MULTILINEAR) of all vectors $x = (x_0, x_1, \cdots, x_n)$ of length $n + 1$ with coordinates x_i in F, then the subspaces of V form a modular lattice under the relation of inclusion, which is called the n-dimensional projective geometry over F, and denoted $PG(F; n)$. (See GEOMETRY, PROJECTIVE.)

In any projective geometry $PG(F; n)$, the universal lower bound 0 is the trivial subspace consisting of the null vector $0 = (0, 0, \cdots, 0)$ alone. Geometrically, it is interpreted as the empty set, and the projective points are the elements that cover it. The (projective) lines are the elements that cover points, and every projective line contains at least three points.

The Fano plane

Figure 6 exhibits the diagram of the lattice $PG(Z_2; 2)$, the so-called Fano plane; it has seven points and seven lines, with exactly three points on each line. Its points are the strings of three binary digits, not all 0: 001, 010, 011, 100, 101, 110, 111; its lines are the sets of points $x_1x_2x_3$ that satisfy an equation of the form $a_1x_1 + a_2x_2 + a_3x_3 \equiv 0 \pmod 2$, in which $a_1a_2a_3$ is one of the strings just listed.

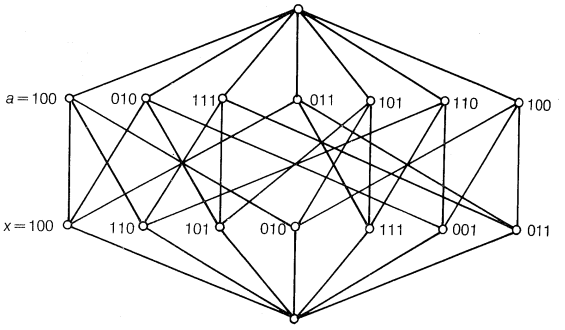


Figure 6: Poset diagram of the Fano plane.

Geometric lattices. In any poset with universal lower bound 0, a point (or atom) is an element p that covers 0. That is, it is necessary that $p > 0$ (meaning that $p \geq 0$ but that $p \neq 0$), while $p > x > 0$ for no x. This definition is essentially the same as Euclid's: "a point is that which has no parts." A lattice of finite length is called a geometric lattice when: (1) each of its elements is a join of points, and (2) it satisfies the covering condition (:) under covering properties above. The partition lattice π_4 of Figure 2 and the Fano plane of Figure 6 are (finite) geometric lattices. More generally, any projective geometry is a (modular) geometric lattice, and the lattice of all partitions of any finite set is a geometric, which is not modular if the set contains more than three elements.

Again, if Ω is any map, then the submaps formed from Ω by obliterating boundaries forms a geometric lattice $G(\Omega)$. The number of ways in which Ω can be coloured in λ colours so that no two adjacent regions have the same colour is called the chromatic polynomial of Ω [and of $G(\Omega)$]. This can also be constructed from the Möbius function of $G(\Omega)$, a function that can be defined on any poset and that has many applications to number theory and combinatorial questions.

OTHER LATTICE-ORDERED STRUCTURES

Relation algebras. Ordering relations play an important role in many algebraic structures having operations that, unlike the meet and join operations, cannot be defined purely in terms of order. One such class of structures is defined by binary relations.

Structures defined by binary relations

By a binary relation on a set S is meant a rule ρ that tells for any two elements x and y of S whether x and y are in the given relation or not. If they satisfy the relation, $x\rho y$ is written; if they do not satisfy the relation, $x\rho' y$ is written. The set of all binary relations on S forms a Boolean algebra under the operations $\rho \wedge \sigma$, $\rho \vee \sigma$, and ρ' corresponding to the words "and," "or," and "not." [Thus, $x(\rho \wedge \sigma)y$ means that $x\rho y$ and $x\sigma y$.] In this Boolean algebra, the order relation $\rho \geq \sigma$ defined by the statement that $\rho \wedge \sigma = \sigma$ (ρ and σ is the same as σ) is clearly tantamount to saying that σ "implies" ρ (the statement that $x\rho y$ "adds nothing" to the statement that $x\sigma y$). In logic, this is written $\sigma \supset \rho$.

The algebra of binary relations on a set S involves also the operation of composition. The composite $\rho\sigma = \rho \circ \sigma$ of ρ and σ is defined by the statement:

$x(\rho\sigma)y$ means that, for some z in S, $x\rho z$ and $z\sigma y$.

For functions of a real variable—*i.e.*, when $S = R$, the real field—the definition of composite is equivalent to the usual definition of $\tau = \rho\sigma$, by $\tau(x) = \sigma(\rho(x))$. As another illustration: if the relation ρ (on the set of all people) is defined by "$x\rho y$ if and only if x is a sibling (brother or sister) of y," and σ by "$x\sigma y$ if and only if y is a son of x," then $x(\rho\sigma)y$ means that y is a nephew of x.

The operation of composition is associative: $\rho(\sigma\tau) = (\rho\sigma)\tau$ for any three binary operations on the same set S. The equality relation e (x e y means $x = y$) is an identity for composition: $e\rho = \rho e = \rho$ for all ρ. These facts are summarized in the statement that the algebra $[S^2, \circ]$ defined by the set S^2 of all binary relations on S is a monoid under composition.

Like $\wedge$ and $\vee$, the operation of composition is order-preserving (isotone): if $\rho \geq \rho_1$ and $\sigma \geq \sigma_1$, then $\rho\sigma \geq \rho_1\sigma_1$. Moreover, $\rho(\sigma \vee \tau) = \rho\sigma \vee \rho\tau$ and $(\rho \vee \sigma)\tau = \rho\tau \vee \sigma\tau$, two laws that imply isotonicity. [It is, however, not in general true that $\rho(\sigma \wedge \tau) = \rho\sigma \wedge \rho\tau$ or that $(\rho \wedge \sigma)\tau = \rho\tau \wedge \sigma\tau$.] Lattices that are also monoids and in which the isotonicity relations hold are called lattice-ordered monoids; it has just been shown that the binary relations on any set S form a lattice-ordered monoid.

Vector lattices. Many other kinds of lattice-ordered monoids arise in mathematics, such as the lattice of all ideals of any ring (under the operations of set inclusion and ideal multiplication; see below *Ring theory*). One example is provided by the set of all continuous func-

(14) $f \geqslant g$ means that $f(x) \geqslant g(x)$ for all x.

(15) $h = f + g$ means that $h(x) = f(x) + g(x)$ for all x.

(16) $f + (g \vee h) = (f + g) \vee (f + h)$

(17) $\begin{cases} x = (x_1, x_2, x_3), \quad x \geqslant y = (y_1, y_2, y_3) \\ \sqrt{c_1^2 + c_2^2 + c_3^2} = \text{length of } c \end{cases}$

(18) $\begin{cases} xu = 1(xu) = (vu(xu)) \\ \quad = ((vu)x)u = (v(ux))u = (v \cdot 1)u = v(1u) = vu = 1 \\ x = 1x = (xu)x = x(ux) = x1 \end{cases}$

(19) $\begin{cases} (x + y) + z = x + (y + z) \quad \text{for all } x, y, z \\ x + y = y + x \text{ for all } x, y \end{cases}$

tions f of a real variable x, order being defined (see 14) and the monoid operation being defined as ordinary addition (see 15). Here a previous equation assumes a simpler form (see 16) because addition is commutative; there is no need to write two equations.

Another example is provided by the set of all real vectors (see 17), the order relation meaning that $x_j \geq y_j$ for all $j = 1, 2, 3$. In this the positive cone of all $x \geq O = (0, 0, 0)$ consists of the first octant: it is a trihedron with the positive quadrants of the coordinate planes as faces. For any fixed vector c, the set of all $x \geq c$ is obtained by translating the positive cone parallel to the vector c through a distance equal to its length (see 17).

These and other vector lattices are, by definition, (real) vector spaces that are also lattices, in which vector addition and order are related (see 16) and in which if $\lambda > 0$ is any positive scalar and $x > 0$ any positive vector, $\lambda x > O$.

The preceding examples should give some idea of the enormous variety of lattices and lattice-ordered algebraic structures; they arise in every branch of mathematics.

(G.Bi.)

II. Groups

Groups, in mathematics, are systems of elements with a composition satisfying certain laws (these are given below). The elements may be operations; for example, the rotations of a sphere. The symmetries of a geometrical figure are best described as a group. The ornamental wall designs of the ancient Egyptians exhibit all possible combinations of symmetries. Euclid studied the properties of the regular polygons and the five regular solids. Not until the late 18th and 19th centuries, however, were groups recognized as mathematical systems. A French mathematician, Joseph-Louis Lagrange, was one of the first to consider them. Another French mathematician, Augustin-Louis Cauchy, began a study of permutation groups. In studying the solution of polynomial equations, a Norwegian mathematician, Niels Henrik Abel, showed that in general the equation of fifth degree cannot be solved by radicals. Then the French mathematician Évariste Galois, using groups systematically, showed that the solution of an equation by radicals is possible only if a group associated with the equation has certain specific properties; these groups are now called solvable groups.

Early work on groups

The group concept is now recognized as one of the most fundamental in all of mathematics and in many of its applications. The German mathematician Felix Klein considered geometry to be those properties of a space left unchanged by a certain specific group of transformations. In topology geometric entities are considered equivalent if one can be transformed into another by an element of a continuous group.

DEFINITION

A group, denoted G, is a nonempty set of elements with a composition defined for any ordered pair x, y of its elements. If the composition is written as the product xy, then the following laws hold:

- G1. The Associative Law: $(xy)z = x(yz)$ for all x, y, z, of G.
- G2. Existence of an identity. There is an identity element 1, such that $1x = x1 = x$ for every x of G.
- G3. Existence of an inverse. For every x of G there is an element x^{-1} such that $x^{-1}x = xx^{-1} = 1$, in which 1 is the identity element.

These laws are redundant. Instead of G2 and G3 it is enough to assume the existence of an identity 1 such that $1x = x$ for any x, and that for any x there is a left inverse u such that $ux = 1$. Then let v be the left inverse of u so that $vu = 1$. The desired consequences follow (see 18).

If a group G satisfies the further law:

- G4. The Commutative Law. If $yx = xy$ for all x, y of G, then G is called an Abelian group.

The composition of a group may be written as addition, though this is not usually done except for Abelian groups. In this case the laws take the form:

- G1. Associative Law (see 19).
- G2. Existence of zero. There is a zero element 0 such that its composition with x yields x for every x of G.
- G3. Existence of negative. For every x of G there is an element $-x$ such that its composition with x yields 0.
- G4. Commutative Law (see 19).

Many of the important applications of groups come from the fact that groups arise naturally from one-to-one mappings (sometimes called bijections). A mapping α of a set S into a set T is a rule (or function) that associates with each element x of S a unique element y of T called the image of x. This mapping may be written $y = (x)\alpha$. Of particular importance is the case in which the set T is the same as the set S. The product $\alpha\beta$ of two mappings α and β of a set S into itself is defined by the rule $(x)\alpha\beta = z$ if $(x)\alpha = y$ and $(y)\beta = z$. This definition of product automatically satisfies the associative law. For if $(x)\alpha = y$, $(y)\beta = z$, $(z)\gamma = w$, then $(x)(\alpha\beta)\gamma = (z)\gamma = w$ and $(x)\alpha(\beta\gamma) = (y)\beta\gamma = w$ so that the two mappings $(\alpha\beta)\gamma$ and $\alpha(\beta\gamma)$ are identical and so $(\alpha\beta)\gamma = \alpha(\beta\gamma)$.

Mappings

A mapping α is said to be a mapping of S onto T (rather than into T) providing that for every y of T there is at least one x of S such that $y = (x)\alpha$. Such a mapping is also called a surjection. If the mapping α also has the property that there is exactly one x of S such that $y = (x)\alpha$ for y in T, then α is called a one-to-one mapping or a one-to-one correspondence, or a bijection.

The one-to-one mappings of a set S onto itself form a group. The identity mapping 1 is naturally defined by the rule $(x)1 = x$ for every x of S. And if α is a one-to-one mapping of S onto itself, then an inverse mapping α^{-1} is defined by saying that if $(x)\alpha = y$ then $(y)\alpha^{-1} = x$. For because α is a one-to-one mapping, then given y in S there is exactly one x in S with $(x)\alpha = y$ so that y determines x uniquely. Also $(x)\alpha\alpha^{-1} = x$ for every x and $(y)\alpha^{-1}\alpha = y$ for every y, from which it follows that $\alpha\alpha^{-1} = 1$, $\alpha^{-1}\alpha = 1$.

A nonempty set H of elements with a composition rule xy satisfying the associative law $(xy)z = x(yz)$ is called a semigroup. As has been shown above, the mappings of a set S into itself form a semigroup. In case S consists of the points of a plane, one mapping is to map all the points onto the points of a single line; for example, by mapping a point onto the foot of the perpendicular to the line through the point. Such a mapping does not have an inverse because a point not on the line is not the image of any point. Semigroups are considerably more general than groups and, by the same token, have fewer properties common to all of them.

Semigroups

A group G is said to be finite or infinite if the number of distinct elements in it is respectively finite or infinite. Many properties can be proved for finite groups but are false for infinite groups. The number of distinct elements in a group G is called the order of G.

When two groups can be considered as essentially the same group under certain conditions, these conditions

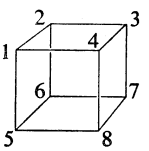
require a precise definition and a name. The groups are said to be isomorphic. The formal definition is that: Two groups G_1 and G_2 are said to be isomorphic if there is a one-to-one correspondence α between elements of G_1 and the elements of G_2, such that if $(x_1)\alpha = x_2$ and $(y_1)\alpha = y_2$, then $(x_1y_1)\alpha = x_2y_2$, in which x_1, y_1 are arbitrary elements of G_1, and x_2, y_2 are their images in G_2. From this definition it is easy to see that the identities of the two groups correspond and that if elements correspond, then also their inverses correspond. In particular, any two groups of order 2 are isomorphic because such a group consists of the identity and one further element the square of which must be the identity. A less obvious example of isomorphism is that between the multiplicative group of positive real numbers and the additive group of all real numbers, the logarithmic function $\log x$ has the property $\log(xy) = \log x + \log y$, and this gives the isomorphism of these two groups.

Isomorphic groups

EXAMPLES OF GROUPS

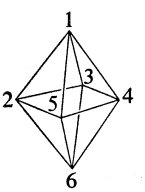
A permutation of a set S is a one-to-one mapping of the elements of S onto themselves. If S consists of the integers $1, 2, \cdots, n$, a permutation α may be written in two rows with a_i below the integer i (see 20) in which this is interpreted to mean that $(i)\alpha = a_i$ for $i = 1, 2, \cdots, n$. As has been observed, the permutations of a set form a group. If S is the set 1, 2, 3, there are 3! (=6) permutations (see 21). Here, for example, the product $\alpha_2\alpha_4 = \alpha_5$ because for $i = 1, 2, 3$ $((i)\alpha_2)\alpha_4 = (i)\alpha_5$ (see 21).

Symmetries of a cube. A rotation taking a cube into itself is fully described by the permutation that this yields



on the 8 vertices. Here a rotation of 90° about an axis through the midpoints of the top and bottom faces is given by a specific permutation (see 22). A rotation about the axis through the vertex 1 and the opposite vertex 7 is given by a specific permutation that holds the integer 7 fixed (see 23). Combinations of α and β give a total of 24 rotations (including, of course, the identity rotation). Apart from these rotations that can be achieved by motions in three-dimensional space, there is also a mirror image reflection about a plane through the midpoints of the edges 14, 23, 67, 58 expressed as a permutation of integers that inverts the order of the first four of eight integers and the last four, similarly (see 24). Combining this reflection with the 24 rotations gives the full group of 48 symmetries of the cube.

Symmetries of the regular octahedron. The midpoints of the faces of the cube may be taken as the vertices of the regular octahedron. For this reason, it is intuitively



clear that the symmetry group of the octahedron is isomorphic to that of the cube. The corresponding symmetries as they affect the vertices of the octahedron are described by three permutations (see 25).

Collineations of the Euclidean plane. If the points of the plane are given by coordinates (x, y) in which x, y are real numbers, the transformations that transform pairs into pairs of linear combinations (see 26) with coefficients a, b, c, d, e, in which $ad - bc$ is not equal to zero, form a group consisting of the one-to-one mappings of the points onto themselves with the property of taking straight lines into straight lines, and this is called the collineation group. The subsystem of collineations β fixing the origin (0, 0) is also a group (see 27). A subsystem of this consists of rotations γ about the origin. The rotation γ corresponding to an angle θ is also expressible as a transformation on pairs, in which coefficients in the linear combinations after transformation depend on θ (see 28).

$$(20)\qquad \alpha = \begin{pmatrix} 1, 2, \cdots, n \\ a_1, a_2, \cdots, a_n \end{pmatrix}$$

$$(21)\qquad \begin{cases} \alpha_1 = \begin{pmatrix} 1,2,3 \\ 1,2,3 \end{pmatrix} & \alpha_4 = \begin{pmatrix} 1,2,3 \\ 1,3,2 \end{pmatrix} \\ \alpha_2 = \begin{pmatrix} 1,2,3 \\ 2,3,1 \end{pmatrix} & \alpha_5 = \begin{pmatrix} 1,2,3 \\ 3,2,1 \end{pmatrix} \\ \alpha_3 = \begin{pmatrix} 1,2,3 \\ 3,1,2 \end{pmatrix} & \alpha_6 = \begin{pmatrix} 1,2,3 \\ 2,1,3 \end{pmatrix} \end{cases}$$

$$(22)\qquad \alpha = \begin{pmatrix} 1,2,3,4,5,6,7,8 \\ 2,3,4,1,6,7,8,5 \end{pmatrix}$$

$$(23)\qquad \beta = \begin{pmatrix} 1,2,3,4,5,6,7,8 \\ 1,4,8,5,2,3,7,6 \end{pmatrix}$$

$$(24)\qquad \gamma = \begin{pmatrix} 1,2,3,4,5,6,7,8 \\ 4,3,2,1,8,7,6,5 \end{pmatrix}$$

$$(25)\qquad \begin{cases} \alpha = \begin{pmatrix} 1,2,3,4,5,6 \\ 1,3,4,5,2,6 \end{pmatrix} \\ \beta = \begin{pmatrix} 1,2,3,4,5,6 \\ 5,1,4,6,2,3 \end{pmatrix} \\ \gamma = \begin{pmatrix} 1,2,3,4,5,6 \\ 1,4,3,2,5,6 \end{pmatrix} \end{cases}$$

$$(26)\qquad (x, y)\alpha = (ax + by + e, cx + dy + f)$$

$$(27)\qquad (x, y)\beta = (ax + by, cx + dy), \qquad ad - bc \neq 0$$

$$(28)\qquad (x, y)\gamma = ((\cos\theta)x + (\sin\theta)y, (-\sin\theta)x + (\cos\theta)y)$$

$$(29)\qquad (xy)(zw) = x(y(zw)) = x((yz)w) = (x(yz))w = ((xy)z)w$$

$$(30)\qquad \begin{cases} (x_1x_2\cdots x_n)^{-1} = x_n^{-1}\cdots x_2^{-1}x_1^{-1} \\ x_1x_2\cdots x_nx_n^{-1}\cdots x_2^{-1}x_1^{-1} = 0 \end{cases}$$

ELEMENTARY GROUP STRUCTURE

In a group G the common value of $(xy)z = x(yz)$ can be written xyz. Similarly with four factors (see 29) and the common value of these products may be written $xyzw$. It is not difficult to prove that any bracketing of $x_1x_2\cdots x_n$ without changing the order of the factors gives the same product. This fact is known as the generalized associative law. A consequence of it is that the inverse of a product is the corresponding product of inverses taken in reverse order (see 30). Similarly, if both the associative and commutative laws hold, a product $x_1x_2\cdots x_n$ has the same value as the product of any rearrangement of these factors.

The generalized associative law

A subset H of the elements of a group G that is itself a group under the composition for G is called a subgroup. In many important cases the elements of a group that have certain properties form a subgroup. For example, the one-to-one mappings of the points of the Euclidean plane onto themselves form a group G. Those mappings that take straight lines into straight lines, the group of collineations, form a subgroup G_1 of G. The group G_1 in turn has a subgroup G_2, called the orthogonal group,

which consists of the one-to-one mappings that leave distances unchanged.

If $a_1, a_2, \cdots, a_r$ are elements of a group G, then all expressions $x_1x_2 \cdots x_n$ in which every value of x is one of the a_i or an inverse a_i^{-1} form a subgroup H that is said to be generated by $a_1, a_2, \cdots, a_r$, and this is written $H = \langle a_1, a_2, \cdots, a_r \rangle$. If H is generated by a single element, $H = \langle a_1 \rangle$, H is called a cyclic group. The product of n factors $a_1a_1 \cdots a_1$ can be written a_1^n and similarly $a_1^{-1}, \cdots, a_1^{-1} = a_1^{-n}$. Then the elements of H are of the form a_1^m with m any integer, positive or negative, and the product rule is $a_1^r a_1^s = a_1^{r+s}$. It may happen that all the powers of a_1 are different, in which case H is essentially the additive group of the exponents that are the integers. If two different powers a_1^r and a_1^s are the same element, then $a_1^{r-s} = 1$. In this case H is finite and there is a positive integer n such that $1, a_1, \cdots, a_1^{n-1}$ are different and $a_1^n = 1$. In this case n is the order of H, and $a_1^r = a_1^s$ if and only if $r - s$ is some multiple of n.

SUBGROUPS

It is not difficult to show that a group G with no subgroup except itself and the identity subgroup must be a finite cyclic group the order of which is a prime number. This describes these groups completely. For a more general group the nature of its subgroups and their interrelations are important in describing the group and its structure.

If H is assumed to be a subgroup of a group G and x some fixed element of G, the set of elements of the form hx with h belonging to H is called a left coset of H and is written Hx. Similarly, the set of elements xh is called a right coset and is written xH. It is easily seen that a left coset Hx contains the same number of elements as the subgroup H because the element hx of Hx can be matched with the element h of H. Similarly a right coset xH has the same number of elements as H. If h_1 and h_2 are any elements of H, and if $x_1 = h_1x$ and $x_2 = xh_2$, then it is easily shown that Hx and Hx_1 are the same set of elements, and also xH and x_2H are the same set of elements. From this it readily follows that any two left (right) cosets of H are identical or disjoint. If $H1 = H$, the subgroup itself and $Hx_2, \cdots, Hx_n$ are all the distinct left cosets of H in G, this fact is represented symbolically by $G = H + Hx_2 + \cdots + Hx_n$. The plus sign has no arithmetical significance here and the expression means that G is the union of the disjoint cosets. For this reason some writers use a different notation. The number of distinct cosets of H in G is called the index of H in G and is written $[G:H]$. The inverses of the elements hx are the elements $x^{-1}h^{-1}$ so that the inverses of the elements in a left coset Hx are precisely the elements of the right coset $x^{-1}H$, because as h runs over all elements of H so also does h^{-1}. Thus even if H is infinite there is a one-to-one correspondence between right and left cosets, and the index $[G:H]$ is the same for right cosets as it is for left cosets. If G is a finite group and the subgroup H is taken to be the identity, then each coset consists of a single element so that the index $[G:1]$ is the order of G; that is, the number of elements in G. A simple but very important consequence of the fact that distinct cosets of a subgroup are disjoint and have the same number of elements is the theorem of Lagrange. It states that the order of a finite group is the product of the order of a subgroup and the index of the subgroup.

The theorem of Lagrange

The theorem of Lagrange is not true for semigroups, illustrating the sharp differences in the theories of groups and semigroups. For a given semigroup S of order n, it is possible to take one additional element u and define $u^2 = u$, $xu = ux = u$ for every x of S, to obtain a semigroup of order $n + 1$ containing S.

From the theorem of Lagrange, the order of a subgroup H of a finite group G divides the order of G. In particular, if the order of G is a prime number p, then G has no subgroups except itself and the identity. As observed above, G must be a cyclic group. The converse of the theorem of Lagrange is false. If G is a finite group of order n and if m is a divisor of n, there may not be a subgroup of G of order m. As a simple example of this, there is a group of order 12 that does not have a subgroup of order 6.

A partial converse to the theorem of Lagrange was proved by the Norwegian mathematician Ludwig Sylow in 1872. In order to describe this theorem fully it is necessary to introduce the concept of conjugate subgroups. If H is a subgroup of the group G and if x is a fixed element of G, then the elements $x^{-1}hx$ in which h ranges over the elements of H form a subgroup that is designated as $x^{-1}Hx$. It is called the conjugate of H by x. There are three main parts to the Sylow theorem. The first part asserts that if p^r is the highest power of the prime p dividing the order of a finite group G, then there is a subgroup P of order p^r, called a Sylow p-subgroup, and that if P_1 is another Sylow p-subgroup, then for some x of G, $P_1 = x^{-1}Px$. Second, the number of distinct groups of order p^r in G is of the form $1 + kp$ for some integers k and is a divisor of the order of G. A third part of Sylow's theorem says that if p^r is the highest power of the prime p dividing the order of G, then there are subgroups of order $p, p^2, \cdots, p^{r-1}$, and each subgroup of order p^i is contained in at least one subgroup of order p^{i+1} for $i = 1, 2, \cdots, r - 1$. These Sylow theorems give a great deal of valuable information on the existence of subgroups of prime power order and their interrelations. For example, if G is of order 12, then it has 1 or 4 subgroups of order 3 and has 1 or 3 subgroups of order 4. Also G has subgroups of order 2, each of which is contained in at least one subgroup of order 4. These facts are not independent. It is not possible for a group G of order 12 to have 3 subgroups of order 4 and also 4 subgroups of order 3.

The mapping $(h)\alpha = x^{-1}hx$ from the subgroup H onto the subgroup $x^{-1}Hx$ is an isomorphism because products are preserved (see 31) and also inverses are preserved (see 32). Thus one of the consequences of the Sylow theorems is that any two Sylow p-subgroups are isomorphic.

(31) $(x^{-1}h_1x)(x^{-1}h_2x) = x^{-1}h_1h_2x$

(32) $(x^{-1}h_1x)^{-1} = x^{-1}h_1^{-1}x$

Of particular importance are those subgroups H of a group G that are conjugate only to themselves; that is, $H = x^{-1}Hx$ for every x of G. This is so important that it is given a special name. A subgroup H of a group G is called a normal subgroup if for every x of G, $H = x^{-1}Hx$. Other names have been used for this property. Normal subgroups have also been called invariant subgroups or self-conjugate subgroups. If the right coset xH contains exactly the same elements as the left coset Hx, then from the equality of the sets $xH = Hx$ it follows that $H = x^{-1}(xH) = x^{-1}Hx$. Conversely, if $H = x^{-1}Hx$, then also $xH = Hx$. Thus H is a normal subgroup of a group G if and only if every right coset xH is also a left coset Hx.

HOMOMORPHISMS; FACTOR GROUPS

It is often possible to map a group G onto a smaller group H, mapping several elements of G onto the same element of H, and to do this so that the product rules are preserved. Such a mapping is called a homomorphism. For example, the real numbers with zero excluded form a group M under multiplication, and mapping every positive number onto $+1$ and every negative number onto -1 gives a homomorphism of M onto the multiplicative group H of the two numbers $+1$ and -1.

A formal definition is: A mapping β of the elements of a group G onto the elements of a group H is called a homomorphism if $(xy)\beta = (x)\beta(y)\beta$ for any x, y of G. It should be noted that in this definition, if it should happen that the mapping β is one-to-one, then β is an isomorphism. Thus an isomorphism is a special case of a homomorphism.

It can be demonstrated that in a homomorphism β the image of the identity of G is the identity of H and that

the image of the inverse of an element is the inverse of the image.

In a homomorphism β of G onto H the set of elements t whose image is the identity of H plays a special role. In particular (see 33) these elements t form a subgroup T. Also for any x of G a simple calculation (see 34) shows that $x^{-1}tx$ is in T, and so $x^{-1}Tx = T$ is a normal subgroup of G. T is called the kernel of the homomorphism. As a result of a calculation (see 35) it is seen that elements in the same coset of T in G have the same image in H, and conversely it is true that elements of G with the same image in H are in the same coset of T in G. As T is a normal subgroup of G, $xT = Tx$. Consequently, considered as sets of elements in G, $TxTy = T(Tx)y = TTxy = Txy$, so that the product of two cosets of T is a definite single coset. This property is not true of subgroups in general. If W is a subgroup of G, then usually the set of elements $WxWy$ make up a number of cosets of W.

If T is a normal subgroup of G, and if $G = T + Tx_2 + \cdots + Tx_n$, the cosets Tx_i may be taken as elements of a new system H. Here if x_ix_j belongs to Tx_k, then it is true that $Tx_iTx_j = Tx_k$, considered as sets of elements of G. This new system H is a group, and H is called the factor group of G with respect to T, and this is written $H = G/T$. If the coset Tx_i is the element h_i of H, there is a homomorphism β of G onto H with T as kernel, and $(tx_i)\beta = h_i$. Thus, given a homomorphism β of G onto a group H, the kernel T of the homomorphism is a normal subgroup of G, and conversely any normal subgroup T of G is the kernel of the homomorphism of G onto the factor group G/T. In this way homomorphisms, normal subgroups, and factor groups go hand in hand.

As an example of a homomorphism, the group G of the 48 symmetries of the cube may be considered. Coordinate axes x, y, z are taken joining the midpoints of opposite faces. Then every symmetry of the cube determines a permutation of x, y, z and every one of the six permutations arises, giving a homomorphic image H of order 6. The kernel T of the homomorphism of G onto H is generated by the reflections in the three coordinate planes. Here T is of order 8 and in terms of coordinates consists of transformations $(x, y, z) \to (\pm x, \pm y, \pm z)$, in which all 8 combinations of the $\pm$ signs arise.

A simple group G is a group that has no proper normal subgroup ("proper" meaning "different from itself or the identity"). The finite cyclic groups of prime order have no proper subgroups and thus are simple groups. There are many other simple groups, the smallest being of order sixty. There have been many investigations on simple groups; some of the results will be discussed later.

The notation $G \supset H$ is used to denote that G contains H as a subgroup. A normal chain for a group G is a sequence (see 36) in which each G_i is a normal subgroup of G_{i-1}. If it is not possible to insert a group T between G_{i-1} and G_i so that T is normal in G_{i-1} for any $i = 1, \cdots, r$, then the chain is said to be a maximal normal chain. In a maximal normal chain every factor group G_{i-1}/G_i, $i = 1, 2, \cdots, r$ is a simple group. A celebrated theorem of Jordan–Hölder states that if a group G has two (or more) maximal normal chains, then the chains have the same number of terms, and the simple factor groups $G_{j-1}{}^*/G_j{}^*$ in a second chain are isomorphic in some order to the simple factor groups G_{i-1}/G_i of the first. This may be considered an analogue of the fundamental theorem of arithmetic, which asserts that the factorization of a positive integer as a product of primes is unique apart from the order of the factors.

The Jordan–Hölder theorem

A field is a system of elements with both the operations of addition and multiplication such that addition is an Abelian group and multiplication of elements (excluding zero) is an Abelian group. Also the distributive law $(x + y)z = xz + yz$ holds. Of particular interest are fields of numbers, such as $\mathfrak{C}$, the field of all complex numbers. The rational numbers form a field usually designated as Q. The numbers $a + b\sqrt{2}$ in which a and b are rational, however, also form a field. If $f(x)$ is a polynomial of degree n with coefficients from the rational field Q, then $f(x)$ has n roots $u_1, u_2, \cdots, u_n$; that is, $f(u_i) = 0$, $i = 1, 2, \cdots, n$, which are numbers in the field $\mathfrak{C}$.

$$(33)\quad \begin{cases}(t_1)\beta = 1, \qquad (t_2)\beta = 1\\ (t_1t_2)\beta = (t_1)\beta(t_2)\beta = 1 \cdot 1 = 1\\ (t_1^{-1})\beta = ((t_1)\beta)^{-1} = (1)^{-1} = 1\end{cases}$$

$$(34)\quad (x^{-1}tx)\beta = ((x)\beta)^{-1} \cdot 1 \cdot (x)\beta = 1$$

$$(35)\quad (tx)\beta = (t)\beta(x)\beta = 1 \cdot (x)\beta = (x)\beta$$

$$(36)\quad G = G_0 \supset G_1 \supset G_2 \cdots G_{r-1} \supset G_r = \{1\}$$

$$(37)\quad w = b_1v + \cdots + b_m \qquad \text{with } b_1, \cdots, b_m$$

$$(38)\quad \begin{cases}(w_1 + w_2)\alpha = (w_1)\alpha + (w_2)\alpha\\ (w_1w_2)\alpha = (w_1)\alpha(w_2)\alpha\end{cases}$$

$$(39)\quad (x_1, x_2)(y_1, y_2) = (x_1y_1, x_2y_2)$$

$$(40)\quad (x_1, 1)(1, x_2) = (x_1, y_2) \in G_1 \times G_2$$

$$(41)\quad (x_1, x_2, \cdots, x_n)(y_1, y_2, \cdots, y_n) = (x_1y_1, x_2y_2, \cdots, x_ny_n)$$

$$(42)\quad (x_1, x_2, \cdots, x_i, \cdots)(y_1, y_2, \cdots, y_i, \cdots) = (x_1y_1, x_2y_2, \cdots, x_iy_i, \cdots)$$

Then with $f(x)$ is associated a field K, in which $K = Q(u_1, \cdots, u_n)$ consists of all numbers obtainable from Q and $u_1, u_2, \cdots, u_n$ by adding, subtracting, multiplying, and dividing. It can be shown that there is one irrational number v and an exponent m such that $v^m = a_1v^{m-1} + \cdots + a_m$ with $a_1, \cdots, a_m$ rational and such that any number w of K has a unique form (see 37) with rational numbers as coefficients of powers of v. Then it is possible to associate with K, and by implication with $f(x)$, a number of mappings $w \to (w)\alpha$ of K onto itself such that sums map into sums and similarly for products (see 38). These mappings form a group G the order of which turns out to be exactly the number m defined above. G is called the group of the field K or respectively of the polynomial $f(x)$. Galois's great contribution to group theory and to the solution of equations consisted in relating this group G to the matter of the possibility of solving for the roots of $f(x)$ by radicals. In a normal chain for G, $G = G_0 > G_1 > \cdots > G_r = 1$ if every factor group G_{i-1}/G_i is a group of some prime order, then $f(x)$ may be solved by radicals, but not otherwise. For this reason such a group is called a solvable group. For an equation of fifth degree G will in general be of order 120 and will have a normal subgroup of order 60, a simple group. In this situation Abel showed that the equation could not be solved by radicals.

DIRECT PRODUCTS; ABELIAN GROUPS

If any two groups G_1 and G_2 are considered, the ordered pairs (x_1, x_2) with x_1 from G_1 and x_2 from G_2 form a group according to the rule that defines products of pairs (see 39). This group is called the direct product of G_1 and G_2 and is written $G_1 \times G_2$. It is easy to see that $G_1 \times G_2$ and $G_2 \times G_1$ are isomorphic. The ordered pairs $(x_1, 1)$ form a subgroup of $G_1 \times G_2$ isomorphic to G_1 and the ordered pairs $(1, x_2)$ a subgroup isomorphic to G_2, so that in this natural way G_1 and G_2 may be considered as subgroups of $G_1 \times G_2$, and because of a simple relation (see 40) the direct product is generated by G_1 and G_2. In the same way, for any finite integer n, given n groups $G_1, G_2, \cdots, G_n$ the finite sequences $(x_1, x_2, \cdots, x_n)$ with x_i from G_i, $i = 1, 2, \cdots, n$, and the product rule (see 41) form the direct product $G_1 \times G_2 \times \cdots \times G_n$ that contains the G_i as subgroups and is generated by them. If there are infinitely many G_i's, it is still possible to define a group by the rule that is suggested by the finite case (see 42)

$$(43)\quad (h_1k_1)(h_2k_2)=h_1h_2(h_2^{-1}k_1h_2)k_2=h_1h_2k_1^{h_2}k_2$$

$$(44)\quad (h_1h_2)^{-1}k(h_1h_2)=h_2^{-1}(h_1^{-1}kh_1)h_2$$

$$(45)\quad (h_1k_1)(h_2k_2)=(h_1h_2)(k_1k_2)$$

$$(46)\quad (h_1k_1)(h_2k_2)=(h_1k_2)(k_1^{h_2}k_2)$$

$$(47)\quad H=\langle h\rangle,\qquad K=\langle k\rangle,\qquad h^4=1,\qquad k^3=1$$

$$(48)\quad h^4=1,\qquad k^3=1,\qquad h^{-1}kh=k^{-1}$$

$$(49)\quad G=H+Hx_2+\cdots+Hx_i+\cdots+Hx_n$$

and the G_i are subgroups of this group G, called the Cartesian product of the G_i. G, however, is no longer generated by the G_i. The subgroup of G generated by the G_i consists of those sequences $(x_1, x_2, \cdots, x_i, \cdots)$ that are the identity in all but a finite number of places, and it is this subgroup that is called the direct product of the G_i. If $G_i = \langle a_i\rangle$ with 2^ith power of a_i equal to 1, $i = 1, 2, \cdots$, then the direct product of the G_i contains only elements of finite order, but the element $a = (a_1, a_2, \cdots, a_i, \cdots)$ is of infinite order in the Cartesian product.

It is easy to see that any relation that holds identically in all groups $G_1, G_2, \cdots$ also holds identically in their direct (or Cartesian) product. In particular, if each of $G_1, G_2, \cdots, G_i, \cdots$ is Abelian and satisfies the identity $yx = xy$, then it is also true that the direct (or Cartesian) product of the G_i is also Abelian.

A fundamental theorem on Abelian groups asserts that a finitely generated Abelian group $G = \langle x_1, \cdots, x_n\rangle$ is the direct product $G = C_1 \times C_2 \times \cdots \times C_m$ of cyclic groups $C_i = \langle c_i\rangle$ in which $m \leq n$. This representation is not unique, because, for example, the cyclic group of order 6 is the direct product of cyclic groups of orders 2 and 3. If the cyclic groups C_i are restricted to be either of infinite order or of finite prime power order p^r, however, then the representation $G = C_1 \times C_2 \times \cdots \times C_m$ is unique apart from the order of the factors. In particular, this result gives a precise description of all finite Abelian groups. This is not to say, however, that all questions of interest are settled with respect to finite Abelian groups.

Infinite Abelian groups

The description of infinite Abelian groups is far from complete. Every subgroup of an Abelian group is normal, but even in an Abelian group the knowledge of a normal subgroup and the factor group does not determine the group. As a very simple example of this, an Abelian group of order p^2 (in which p is a prime) has a normal subgroup of order p and a factor group of order p, but the group may be either the cyclic group of order p^2 or the direct product of two cyclic groups of order p. Infinite Abelian groups are usually written with the group operation as addition and terms are changed accordingly, so that the term direct product is replaced by direct sum.

In an infinite Abelian group G the elements of finite order form a subgroup T that is called the torsion subgroup. Here the factor group G/T contains no elements of finite order except the zero element, and such a group is said to be torsion free. In general it is not possible to express G as the direct sum of the torsion subgroup and a torsion-free subgroup, but most study has been devoted either to torsion groups or torsion-free groups.

An Abelian group D (in additive notation) is said to be divisible if for any y of D and any integer n there is an x of D with $y = nx$. Up to isomorphism there are two basic kinds of divisible groups, first the additive group R^+ of the rational numbers, and $Z(p_\infty)$ the multiplicative group of p^nth roots of 1, in which p is a fixed prime and n ranges over all positive integers. A divisible group D is a direct summand of any Abelian group G containing it, and D is in turn a direct sum of groups each of which is isomorphic to R^+ or some $Z(p_\infty)$.

A primary Abelian group P is a group in which every element has order some power of a fixed prime p. Every torsion group G is the direct sum of its primary subgroups, and this effectively reduces the study of torsion groups to the study of primary groups. Primary groups with at most a countable number of elements have been completely described in terms of certain numerical invariants, but this approach is inadequate for primary groups with more than a countable number of elements. Much less is known on the general theory of torsion-free groups.

AUTOMORPHISMS OF GROUPS; SEMI-DIRECT PRODUCTS

An automorphism α of a group G is a one-to-one mapping of the elements of G onto themselves such that $(g_1g_2)\alpha = (g_1)\alpha(g_2)\alpha$ for any g_1, g_2 of G. This is to say that α is an isomorphism of G with itself. As with any set, the one-to-one mappings form a group, and because the product of two automorphisms is again an automorphism, and the inverse of an automorphism is an automorphism, the automorphisms of G themselves form a group $A(G)$. For a fixed element x of G, the mapping $g \to x^{-1}gx$ for all g of G is an automorphism called an inner automorphism, which is designated as α_x. It is easy to check that $\alpha_x\alpha_y = \alpha_{xy}$ and that $(\alpha_x)^{-1}$ is the inner automorphism associated with x^{-1}. Here $\alpha_x = 1$ if, and only if, $x^{-1}gx = g$ or $gx = xg$ for every g of G. The elements with this property of commuting with every element of G themselves form a subgroup called the centre $Z(G)$ of G. Clearly $Z(G)$ is an Abelian group and is normal in G. The inner automorphisms form a group $I(G)$ that is isomorphic to $G/Z(G)$. $I(G)$ is a normal subgroup of $A(G)$, and the factor group $A(G)/I(G)$ is called the group of outer automorphisms of G. If G is Abelian, then $Z(G) = G$ and $I(G) = 1$, so that all automorphisms of an Abelian group are outer automorphisms.

If K is a normal subgroup of a group G, then for any x belonging to G, $x^{-1}Kx = K$, and the mapping $k \to x^{-1}kx$ for k in K is an automorphism of K. It may happen that in $G = K + x_2K + \cdots + x_nK$, it is possible to choose the coset representatives $1, x_2, \cdots, x_n$ so that they form a subgroup H. In this case every element h of H is associated with an automorphism $k \to h^{-1}kh$ of K, and it is convenient to use an exponential notation, writing $h^{-1}kh = k^h$. Then every element x of G has a unique expression in the form $x = hk$ with h from H and k from K. An identity (see 43) gives the product rule for G in terms of the product rule in H, the product rule in K, and the automorphisms of K associated with elements of H. The only requirement on the automorphisms comes from the identity (see 44), which in exponential notation takes a familiar form. This means that the group H is mapped homomorphically onto a group of automorphisms of K. If the automorphism $k \to k^h$ is always the identity automorphism, *i.e.*, $k^h = k$ in every case, then the multiplication rule takes a form (see 45) so that G is the direct product $H \times K$. In general the product rule (see 46), a slight modification of those considered, and subject to the exponential condition given above, determines a group G known as the semi-direct product of K by H, in which K is a normal subgroup of G, and H is isomorphic to the factor group G/K. This is a useful rule that gives the construction for many groups. An easy example is that in which the fourth power of h and the third power of k are 1 (see 47) and h and h^3 correspond to the automorphism $k \to k^{-1}$ while 1 and h^2 correspond to the automorphism $k \to k$. Here $G = HK$ is of order 12 and is fully determined by three particular relations (see 48).

PERMUTATION GROUPS

If G is a permutation group on a set $S = \{1, 2, \cdots, n\}$, then G is said to be transitive on S if for every i and j of S there is a permutation x of G for which $(i)x = j$. It is easy to see that if G is transitive, those permutations h that fix 1—*i.e.*, $(1)h = 1$—form a subgroup H and that G is analyzed as a sum of terms of the type Hx_i (see 49), in which x_i is a permutation such that $(1)x_i = i$. Thus, in a transitive group G on symbols $1, 2, \cdots, n$, each symbol corresponds to a coset of a subgroup. Conversely, given

a group G and a subgroup H, in which $G = H + Hx_2 + \cdots + Hx_n$, there is a mapping of the elements of H onto permutations of 1, 2, $\cdots$, n corresponding to the cosets. In general this mapping is a homomorphism and its kernel K is the largest normal subgroup of G contained in H.

If H is the identity subgroup, then each coset consists of a single element and for each element x of G there is a permutation α_x such that $(y)\alpha_x = yx$ for y belonging to G. The correspondence between x and α_x is easily seen to be an isomorphism so that every G is isomorphic to a permutation group of its own elements. Thus every group may be considered to be a permutation group.

Define the polynomial $\Delta(x_1, \cdots, x_n)$ in terms of products (see 50), this being the product of all $x_i - x_j$ with i less than j. Every permutation of $x_1, \cdots, x_n$ replaces Δ by Δ or $-\Delta$, and the permutation that interchanges x_1 and x_2 and fixes the rest takes Δ into $-\Delta$. Those permutations that take Δ into Δ form a group A_n, called the alternating group. The group of all $n!$ permutations on $x_1, \cdots, x_n$ is called the symmetric group S_n. Here the group A_n consists of half the elements of S_n and is of order $n!/2$.

The permutations that leave a given array invariant will form a group. The array of seven triples has the property that every pair of elements occurs in exactly one triple. It can be considered as the set of lines in a finite plane (see 51).

Four permutations leave the array invariant (see 52), and they are symbolized by x, y, z, and w.

Because the group G is transitive on 1, $\cdots$, 7, the subgroup H is of index 7, $[G:H] = 7$. The subgroup H containing y and z is transitive on 2, $\cdots$, 7, and so the subgroup T of H that fixes 2 as well as 1 is of index 6 in H, $[H:T] = 6$. T is generated by z and w and is of order 4, $[T:1]$. Hence G is of order 168 as a computation shows (see 53).

Given any two distinct numbers i, j and two other distinct numbers i^*, j^*, there is an element u of G such that $(i)u = i^*$ and $(j)u = j^*$. For this reason G is said to be doubly transitive. In general, a permutation group G is said to be k-ply transitive if given k distinct numbers $i_1, i_2, \cdots, i_k$ and another set of k distinct numbers $i_1^*, i_2^*, \cdots, i_k^*$, then there is an element u in G such that u post-multiplies each number of the first set to give another in the second (see 54).

The symmetric group S_n is n-ply transitive and the alternating group A_n is n-2-ply transitive. There are many triply transitive permutation groups, but apart from the groups A_n and S_n, there are only four permutation groups known that are as much as quadruply transitive. These are known as the Mathieu groups after the French mathematician Émile-Léonard Mathieu who discovered them in 1860 and 1873. These groups are M_{11}, of order 7920 quadruply transitive on 11 letters, M_{12} of order 95040 quintuply transitive on 12 letters, M_{23} of order 10,200,960 quadruply transitive on 23 letters, and M_{24} of order 244,922,040 quintuply transitive on 24 letters.

GROUP REPRESENTATION

It has been remarked earlier that transformations $(x, y) \rightarrow (ax + by, cx + dy)$ in which $ad - bc \neq 0$ form a group. The transformation value $(x, y)\alpha$ is a pair of real numbers which may be called $((x, y)\alpha)_1$ and $((x, y)\alpha)_2$. This transformation may then be written in the form of two equations (see 55). As an example, take transformations α and β in which the corresponding equations are of the above type (see 56). The group $G = \langle \alpha, \beta \rangle$ is a finite group of order 6 in which the basic properties are easily found (see 57).

Linear transformations of n variables that are one-to-one (and so have inverses) will form a group. If the symbol x denotes a sequence $(x_1, \cdots, x_n)$ of numbers, then x_j is the jth term of x. A transformation α taking n-term sequences into n-term sequences can be formed in a fashion similar to the two-term use above. Then $(x)\alpha$ is again a sequence with terms $((x)\alpha)_1, \cdots, ((x)\alpha)_n$. A linear transformation is a transformation α that can be defined by a linear system of equations (see 58). If another transformation β is expressed the same way (see 59), then the product $\alpha\beta$ of these transformations is given by another set of n linear equations (see 60) in which in every case the coefficient c_{ij} is given by the rule that identifies a typical coefficient in the equations of the product transformation with sums of products of coefficients in the equations expressing α and β (see 61).

Clearly these transformations are determined by the array of coefficients. The square array that results (see 62) is called a matrix, and a product rule is defined, such

(50) $$\Delta = (x_1 - x_2)(x_1 - x_3) \cdots (x_1 - x_n)(x_2 - x_3) \cdots (x_{n-1} - x_n)$$

(51)
$$\begin{cases} 1, 2, 4 \\ 2, 3, 5 \\ 3, 4, 6 \\ 4, 5, 7 \\ 5, 6, 1 \\ 6, 7, 2 \\ 7, 1, 3 \end{cases}$$

(52)
$$\begin{cases} x = \begin{pmatrix} 1,2,3,4,5,6,7 \\ 2,3,4,5,6,7,1 \end{pmatrix} \\ y = \begin{pmatrix} 1,2,3,4,5,6,7 \\ 1,3,5,7,2,4,6 \end{pmatrix} \\ z = \begin{pmatrix} 1,2,3,4,5,6,7 \\ 1,2,5,4,3,7,6 \end{pmatrix} \\ w = \begin{pmatrix} 1,2,3,4,5,6,7 \\ 1,2,6,4,7,3,5 \end{pmatrix} \end{cases}$$

(53) $$[G:1] = [G:H][H:T][T:1] = 7 \cdot 6 \cdot 4$$

(54) $$(i_1)u = i_1^*, \quad \cdots, \quad (i_k)u = i_k^*$$

(55)
$$\begin{cases} ((x, y)\alpha)_1 = ax + by \\ ((x, y)\alpha)_2 = cx + dy \end{cases}$$

(56)
$$\begin{cases} ((x, y)\alpha)_1 = y, \quad ((x, y)\beta)_1 = y \\ ((x, y)\alpha)_2 = -x - y, \quad ((x, y)\beta)_2 = x \end{cases}$$

(57) $$\alpha^3 = 1, \qquad \beta^2 = 1, \qquad \beta\alpha\beta = \alpha^{-1}$$

(58)
$$\begin{cases} ((x)\alpha)_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1j}x_j + \cdots + a_{1n}x_n \\ ((x)\alpha)_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2j}x_j + \cdots + a_{2n}x_n \\ ((x)\alpha)_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{ij}x_j + \cdots + a_{in}x_n \\ ((x)\alpha)_n = a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nj}x_j + \cdots + a_{nn}x_n \end{cases}$$

(59)
$$\begin{cases} ((x)\beta)_1 = b_{11}x_1 + b_{12}x_2 + \cdots + b_{1j}x_j + \cdots + b_{1n}x_n \\ ((x)\beta)_2 = b_{21}x_1 + b_{22}x_2 + \cdots + b_{2j}x_j + \cdots + b_{2n}x_n \\ ((x)\beta)_i = b_{i1}x_1 + b_{i2}x_2 + \cdots + b_{ij}x_j + \cdots + b_{in}x_n \\ ((x)\beta)_n = b_{n1}x_1 + b_{n2}x_2 + \cdots + b_{nj}x_j + \cdots + b_{nn}x_n \end{cases}$$

(60)
$$\begin{cases} ((x)\alpha\beta)_1 = c_{11}x_1 + c_{12}x_2 + \cdots + c_{1j}x_j + \cdots + c_{1n}x_n \\ ((x)\alpha\beta)_2 = c_{21}x_1 + c_{22}x_2 + \cdots + c_{2j}x_j + \cdots + c_{2n}x_n \\ ((x)\alpha\beta)_i = c_{i1}x_1 + c_{i2}x_2 + \cdots + c_{ij}x_j + \cdots + c_{in}x_n \\ ((x)\alpha\beta)_n = c_{n1}x_1 + c_{n2}x_2 + \cdots + c_{nj}x_j + \cdots + c_{nn}x_n \end{cases}$$

(61) $$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj}, \qquad i, j = 1, 2, \cdots, n$$

(62) $$A = \begin{bmatrix} a_{11}, a_{12}, \cdots, a_{1n} \\ a_{n1}, a_{n2}, \cdots, a_{nn} \end{bmatrix} = [a_{ij}], \qquad i, j = 1, 2, \cdots n$$

that ordinary matrix multiplication (see 63) applies in every instance (see 64). The matrices, six in number, with elements composed of -1, 0, and 1 (see 65) and containing the identity matrix, as well as two matrices of specific structure, correspond to the group of order 6, $G = \langle \alpha, \beta \rangle$ mentioned above, and themselves form an isomorphic group.

A matrix representation, or, more briefly, representation, of a group G is a homomorphism of G onto a subgroup of the group of matrices of dimension n that have inverses. The elements in the matrices may be complex numbers or may be from any field.

As an example of matrix representation, the symmetries of a cube given above should be considered. The vertices may be given coordinates composed of 1 and -1 (see 66) so that the cube is symmetric about the origin. The symmetries α, β, γ take the form of matrices of nine elements each (see 67), composed of the numbers -1, 0, 1.

The determinant of a matrix A, written det A, is a function of the entries of A that can be defined recursively in terms of the dimension n of A. If $n = 1$ and $A = [a_{11}]$, then det $A = a_{11}$. For n greater than one, the det A has an expansion in det A_{1j} (see 68), in which for $j = 1, 2, \cdots, n$, A_{1j} is the matrix of dimension $n - 1$ obtained from A by deleting the first row and jth column. Thus for $n = 2$, $\det A = a_{11}a_{22} - a_{12}a_{21}$. Thus the determinant of a matrix the entries of which are in a field F is an element of that field. The most useful property of determinants is the fact that $\det (AB) = \det A \cdot \det B$. The identity transformation 1, with $(x)1 = x$ for all x, corresponds to the identity matrix I, and if $I = [e_{ij}]$, then $e_{11} = e_{22} \cdots = e_{nn} = 1$ and $e_{ij} = 0$ if $i \neq j$. In this det $(I) = 1$, and a matrix A has an inverse if and only if $\det A \neq 0$.

A useful operation on a matrix is called the transpose. If $A = [a_{ij}]$, the transpose A^T is defined as $A^T = [b_{ij}]$, in which $b_{ij} = a_{ji}$.

Matrices may also be added if when A and B are expressed in terms of their elements (see 69) the sum $A + B = C$ is defined by addition of corresponding elements (see 70). If u is an element of the field F, the scalar product uA is defined as the matrix $[ua_{ij}]$. Matrix addition is associative and commutative; matrix multiplication is associative but not commutative ($(AB)C = A(BC)$, but not $BA = AB$). Multiplication is left and right distributive over addition (see 71).

A representation M of a group G is given by matrices $M(x)$ for x in G such that the matrix corresponding to 1 is the identity and the matrix corresponding to a product of elements in the group is a product of corresponding matrices (see 72). If R is any fixed matrix with an inverse R^{-1}, then $R^{-1}M(x)R$ also gives a representation that is considered equivalent to the representation given by $M(x)$. If it happens that for every x of G $M(x)$ is of the form of four blocks, upper left being $M_1(x)$, lower left being $M_{12}(x)$, and lower right being $M_2(x)$ (see 73), with $M_1(x)$ of dimension r, $M_2(x)$ of dimension s, and the block [0] in the upper right corner consisting entirely of zeros, it is said that M is reducible, and if also $M_{12}(x)$ is identically zero it is said that M is the sum of M_1 and M_2. If no representation equivalent to M is reducible, then M is said to be irreducible.

For representations of all finite groups over the complex numbers and certain classes of infinite groups, it is true that, up to equivalence, there are only a finite number of irreducible representations and every representation is equivalent to the sum of a number of irreducible representations.

If $M(x) = A = [a_{ij}]$, then the quantity $a_{11} + a_{22} + \cdots + a_{nn}$ is called the trace of A, written tr A, and it is known that A and $R^{-1}AR$ have the same trace. If $M(x) = A$, then tr A is called the character of x and may be denoted as $\chi(x)$. Here a representation M of a group G is associated with a character $\chi(x)$. If M is the sum of M_1 and M_2, then obviously $\chi(x) = \chi_1(x) + \chi_2(x)$ for the corresponding traces. From this it follows that, over the complex numbers, the representations of all finite groups (and certain infinite groups) are completely determined by the characters of the irreducible representations. The nature of these characters and the relations on them have been the subject of much investigation, and they have been a very valuable technical tool in the study of groups. Some of these relations are not obviously related to the groups themselves and appear to be a sort of "numerology" or "astrology" of groups.

(63) $$\begin{cases} B = [b_{ij}], & i,j = 1, 2, \cdots, n \\ AB = C = [c_{ij}], & i,j = 1, 2, \cdots, n \end{cases}$$

(64) $$\begin{cases} c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj} \\ i,j = 1, 2, \cdots, n \end{cases}$$

(65) $$\begin{cases} 1 = \begin{bmatrix} 1, & 0 \\ 0, & 1 \end{bmatrix} \quad A = \begin{bmatrix} 0, & 1 \\ -1, & -1 \end{bmatrix} \quad A^2 = \begin{bmatrix} -1, & -1 \\ 1, & 0 \end{bmatrix} \\ B = \begin{bmatrix} 0, & 1 \\ 1, & 0 \end{bmatrix} \quad AB = \begin{bmatrix} 1, & 0 \\ -1, & -1 \end{bmatrix} \quad A^2B = \begin{bmatrix} -1, & -1 \\ 0, & 1 \end{bmatrix} \end{cases}$$

(66) $$\begin{cases} 1 = (-1, 1, -1) & 5 = (-1, 1, 1) \\ 2 = (-1, -1, -1) & 6 = (-1, -1, 1) \\ 3 = (1, -1, -1) & 7 = (1, -1, 1) \\ 4 = (1, 1, -1) & 8 = (1, 1, 1) \end{cases}$$

(67) $$\alpha = \begin{bmatrix} 0, & 1, & 0 \\ -1, & 0, & 0 \\ 0, & 0, & 1 \end{bmatrix} \quad \beta = \begin{bmatrix} 0, & -1, & 0 \\ 0, & 0, & -1 \\ 1, & 0, & 0 \end{bmatrix} \quad \gamma = \begin{bmatrix} -1, & 0, & 0 \\ 0, & 1, & 0 \\ 0, & 0, & 1 \end{bmatrix}$$

(68) $$\det A = a_{11} \det A_{11} - a_{12} \det A_{12} + \cdots + (-1)^{j+1} a_{1j} \det A_{ij} + \cdots + (-1)^{n+1} a_{1n} \det A_{1n}$$

(69) $$A = [a_{ij}], \quad B = [b_{ij}], \quad i,j = 1, 2, \cdots, n$$

(70) $$c_{ij} = a_{ij} + b_{ij}, \quad i,j = 1, \cdots, n$$

(71) $$\begin{cases} A + B = B + A \\ (A + B) + C = A + (B + C) \\ A(B + C) = AB + AC \\ (A + B)C = AC + BC \end{cases}$$

(72) $$M(1) = I \quad \text{and} \quad M(x_1 x_2) = M(x_1)M(x_2)$$

(73) $$M(x) = \left[\begin{array}{c|c} M_1(x) & 0 \\ \hline M_{12}(x) & M_2(x) \end{array}\right]$$

(74) $$(AB)^T = B^T A^T \quad \text{and} \quad \det(A^T) = \det A$$

THE CLASSICAL GROUPS; CONTINUOUS GROUPS

The set of all matrices of dimension n over a field F the determinant of which is not zero forms a group called the general linear group of dimension n and is designated as $GL_n(F)$. Because determinants satisfy $\det (AB) = \det A \cdot \det B$, it follows that all matrices of determinant unity, $\det A = 1$, form a subgroup of $GL_n(F)$ known as the special linear group of dimension n and designated $SL_n(F)$. The transpose operation on matrices has the property that it permutes the order of multiplication when applied to a product and leaves the determinant of a matrix unchanged (see 74). From this it follows easily that matrices X with the property $X^TX = I$, I being the identity matrix, form a group, and this is called the orthogonal group $O_n(F)$. The orthogonal group is also the group of linear transformations that leave $x_1^2 + x_2^2 + \cdots + x_n^2$ invariant. For matrices of even dimension $n = 2t$, the particular matrix composed of four blocks, 0's being on the diagonal, and I_t and its negative

being in the other two blocks (see 75), is of importance in which I_t is the t-dimensional identity matrix and 0 stands for a block consisting entirely of zeros. A has the property $A^T = -A$. Those matrices X such that $X^TAX = A$ form a group known as the symplectic group, designated as $Sp_n(F)$ with $n = 2t$.

If the field F is the field $\mathbb{C}$ of complex numbers, the mapping $a + bi \to a - bi$ is an automorphism of order 2 and can be written $x \to x^*$. Here if x is a complex number $a + bi$, then x^* is the conjugate $a - bi$. In general, if a field F has an automorphism $x \to x^*$ of order 2, those matrices X such that $Y^TX = I$ form a group, in which Y is obtained from X by replacing every entry x_{ij} by its conjugate x^*_{ij}. This group is called the unitary group $U_n(F)$. The unitary group is a generalization of the orthogonal group and may be considered as the group of linear transformations leaving a sum of products of n complex numbers times their conjugates invariant (see 76).

The general linear group, the special linear group, the orthogonal, symplectic, and unitary groups are known as the classical groups.

Only a little will be said here of continuous groups because they are discussed in the article TOPOLOGICAL GROUPS AND DIFFERENTIAL TOPOLOGY. In some groups G it is possible to define a distance between elements of the group. For example, if G consists of the rotations of the surface of a sphere, and if x and y are two of these rotations, the distance between x and y, $d(x, y)$ can be defined as the maximum distance between $(p)x$ and $(p)y$ as p ranges over all points of the sphere. In such a case a neighbourhood of a group element x is defined as the set of all elements y such that $d(x, y) < e$ for some fixed e. Neighbourhoods may also be defined in other ways. The group G is said to be continuous if the product operation xy in G is continuous either in terms of the distance defined or in terms of the neighbourhoods. In some instances continuous groups may be given in terms of continuous parameters (variables that distinguish the elements of the group). For example, the group of orthogonal matrices (an orthogonal matrix is a matrix in which the columns are orthogonal vectors; see ALGEBRA, LINEAR AND MULTILINEAR) over the real field that have determinant one can be parameterized in terms of a number of angles of rotation, and it is a continuous group in terms of these parameters. Continuous groups expressed in terms of parameters are known as Lie groups, after Sophus Lie, who studied them at great length.

FREE GROUPS; THE WORD PROBLEM; THE BURNSIDE PROBLEM

If $a_1, a_2, \cdots, a_r$ are elements of a group G, then $H = \langle a_1, a_2, \cdots a_r\rangle$ is said to be the subgroup generated by $a_1, \cdots a_r$ and consists of all finite words, products of the elements $x_1x_2, \cdots x_t$, in which every x_j is some a_i or a_i^{-1}, and also the identity 1 that may be considered the empty word. If x_jx_{j+1} is of the form $a_ia_i^{-1}$ or $a_i^{-1}a_i$, then the same group element is given by $x_1x_2 \cdots x_{j-1}x_{j+2} \cdots x_t$. If no x_jx_{j+1} is of the form $a_ia_i^{-1}$ or $a_i^{-1}a_i$, the word $x_1x_2 \cdots x_t$ is said to be reduced. By deleting consecutive inverses every word is equal to a reduced word, including 1 as a reduced word. It can be shown that the order in which these deletions are carried out does not matter and that in this way every word is equal to a unique reduced word. It can be proved that given generators $a_1, a_2, \cdots, a_r$, there is a unique group in which all the distinct reduced words are distinct elements. This group is called the free group F_r-generated by $a_1, a_2, \cdots, a_r$, and this is written $F_r = \langle a_1, a_2, \cdots, a_r\rangle$.

It is not difficult to show that if G is any group generated by r elements $G = \langle b_1, b_2, \cdots, b_r\rangle$, then G is a homomorphic image of $F_r = \langle a_1, \cdots, a_r\rangle$ under a homomorphism β with $(a_1)\beta = b_1, \cdots, (a_r)\beta = b_r$. The kernel K of this homomorphism consists of those words in the a's that become the identity in G when the a's are replaced by b's. K is, of course, a normal subgroup of F_r and, of course, any normal subgroup of F_r is the kernel of an appropriate homomorphism. Thus every group G is a factor group of a free group F. It is a remarkable fact that every subgroup of a free group is also a free group on an appropriate set of generators.

$$(75)\quad A = \left[\begin{array}{c|c} 0 & I_t \\ \hline -I_t & 0 \end{array}\right]$$

$$(76)\quad x_1\bar{x}_1 + x_2\bar{x}_2 + \cdots + x_n\bar{x}_n$$

$$(77)\quad F,\ W^{-1}x^4W,\ W^{-1}y^3W,\ W^{-1}x^4yxyW$$

$$(78)\quad R_1(b_1, \cdots, b_r) = 1, \cdots$$

$$(79)\quad R_1(a_1, \cdots, a_r), \cdots, R_n(a_1, \cdots, a_r)$$

$$(80)\quad N(r) = r + r(r-1)/2 + r(r-1)(r-2)/6$$

Free groups give a precise basis for describing defining relations on groups. To say that a group G of order 12 is defined by relations $h^4 = 1$, $k^3 = 1$, and $h^{-1}kh = k^{-1}$, is equivalent to taking the free group $F_2 = \langle x, y\rangle$ and the least normal subgroup K of F containing x^4, y^3 and $x^{-1}yxy$, and defining G as F_2/K, mapping $(x)\beta = h$, $(y)\beta = k$. Here, as W ranges over all words of F, terms of the type $W^{-1}UW$, in which U is one of the three listed elements of K (see 77), are elements of K.

The word problem in groups involves the case in which G is generated by a finite number of elements $b_1, b_2, \cdots, b_r$ and that these satisfy a finite number of relations (see 78) and no other relations not consequences of these. Is it possible to decide whether or not two different words W_1 and W_2 represent the same element of G, or equivalently whether or not $W = W_1W_2^{-1}$ is the identity? In terms of free groups this is the question of deciding whether or not in the free group $F_r = F_1(a_1, \cdots, a_r)$ a word W is contained in the least normal subgroup K of F_r that contains the finite number of functions (see 79) that were used to express the relations.

The word problem is in fact unsolvable. No method can exist that will in general solve all word problems. This was proved in 1955 by the Soviet mathematician Pyotr Sergeyevich Novikov. In 1959 the U.S. mathematician William W. Boone exhibited a particular set of relations $R_1, \cdots, R_n$ for which the word problem is unsolvable. This unsolvability has been a major difficulty in various natural problems of group theory.

In 1902 the British mathematician William Burnside raised the following question. If a group G is generated by r elements, $G = \langle b_1, \cdots, b_r\rangle$, and if for a fixed n, $x^n = 1$ for every element of G, is G necessarily finite? If $F_r = \langle a_1, \cdots, a_r\rangle$ is the free group with r generators and if $K = K_n$ is the subgroup of F generated by x^n for every x of F, then K is necessarily a normal subgroup of F, and $B(n, r) = F_r/K_n$ is a group with r generators and $x^n = 1$ for every element of $B(n, r)$. Any group such as G is necessarily a homomorphic image of $B(n, r)$, and so for given n and r the question is whether or not the group $B(n, r)$ is finite. If r is 1, then $B(n, r) = B(n, 1)$ is the cyclic group of order n. Burnside showed that for $n = 2$, $B(2, r)$ is of order 2^r and is the direct product of r groups of order 2. He also showed that for $n = 3$ the groups are finite, and that for $n = 4$, $r = 2$, $B(4, 2)$ is of order at most $2^{12} = 4096$. In 1933 the mathematicians F. Levi and B.L. van der Waerden found the exact order and structure of the groups $B(3, r)$. The order of $B(3, r)$ is $3^{N(r)}$, in which the exponent of 3 is a specific algebraic function of r (see 80).

In 1940 I.N. Sanov, a mathematician, showed that the groups $B(4, r)$ are finite. It has been shown that Burnside was right about $B(4, 2)$ and that its order is exactly $2^{12} = 4096$, but for $r = 3$ or greater the order is not known. A mathematician, C. Wright, in 1962 found more reasonable bounds on the order and structure of $B(4, r)$ than Sanov had found, but precise information was still lacking in the 1970s.

In 1958 it was shown by the U.S. mathematician Mar-

(81, 82)
$$\begin{cases} GF(4),\ 1+1=0, & u+u=0 \\ u^2=u+1, & u(u+1)=1 \end{cases}$$

(83) Mathieu groups:
$$M_{11}\ \ 2^4\cdot 3^2\cdot 5\cdot 11$$
$$M_{12}\ \ 2^6\cdot 3^3\cdot 5\cdot 11$$
$$M_{22}\ \ 2^7\cdot 3^2\cdot 5\cdot 7\cdot 11$$
$$M_{23}\ \ 2^7\cdot 3^2\cdot 5\cdot 7\cdot 11\cdot 23$$
$$M_{24}\ \ 2^{10}\cdot 3^3\cdot 5\cdot 7\cdot 11\cdot 23$$

(84)
$$2^3\cdot 3\cdot 5\cdot 7\cdot 11\cdot 19 \quad \text{Janko}$$
$$2^7\cdot 3^3\cdot 5^2\cdot 7 \quad \text{Hall-Janko}$$
$$2^9\cdot 3^2\cdot 5^3\cdot 7\cdot 11 \quad \text{Higman-Sims}$$
$$2^{13}\cdot 3^7\cdot 5^2\cdot 7\cdot 11\cdot 13 \quad \text{Suzuki}$$
$$2^7\cdot 3^6\cdot 5^3\cdot 7\cdot 11 \quad \text{McLaughlin}$$
$$2^7\cdot 3^5\cdot 5\cdot 17\cdot 19 \quad \text{Higman-McKay-Janko}$$
$$2^{10}\cdot 3^3\cdot 5^2\cdot 7^3\cdot 17 \quad \text{Held}$$

(85)
$$2^{21}\cdot 3^9\cdot 5^4\cdot 7^2\cdot 11\cdot 13\cdot 23 \quad \text{Conway .1}$$
$$2^{18}\cdot 3^6\cdot 5^3\cdot 7\cdot 11\cdot 23 \quad \text{Conway .2}$$
$$2^{10}\cdot 3^7\cdot 5^3\cdot 7\cdot 11\cdot 23 \quad \text{Conway .3}$$
$$2^{17}\cdot 3^9\cdot 5^2\cdot 7\cdot 11\cdot 13 \quad \text{Fischer } F_{22}$$
$$2^{18}\cdot 3^{13}\cdot 5^2\cdot 7\cdot 11\cdot 13\cdot 17\cdot 23 \quad \text{Fischer } F_{23}$$
$$2^{22}\cdot 3^{16}\cdot 5^2\cdot 7^3\cdot 11\cdot 13\cdot 17\cdot 23\cdot 29 \quad \text{Fischer } F_{24}$$

(86)
$$2^8\cdot 3^7\cdot 5^6\cdot 7\cdot 11\cdot 31\cdot 37\cdot 67 \quad \text{Lyons-Sims}$$
$$2^{14}\cdot 3^3\cdot 5^3\cdot 7\cdot 13\cdot 29 \quad \text{Rudvalis-Conway-Wales}$$

shall Hall, Jr., that the groups $B(6, r)$ are finite; and from results of the mathematicians P. Hall and G. Higman, it follows that the order and structure of $B(6, r)$ are fully known. $B(6, r)$ is of order $2^A 3^B$ in which $A = 1 + (r-1)3^{N(r)}$ and with $s = 1 + (r-1)2^r$, $B = N(s)$, in which the exponent of 3 is as above (see 80).

In 1968 it was shown by the mathematicians Novikov and Adjan that $B(n, r)$ is infinite if r is at least 2 and n is odd and at least 4381.

SIMPLE GROUPS

A simple group is a group that has no normal subgroup except itself and the identity subgroup. The Jordan–Hölder theorem shows that these are the building blocks for all groups. Many important questions about groups in general can easily be reduced to problems on simple groups.

The nature of the simple groups remains a challenging and difficult matter. In one case the answer was found by the French mathematician Élie-Joseph Cartan in 1894. For groups that are continuous in terms of real or complex parameters, the Lie groups, a complete list has been found. The classical groups will in general have a centre Z, consisting of those scalar multiples of the identity that are in the group, and in the case of the orthogonal group there may be a normal subgroup with an Abelian factor group. Such a normal subgroup up to certain identification with its centre (that is, modulo its centre) will be in general a simple group. There are four infinite families, A_t corresponding to the special linear group of dimension $t+1$ modulo its centre, B_t and D_t coming from orthogonal groups respectively of dimensions $2t+1$ and $2t$, and C_t coming from another group of dimension $2t$, called symplectic.

There are also five exceptional groups, G_2, F_4, E_6, E_7, and E_8, which can be expressed with respectively 14, 52, 78, 133, and 248 parameters. The unitary groups are not technically a part of this list because complex conjugation is part of their definition.

For finite simple groups the situation is much more complicated and is currently the subject of intense research effort. To describe the situation it is first necessary to describe finite fields. Let p be a prime number. It is said that integers r and s are congruent modulo p if the difference $r - s$ is a multiple of p, and this is written $r \equiv s \pmod p$. It is also said that r and s are in the same residue class modulo p, and that either r or s is a representative of the residue class. Hence $0, 1, 2, \cdots, p-1$ are representatives of all p residue classes. If $r \equiv s \pmod p$ and $t \equiv u \pmod p$, then $r + t \equiv s + u \pmod p$ and $rt \equiv su \pmod p$. The p residue classes form a field, written $GF(p)$. If $q = p^r$ is a power of the prime p, there is always a unique field with q elements, $GF(q)$, the elements of which are polynomials of a single element with coefficients from $GF(p)$. Thus $GF(4)$ has elements 0, 1, u, $u+1$, in which 0, 1 are the residue classes modulo 2. In $GF(4)$, four basic relations hold (see 81, 82) and other relations are consequences of these.

The mathematician Claude Chevalley showed in 1955 that all of the Lie groups have finite analogues if the field of real or complex numbers is replaced by any finite field, and that with a very few exceptions, corresponding to fields with 2 or 3 elements, the analogues of the simple Lie groups are simple. This is also true for the unitary groups. Somewhat like the unitary group, there are several families known as twisted Lie groups corresponding to subgroups of the Lie types fixed by a certain automorphism. For instance, if $q = 2^{2m+1}$ is at least 8, there are simple groups of order $(q^2+1)q^2(q-1)$, first discovered by the mathematician M. Suzuki in 1960 in another way. The finite simple groups include the cyclic groups of prime order and the alternating groups A_n in which n is 5 or greater.

The most striking development in the study of finite simple groups was the proof in 1963 by the mathematicians Walter Feit and John Thompson that, apart from the groups of prime order, every finite simple group has even order. Since that time Thompson has determined all minimal simple groups.

Sporadic simple groups

Apart from the families of finite simple groups, there are now known 20 finite simple groups, called sporadic simple groups, that do not belong to the known families. It is not clear whether they belong to families not yet discovered or whether they are truly exceptional, and it is not clear whether further finite simple groups exist.

The designations of the known 20 sporadic groups and their orders in factored form are the Mathieu groups (see 83), the groups, respectively, of Janko, Hall–Janko, Higman–Sims, Suzuki, McLaughlin, Higman–McKay–Janko, and Held (see 84), and the groups identified, respectively, as Conway .1, Conway .2, Conway .3, Fischer F_{22}, Fischer F_{23}, and Fischer F_{24} (see 85).

A 19th group predicted by Lyons was constructed by Sims. The 20th, predicted by Rudvalis, was constructed by Conway and Wales (see 86). (Ma.H.)

III. Fields

Broadly speaking, a field is an algebraic system consisting of elements that are commonly called numbers, in which the four familiar operations of addition, subtraction, multiplication, and division are universally defined (except for division by zero) and have all their usual properties. Much of the general theory of vectors and matrices (see ALGEBRA, LINEAR AND MULTILINEAR) can be developed over an arbitrary field. In particular, this is true of the general theory of simultaneous linear equations and of their solution by a method known as Gaussian elimination (see below). The study of various special fields also explains which geometric constructions, such as angle bisection and angle trisection, can be made with ruler and compass and which cannot. Moreover, finite fields of 2^n elements have been used to construct the best known error-correcting codes.

Although many particular fields—including especially the rational, real, and complex fields, finite fields and algebraic number fields—were intensively studied in the 17th, 18th, and early 19th centuries, the idea of investigating all possible fields seems not to have been conceived until 1910, when the mathematician Steinitz proposed a systematic scheme for classifying them.

THREE FAMILIAR FIELDS

To illustrate the meaning of the first statement of this section, the three most familiar fields are discussed, namely, the rational field, the real field, and the complex field.

The rational field. The most familiar field of all is the rational field, commonly denoted Q. Its members are the positive rational numbers m/n (in which m and n are positive integers), the negative rational numbers $-m/n$, and 0. Positive rational numbers are added, subtracted, multiplied, and divided by the rules familiar in the manipulation of fractions (see 87). The behaviour of 0 is somewhat exceptional in that, when added to a rational it yields the same rational and when multiplied by a rational it yields 0 (see 88). The rules for adding, subtracting, multiplying, and dividing negative rational numbers (with each other and with positive rational numbers and zero) are very like those for positive rational numbers listed above.

Because $m/n = km/kn$ for any positive integer k, the same rational number can be represented as a quotient of integers in many ways. Of these, only one is in lowest terms; that is, has numerator and denominator without common factors (relatively prime) and a positive denominator.

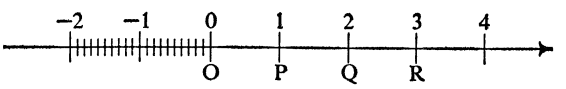
The real field. Much more sophisticated than the rational field is the real field, commonly designated by $\mathfrak{R}$. Positive real numbers are most easily visualized as unlimited decimals (see 89). There are also negative real numbers such as $-1.4142\cdots$ and zero, designated $0 = 0.0000000\cdots$. Among real numbers, rational numbers can be recognized by the characteristic property that their (unlimited) decimal representation is repeating: from some point on, it consists simply of indefinite repetitions of the same block. Thus $1/24 = 0.041666666\cdots$ (block length one), and $5/7$ has a similar representation with block length 6 (see 90). The fact that every rational number is also a real number (with equivalent addition, multiplication, subtraction, and division rules) can be summarized in the statement that Q is a subfield of the rational field—or in the simple formula $Q \subset R$.

Actually, the rules for adding, subtracting, multiplying, and dividing real numbers are quite sophisticated; they can best be understood in terms of the idea of approximation. Thus, when $\sqrt{2} = 1.4142\cdots$ is written, what is meant is that $14{,}142/10{,}000 \leq \sqrt{2} \leq 14{,}143/10{,}000$. But now it is easy to add or multiply any two decimal fractions (see 91) by applying the rules of addition and multiplication of rationals (see 92). Subtraction is equally easy, although long division is considerably harder (see ARITHMETIC).

To pass from the addition and multiplication of decimal fractions (finite decimals) to that of real numbers (unlimited decimals) is best thought of as a process of successive approximation: arbitrarily close decimal fraction approximations to any two real numbers can be added or multiplied together. Thus an arbitrary close approximation to their sum or product can be obtained.

The geometrical representation of real numbers

Real numbers also have a very important and basic geometrical representation as coordinates of points on a straight line. After choosing an origin O and assigning to it the coordinate 0 (zero), any other point P can be chosen and the coordinate $1 = 1.0000000\cdots$ can be assigned to it. By laying off in the same direction equal segments $\overline{PQ} = \overline{OP}$, etc., infinite sequences of points P, Q, R, $\cdots$ (or P_1, P_2, P_3, $\cdots$) are obtained having positive integral coordinates. The coordinates -1, -2, -3, $\cdots$ can then be assigned to a similar sequence of equally spaced points going in the opposite direction. Each segment can also be divided into ten equal parts, assigning (for example) to the new points between P and Q the (decimal) coordinates 1.1, 1.2, 1.3, $\cdots$, 1.9. By repeated decimal subdivision, in this way points corresponding to all decimal fractions can be obtained and thus arbitrarily close (decimal) approximations to any real point produced. This gives a simple and natural geometrical interpretation of the decimal (fraction) approximations to arbitrary real numbers.



The complex field. Much more sophisticated even than the real field is the complex field. Indeed, it is doubtful whether this field would ever have been thought of if man had not been curious about polynomial equations, such as $z^2 = -1$ (or, equivalently, $z^2 + 1 = 0$). It is easy to show that no real number z can satisfy this equation, because the square of any two positive or negative real numbers is necessarily positive (see below *Ordered fields*), while $0^2 = 0$. To remedy this deficiency of the real field $\mathfrak{R}$, mathematicians invented a new number $i = \sqrt{-1}$, defined by the properties of the equation $i^2 = -1$.

In order to get a field containing $\sqrt{-1}$ (that is, to be able to carry out addition, subtraction, multiplication, and division subject to the usual rules), all complex numbers of the form $x + yi = x + iy = x + y\sqrt{-1}$ were considered more generally. It was decided to add and subtract these by the rules of adding and subtracting separately the terms not involving i and the terms involving i (see 93). Multiplication was made according to the rule that the first part of the resultant product was a difference of products not involving i, while the second part included i as a factor and a sum of products, the two parts being added to form a new number in the field (see 94). Finally, it was shown that division (except by zero) was always possible. Specifically, setting $c = a + ib$, $z = x + iy$, and $w = u + iv$, the solution of the equation $cz = w$ can always be evaluated by setting z equal to a complex number, both summands of which have fractional representation with $a^2 + b^2$ in the denominators (see 95) for $c \neq 0$, $a^2 + b^2 > 0$; hence this quotient always exists as a complex number of the prescribed form.

It is quite straightforward to show that the preceding definitions of addition, multiplication, subtraction, and division of complex numbers are the only ones that are compatible with the usual laws of algebra (see below *Postulates for a field*) and the basic equation $i^2 = -1$ given initially above. It can be proved that, conversely, the above operations do obey all the usual algebraic laws; therefore, it is legitimate to speak of the number system just constructed as the complex field $\mathfrak{C}$.

(87)
$$\begin{cases} \dfrac{m}{n} \pm \dfrac{m'}{n'} = \dfrac{mn' \pm nm'}{nn'}, \qquad \dfrac{m}{n} \cdot \dfrac{m'}{n'} = \dfrac{mm'}{nn'} \\ \dfrac{m}{n} \div \dfrac{m'}{n'} = \dfrac{mn'}{nm'} \end{cases}$$

(88)
$$\frac{m}{n} + 0 = \frac{m}{n}, \qquad \frac{m}{n} \cdot 0 = 0 \quad \text{for all } \frac{m}{n}$$

(89)
$$1.414214\cdots, 3.14159\cdots, 2.718281828459\cdots$$

(90)
$$\frac{5}{7} = 0.714285714285714285\cdots$$

(91)
$$\begin{cases} 1.4142 = 14{,}142/10{,}000 \text{ or} \\ 3.14159 = 314{,}159/100{,}000 \end{cases}$$

(92)
$$\begin{cases} \dfrac{14{,}142}{10{,}000} + \dfrac{314{,}159}{100{,}000} = \dfrac{455{,}579}{100{,}000} \\ \dfrac{14{,}142}{10{,}000} \cdot \dfrac{314{,}159}{100{,}000} = \dfrac{4{,}442{,}836{,}578}{1{,}000{,}000{,}000} \end{cases}$$

(93)
$$(x + iy) \pm (x' + iy') = (x \pm x') + i(y \pm y')$$

(94)
$$(x + iy)(x' + iy') = (xx' - yy') + i(xy' + yx')$$

(95)
$$z = \frac{u + iv}{a + ib} = \frac{(u + iv)(a - ib)}{(a + ib)(a - ib)} = \frac{ua + vb}{a^2 + b^2} + i\left(\frac{va - ub}{a^2 + b^2}\right)$$

Just as a real number x can be thought of as the coordinate of a point on an infinite straight line, so each complex number $z = x + iy$ can be thought of as representing a point (x, y) in a plane, the complex plane. The addition and subtraction of two complex numbers $z = x + iy$ and $z' = x' + iy'$ as defined above corresponds to ordinary vector addition and subtraction.

What is more remarkable, the multiplication of two complex numbers also has a simple geometric interpretation, which is most simply explained in terms of polar coordinates. Namely, the vector from the origin $O = 0 + i0$ to $z = x + iy$ has length $r = |z| = \sqrt{x^2 + y^2}$ and makes a certain angle θ with the positive x-axis, called the argument of z. By trigonometry, therefore, $x = r\cos\theta$ and $y = r\sin\theta$, and so z has a representation in terms of r and θ (see 96). If also z' has such a representation, then the multiplication rule (see 94) implies that the product of z and z' is the product of the corresponding r and r' and a complex number whose argument is the sum of corresponding θ and θ' (see 97). In other words, to multiply two complex numbers, multiply their absolute values and add their arguments. These rules can be written in symbols (see 98). As a corollary, it may be stated that for any positive integer n, the absolute value of the nth power of z is the nth power of the absolute value of z, and the argument of the nth power of z is n times the argument of z (see 99). From this, in turn, it follows that any nonzero complex number $c = a + ib$ has exactly n distinct nth roots $c^{1/n}$, all having the same magnitude $|c|^{1/n} = (a^2 + b^2)^{1/n}$ and having equally spaced arguments $\theta_j = (\arg c)/n + 2\pi j/n$, $j = 0, 1, \cdots, n - 1$. These nth roots are thus the vertices of a regular polygon of n sides, with its centre at the origin.

ELEMENTARY THEORY OF FIELDS

Postulates for a field. The three number systems Q, $\mathfrak{R}$, and $\mathfrak{C}$ defined above have in common a great many properties. In particular, they all admit two binary operations of addition and multiplication, which satisfy the five postulates following:

F1. $a + b = b + a$ and $ab = ba$ for all values of a, b.
F2. $a + (b + c) = (a + b) + c$ and $a(bc) = (ab)c$, for all values of a, b, c.
F3. There exist distinguished elements 0, 1 such that $a + 0 = 0 + a = a$ and $a1 = 1a = a$ for all values of a.
F4. Given a, there exist elements $-a$ and (if $a \neq 0$) a^{-1} such that $a + (-a) = 0$ and $aa^{-1} = 1$.
F5. $a(b + c) = ab + ac$ and $(a + b)c = ac + bc$, for all values of a, b, c.

Postulate F1 asserts that addition and multiplication are commutative; F2 that they are associative; F3 that any field has identities for addition and multiplication. Postulate F4 asserts that addition and multiplication (except by 0) are invertible operations; the last postulate, F5, is called the distributive law. Fields may now be formally defined.

A field is an algebraic system the members of which, commonly called numbers, can be combined by two universally defined binary operations (addition and multiplication) that satisfy postulates F1–F5 above.

With a little manipulation, it can be shown that (in the presence of F5), F1–F4 can also be summarized in two statements:

1. The numbers of any field form a commutative group (see above *Groups*) under addition.
2. The nonzero numbers of any field form another commutative group under multiplication.

This gives a second algebraic characterization of fields.

Many familiar rules of algebra follow directly from postulates F1–F5, which are already redundant. For example, $a + b = a + c$ implies successively several equalities (see 100), which imply the following cancellation law:

In any field, $a + b = a + c$ implies $b = c$.

A similar argument proves a second cancellation law:

In any field, $ab = ac$ implies $b = c$ if $a \neq 0$.

Next, it can be proved that the additive identity 0 is also a zero for multiplication, in the sense that $a0 = 0$ for all values of a. Indeed, $a + 0 = a = a1 = a(1 + 0) = a1 + a0 = a + a0$, from which it follows that $0 = a0$ by the cancellation law. Again, from $1 = (-1) = 0$ there follows $a + (-a) = 0 = 0a = 1 + (-1)a = 1a + (-1)a = a + (-1)a$, from which by cancellation is derived $-a = (-1)a$. Finally, two calculations (see 101 and 102) imply $(-1)(-1) = 1 \cdot 1 = 1$ and a general rule of multiplication of negatives (see 103).

Gaussian elimination. The rules derived above from postulates F1–F5 lie fairly near the surface and would probably be assumed as postulates if they did not follow anyway. A collection of much deeper results concerning simultaneous linear equations is now considered. The study of these leads in turn to the theory of vectors and matrices with components, or entries, in any field F.

(96) $$z = r[\cos\theta + i\sin\theta]$$

(97) $$zz' = rr'[\cos(\theta + \theta') + i\sin(\theta + \theta')]$$

(98) $$|zz'| = |z| \cdot |z'|, \qquad \arg(zz') = \arg z + \arg z'$$

(99) $$|z^n| = |z|^n \quad \text{and} \quad \arg z^n = n \arg z$$

(100) $$\begin{cases} b = 0 + b = [(-a) + a] + b = (-a) + [a + b] \\ \quad = (-a) + [a + c] = [(-a) + a] + c = 0 + c = c \end{cases}$$

(101) $$(-1)1 + (-1)(-1) = (-1)[1 + (-1)] = (-1)0 = 0$$

(102) $$(-1)1 + 1 \cdot 1 = [(-1) + 1]1 = 01 = 0$$

(103) $$(-x)(-y) = [(-1)x][(-1)y] = (-1)(-1)(xy) = xy$$

(104) $$ax + by = e, \qquad cx + dy = f$$

(105) $$x = \frac{(ed - bf)}{(ad - bc)}, \qquad y = \frac{(af - ce)}{(ad - bc)}$$

(106) $$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n = k_i, \qquad i = 1, 2, \cdots, n$$

(107) $$\begin{cases} \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc \\ \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & k \end{vmatrix} = \begin{cases} aek + dhc + bfg \\ -bdh - ceg - fha \end{cases} \end{cases}$$

The simplest nontrivial case in point concerns two simultaneous linear equations in two unknowns (see 104), the coefficients and constants in which a, b, c, d, e, f are known numbers; the problem is to determine the unknowns x, y. The answer is given, if the coefficients satisfy $ad \neq bc$, by the formulas that, except for a factor common to both x and y, expresses x in terms of the difference $ed - bf$ and y in the terms $af - ce$ (see 105). This answer is easily checked and most easily found by (i) multiplying the first equation by d, the second equation by b, subtracting the results to eliminate y, and then solving for x; and after this (ii) multiplying the second equation by a, the first by c, subtracting the results to eliminate x, and then solving for y.

The above procedure is called Gaussian elimination, and it can be generalized to the case of n simultaneous linear equations (see 106) in n unknowns $x_j (j = 1, 2, \cdots, n)$. The results can also be presented in the compact notation of determinants (see 107). The reader is referred to the article ALGEBRA, LINEAR AND MULTILINEAR for details. The main point to be made here is that the entire theory can be developed over an arbitrary field F. In particular, this is true of the properties of the rank and linear dependence of vectors and matrices.

SOME GENERAL CONCEPTS

Commutative rings; integral domains. First of all, it should be realized that, if the part of postulate F4 that asserts the existence of a^{-1} is omitted, a much wider class of algebraic systems is formally defined as follows:

A commutative ring is an algebraic system having two binary operations (addition and multiplication) that satisfy postulates F1, F2, F3, F5, and F4′:

F4′. Every element a has an additive inverse $(-a)$ such that $a + (-a) = 0$.

A commutative ring is an integral domain when postulate F4″ is satisfied:

F4″. If $ab = 0$, then either $a = 0$ or $b = 0$.

The most familiar integral domain consists of the integers $0, \pm 1, \pm 2, \cdots$ under ordinary addition and multiplication; it is commonly designated Z. From the domain Z can be constructed a very important class of modular number systems Z_n, one for each positive integer $n > 1$, as follows.

The members of Z_n are the integers $0, 1, \cdots, n-1$, which should be thought of as remainders after division by n. To add two members (numbers) of Z_n, their ordinary sum is taken and then its remainder after division by n (the remainder modulo n) is the answer. Likewise, to multiply two members of Z_n, their ordinary product is taken and again the remainder modulo n is the answer. Thus, in Z_{11}, one has $7 + 9 = 5$, $7 \times 9 = 8$, $5 + 10 = 4$, $5 \times 10 = 6$, and so on.

The proof that every Z_n is a commutative ring is omitted; *i.e.*, that F1–F3, F4′, and F5 hold in Z_n for every integer $n > 1$. When n is a prime number p, multiplication is also invertible and the preceding construction gives a field (*e.g.*, Z_{11} is a field). The modular fields Z_p are the simplest finite fields; others will be constructed later.

When $n = rs$ $(1 < r, s < n)$ is composite, however, it is possible to have $rs = 0$ in Z_n even though $r \neq 0$ and $s \neq 0$. Hence Z_n is not even an integral domain, but only a commutative ring unless n is a prime.

Fields of quotients. There are many integral domains besides Z, however, that are not fields. For example, if J is any integral domain (*e.g.*, if $J = Z$ or if J is a field F), then the domain $J[x]$ of all polynomials over J in the indeterminate x can be formed. Each such polynomial either reduces to 0 or has a canonical form (see 108). (Here and below, the convenient symbol $\in$ is used for "belonging to" or "is a member of.") The nonnegative integer m, the highest power of x, is called the degree of the polynomial $a(x)$. Given $a(x)$ and a second polynomial, the sum can be formed [see 109; the case $m = n$ terminates with $(a_m + b_n)x^m$], and the product (see 110), which is always a nonzero polynomial of degree $m + n$ for $a(x)$ nonzero of degree m and $b(x)$ nonzero of degree n. The domain $J[x]$ satisfies postulates F1–F3, F4′, and F5 and therefore is always an integral domain.

From any integral domain J, for example, from the domain $Z[x]$ of all polynomials with integral coefficients or the domain $R[x]$ of all polynomials with real coefficients, a field of quotients $Q(J)$ can be constructed the elements of which are the formal quotients a/b $(a, b \in J$ with $b \neq 0)$, by adopting definitions of equality, sums, differences, and products (see 111). This construction yields from the integral domain Z of all integers the rational field $Q = Q(Z)$, and from the domain $R[x]$ of real polynomials the field $R(x) = Q(R[x])$ of all real rational functions.

Ideals and quotient-rings. It is possible to greatly generalize the construction of the modular fields Z_p and commutative rings Z_n from the familiar integral domain Z of all integers. The key idea of this generalization is contained in the following basic concepts.

In a commutative ring R, an ideal is a nonvoid subset H such that $h, h' \in H$ imply $h \pm h' \in H$, and $h \in H$ and $r \in R$ imply $rh = hr \in H$. A residue class, or additive coset, of the ideal H is a set $H + x$ consisting of all sums $h + x$ of a fixed element $x \in R$ with variable h of H.

The ideals of the integral domain Z are known to consist of the trivial ideals Z and $(0) = \{0\}$, the subset consisting of zero, and the proper ideals (n) consisting of the set of all multiples kn $(k \in Z)$ of a fixed integer $n > 1$. The residue classes $r + (n)$ of any proper ideal (n) are the sets of integers x having a fixed remainder r under division by n, so that $x = qn + r$ for some integer q.

Z_n can be regarded as the set of all these residue classes, added and multiplied by specific rules (see 112). When constructed in this way, Z_n is called the quotient-ring $Z/(n)$ of Z by the ideal (n).

Morphisms of rings. The preceding basic construction can also be interpreted in terms of another basic concept, that of morphism. If R and S are any two (commutative) rings, a morphism of rings from R to S is a function $\theta : R \to S$, which assigns to each $r \in R$ a value $s = \theta(r)$ in S, such that $\theta(1) = 1$ and morphisms of sums, differences, and products are sums, differences, and products of morphisms of corresponding elements (see 113). Such a morphism is called an epimorphism when it is onto, in the sense that every $s \in S$ is the image $s = \theta(r)$ of some $r \in R$; if it is also one-one, in the sense that $r \neq r'$ in R implies $\theta(r) \neq \theta(r')$ in S, the (epi)morphism is called an isomorphism. Morphisms of rings are related to ideals and quotient rings by the following basic results:

Theorem: If $\theta : R \to S$ is any morphism of rings, then the set $\theta^{-1}(0)$ of all elements $r \in R$ such that $\theta(r) = 0$ in S is an ideal J_θ of R. If θ is an epimorphism, then S is isomorphic to the quotient-ring R/J_θ. Conversely, given any ideal J of a ring R, there exists an epimorphism $\theta : R \to R/J$ under which J is the ideal of all $r \in R$ such that $\theta(r) = 0$ in S.

Subrings and subfields. Here is considered the case in which R is any commutative ring with 1 as its unity element, as postulated in F3. This element is unique because if $1'$ is any other element having the property postulated in F3, then $1' = 11' = 1'1 = 1$. S is a set $S = 0, \pm 1, \pm(1 + 1), \cdots$ obtained from 1 by repeated addition and subtraction. These elements are not necessarily distinct (for example, in Z_{11}). For any positive integer n, $n1$ (or just n) is used to designate the sum $1 + \cdots + 1$ (with n summands), and $(-n)1$ designates the sum $(-1) + \cdots + (-1)$ (again n summands). A final condition is that $0 \cdot 1 = 0$. Using mathematical induction it is then quite easy to prove from the postulates F1–F3, F4′, and F5 for a commutative ring that, for all $m, n \in Z$ (see 114), the function μ: $n \to n1$ is a morphism of rings

$$
\begin{cases} a(x) = a_0 + a_1 x + \cdots + a_m x^m \\ \qquad a_m \neq 0, \quad \text{all } a_j \in J^* \\ b(x) = b_0 + b_1 x + \cdots + b_n x^n \end{cases} \tag{108}
$$

$$
a(x) + b(x) = (a_0 + b_0) + (a_1 + b_1)x + \\ + \cdots + \begin{cases} a_m x^m, & m > n \\ b_n x^n, & n > m \end{cases} \tag{109}
$$

$$
a(x)\, b(x) = a_0 b_0 + (a_0 b_1 + a_1 b_0)x + \cdots + \\ + \left(\sum_{j=0}^{k} a_j b_{k-j} \right) x^k + \cdots + a_m b_n x^{m+n} \tag{110}
$$

$$
\begin{cases} \dfrac{a}{b} = \dfrac{c}{d} \text{ means } ad = bc \\ \dfrac{a}{b} \pm \dfrac{c}{d} = \dfrac{ad \pm bc}{bd} \\ \dfrac{a}{b} \cdot \dfrac{c}{d} = \dfrac{ac}{bd}, \quad \text{and} \quad \dfrac{a}{b} \div \dfrac{c}{d} = \dfrac{ad}{bc} \quad \text{if } c \neq 0 \end{cases} \tag{111}
$$

$$
\begin{cases} [r + (n)] + [r' + (n)] = (r + r') + (n) \\ [r + (n)] \cdot [r' + (n)] = rr' + (n) \end{cases} \tag{112}
$$

$$
\theta(r \pm r') = \theta(r) \pm \theta(r') \quad \text{and} \quad \theta(rr') = \theta(r)\, \theta(r') \tag{113}
$$

$$
m1 \pm n1 = (m \pm n)1 \text{ and } (m1)(n1) = (mn)1 \tag{114}
$$

from Z to R. The set $\{n1\}$ of all natural multiples $n1$ of 1 in R is thus closed under the basic ring operations of addition, subtraction, and multiplication, and it includes the unity 1 of R. A subring of a ring R can be defined to be any subset of R with these properties, and the prime subring can be defined to be the subring $\{n1\}$ just constructed; this prime subring is evidently contained in every other subring: it is the least subring of R.

Characteristic of a ring. Using the results about ideals and quotient-rings in Z stated earlier, one can prove quite easily that the prime subring of any ring is isomorphic to Z or to Z_n for some unique integer $n > 1$. (The case $n = 1$ gives the degenerate ring consisting of 0 alone, in which 0 acts as unity, so that $1 = 0$; this is usually excluded from consideration.) This integer n is called the characteristic of the ring; it is the least positive integer n such that $x + \cdots + x = 0$ whenever the number of summands is a multiple of n. Rings the prime subring of which is isomorphic to Z are said to be of characteristic zero (or, more logically, of characteristic infinity, denoted ∞). Rings of characteristic two are quite special; in them, $y = -y$ (because $y + y = 0$), hence $x - y = x + (-y) = x + y$, so that addition is the same as subtraction.

It is easy to show that the characteristic of any field F is a prime p or infinity. In the former case, the prime subring Z_p is already a field (*i.e.*, closed under division). In the latter case the prime subring Z can be (uniquely) extended to its field of quotients Q, which is thus embedded isomorphically in F as its prime subfield; *i.e.*, as a subfield of F contained in every other subfield. To summarize: every field F contains as its smallest prime subfield a copy of Z_p for some prime p (its characteristic), or of Q if of characteristic ∞.

FIELD EXTENSIONS

Simple algebraic extensions. The preceding concepts make possible a far-reaching generalization of the method used to extend the real field $\Re$ to the complex field $\mathfrak{C} = \Re[i]$. Namely, if F is assumed to be any field, and if it is assumed that $p(x)$ is any monic polynomial—*i.e.*, a polynomial with leading coefficient 1 (see 115)—that is irreducible in the sense that it cannot be factored into polynomials of lower degree with coefficients in F, then the multiples of $p(x)$ form an ideal $(p(x))$ in the integral domain $F[x]$ of all polynomials with coefficients in F. Finally, the quotient-ring $F[x]/(p(x))$ the elements of which are the residue classes of this ideal is a field.

Any field constructed in this way is called a simple algebraic extension of F. Each residue class (additive coset) of the ideal $(p(x))$ contains exactly one polynomial of degree less than n, much as each residue class $(p) + n$ of (p) in Z contains exactly one of the remainders $0, 1, \cdots, p - 1$ commonly used to represent elements of $Z/(p) = Z_p$. Similarly, the polynomial $r(x)$ (see 116) belonging to any given coset $(p(x)) + f(x)$ can be computed from any member of the coset—*e.g.*, from $f(x)$—as its remainder after polynomial long division by $p(x)$.

Hence the polynomials of degree less than n, such as $r(x)$ (see 116) and $s(x)$ (see 117), can be used to represent the elements of $F[x]/(p(x))$ symbolically. They can be added like ordinary polynomials (see 118). To multiply them, their ordinary product is formed (110), then the remainder under division by n is taken. Thus, if $p(x) = x^3 - 2$, the elements of $Q[2^{1/3}] = Q[x]/(x^3 - 2)$ can be taken as all expressions of the form that results (see 119) by replacing x^3 by 2 wherever it occurs; the product (see 120) is obtained in which $2^{1/3}$ could be written for x and $2^{2/3}$ could be written in place of x^2, throughout.

Finite fields. The preceding construction gives all finite fields: any finite field of characteristic p is a simple algebraic extension of its prime subfield Z_p. In more detail, any such field contains p^n elements for some positive integer n and is $Z_p[x]/(p(x))$ for some irreducible monic polynomial of degree n. Moreover, all fields of the same finite order p^n are isomorphic; because they were first constructed by Galois, it is therefore reasonable to speak of the Galois field $GF(p^n)$ of any given prime-power order p^n.

As an example, $GF(16) \cong Z_2[x]/(x^4 + x + 1)$ consists of 0 and the 15 different powers of x (see 121). Under multiplication, the nonzero elements of $GF(16)$ thus form a cyclic group, consisting of the powers of x, or any of the other primitive roots of 1 (see 122) the exponents of which are relatively prime to 15.

The preceding example is typical: in any finite field $GF(p^n)$, the $p^n - 1$ nonzero elements are all powers of some one primitive $(p^n - 1)$th root of unity, which can be taken to be x.

Transcendental extensions. Fields have many extensions besides the simple algebraic extensions discussed above. For example, if F is assumed to be any field, and $F[x]$ is assumed to be the integral domain of all polynomials $p(x)$ in the symbol (indeterminate) x with coefficients in F, then because $F[x]$ is an integral domain, the formal quotients of polynomials (rational forms; see 123) constitute a field $F(x)$ under the definitions (see 111) given earlier for constructing the field of quotients of the domain $F[x]$. An extension E of F that can be constructed from F in this way using a suitable element x of E is called a simple transcendental extension of F. When F is infinite, one can interpret the expressions $R(x)$ (see 123) as rational functions from F to F, defined except for a finite set of values x_i of F in which $q(x_i) = 0$.

By repeating the preceding construction, multiple transcendental extensions of F can be obtained. Thus, for any positive integer n, the field $F(x_1, \cdots, x_n)$ can be constructed of all rational expressions (or forms) in the n indeterminates $x_1, \cdots, x_n$, as the field of (formal) quotients of polynomials in $x_1, \cdots, x_n$ with coefficients in F: the field of quotients of the polynomial ring $F[x_1, \cdots, x_n]$.

Multiple algebraic extensions can be constructed similarly. Before considering the general nature of such multiple algebraic extensions, a special case of multiple quadratic extensions of the rational field Q that has applications to geometry must be considered.

Pythagorean numbers. A real or complex number is called Pythagorean when it can be expressed in terms of rational numbers through a finite number of rational operations and square root extractions; *i.e.*, when it belongs to a multiple quadratic extension of Q. Because the mapping $z \to z^*$ that carries each complex number $x + iy$ into its conjugate $z^* = x - iy$ is an automorphism of the complex field $\mathfrak{C}$, and $x = (z + z^*)/2$ and $y = (z - z^*)/2$, the real and imaginary parts of any

(115) $$p(x) = x^n + b_1 x^{n-1} + \cdots + b_n$$

(116) $$r(x) = r_0 + r_1 x + \cdots + r_{n-1} x^{n-1}$$

(117) $$s(x) = s_0 + s_1 x + \cdots + s_{n-1} x^{n-1}$$

(118) $$r(x) + s(x) = (r_0 + s_0) + (r_1 + s_1)x + \\ + \cdots + (r_{n-1} + s_{n-1})x^{n-1}$$

(119) $$a + bx + cx^2 = a + b\sqrt[3]{2} + c\sqrt[3]{4}, \qquad x = \sqrt[3]{2}$$

(120) $$\begin{cases}(a + bx + cx^2)(a' + b'x + c'x^2) = (aa' + 2bc' + \\ + 2b'c) + (ab' + ba' + 2cc')x + (ac' + bb' + ca')x^2\end{cases}$$

(121) $$\begin{cases}x, x^2, x^3, x^4 = x + 1, & x^5 = x^2 + x, & x^6 = x^3 + x^2 \\ x^7 = x^3 + x + 1, & x^8 = x^2 + 1 \\ x^9 = x^3 + x, & x^{10} = x^2 + x + 1 \\ x^{11} = x^3 + x^2 + x, & x^{12} = x^3 + x^2 + x + 1 \\ x^{13} = x^3 + x^2 + 1 \\ x^{14} = x^3 + 1, & x^{15} = x^0 = 1\end{cases}$$

(122) $$GF(16)\text{: } x, x^2, x^4, x^7, x^8, x^{11}, x^{13}, x^{14}$$

(123) $$R(x) = p(x)/q(x) \quad \text{with } q(x) \neq 0$$

$$\text{(124)}\quad \begin{cases} x^2+y^2=\sqrt{a^2+b^2} & \text{and} \quad 2xy=b \\ x\pm y=[\sqrt{a^2+b^2}\pm b]^{\frac{1}{2}} \end{cases}$$

$$\text{(125)}\quad x^2+y^2+Ax+By=C, \qquad 4C>A^2+B^2$$

$$\text{(126)}\quad C=R[\sqrt{-1}]=\frac{R(x)}{x^2+1}$$

$$\text{(127)}\quad p(z)=c_0+c_1z+c_2z^2+\cdots+c_nz^n, \qquad c_n\neq 0$$

$$\text{(128)}\quad p(z)=c_n(z-z_1)(z-z_2)\cdots(z-z_n), \quad \text{all } z_k\in C$$

$$\text{(129)}\quad (z-z_k)(z-z_k^*)=z^2-(z_k+z_k^*)z+z_k\,z_k^*$$

$$\text{(130)}\quad \begin{cases} x=0010 & x^6=1100 & x^{11}=1110 \\ x^2=0100 & x^7=1011 & x^{12}=1111 \\ x^3=1000 & x^8=0101 & x^{13}=1101 \\ x^4=0011 & x^9=1010 & x^{14}=1001 \\ x^5=0110 & x^{10}=0111 & x^{15}=0001 \end{cases}$$

Pythagorean number are themselves Pythagorean. Conversely, $z^2=c=a+ib$ is equivalent to two simultaneous quadratic equations (see 124). Therefore a complex number $z=x+iy$ is Pythagorean if, and only if, x and y (its real and complex parts) are both Pythagorean.

The subfield of all Pythagorean numbers has an interesting relation to ruler-and-compass constructions in plane geometry. In the complex z-plane, a point is called Pythagorean when its coordinate is a Pythagorean number. A straight line that passes through two Pythagorean points is called a Pythagorean line; the angle between two Pythagorean lines, a Pythagorean angle; and a circle that has a Pythagorean centre and a Pythagorean radius, a Pythagorean circle.

It can be proved very easily that a straight line is Pythagorean if and only if its standard equation ($y=ax+b$ or $x=c$) has (real) Pythagorean coefficients. For, any line having such an equation passes through the Pythagorean points $(0, b)$ and $(1, a+b)$, while if $y_i=ax_i+b$ for two Pythagorean points (x_1, y_1) and (x_2, y_2), then $a=(y_2-y_1)/(x_2-x_1)$ and $b=y_2-ax_2$ must be Pythagorean numbers.

A circle is Pythagorean if and only if its equation, written in the customary way as a sum of squares and a linear component (see 125), has coefficients that are (real) Pythagorean numbers. The circle through any three Pythagorean points z_1, z_2, z_3 is Pythagorean; this can be demonstrated by setting $x=x_i$ and $y=y_i$ $(i=1, 2, 3)$ (see 125) and solving for the unknowns A, B, C.

Likewise, it is easy to show that two Pythagorean lines or circles can only intersect in Pythagorean points. To show this, it suffices to solve for the coordinates (x_i, y_i) of the points of intersection in terms of the (Pythagorean) coefficients of the equations.

From the preceding, it follows that the set of all Pythagorean points, lines, and circles is precisely the set of all points, straight lines, and circles that can be constructed from a given line segment [drawn between (0, 0) and (1, 0)], using a ruler and compass.

Finally, the set of all Pythagorean angles must be determined. Using Galois theory (see below), it can be shown that $1^{1/n}$ is a Pythagorean number if and only if n is the product of a power of two and a Pythagorean prime p_i, such that p_i is one plus a certain power of two. The first few Pythagorean primes are $2=2^0+1$, $3=2^1+1$, $5=2^2+1$, $17=2^4+1$, and $257=2^8+1$. It can be shown that they are all of the form $2^{2^x}+1$, in which $x=2^n$. Correspondingly, 90°, 72°, 60°, and 45° are Pythagorean angles, but 20° is not.

Fundamental Theorem of Algebra. Now the most remarkable property of the complex field $\mathfrak{C}$ is reached, which can also be characterized as the simple algebraic extension (see 126) of the real field $\mathfrak{R}$. This property will be stated now as a theorem, omitting its proof (which is not algebraic, but depends basically on considerations from analysis). This theorem is called the Fundamental Theorem of Algebra.

For any polynomial of degree $n>0$ (see 127), with complex coefficients $c_k\in\mathfrak{C}$, there exists a complex number $z_i\in\mathfrak{C}$ such that $p(z_i)=0$.

A corollary is that the polynomial $p(z)$ (see 127) can always be (uniquely) factored into linear factors (see 128).

As a further corollary, the only polynomials that are irreducible over $\mathfrak{C}$ are linear polynomials, from which it follows that $\mathfrak{C}$ has no proper algebraic extensions.

When the coefficients of $p(z)$ (see 127) are real, its complex roots z_k occur in conjugate pairs, giving real quadratic factors (see 129) with negative discriminant $4x_k^2-4(x_k^2+y_k^2)<0$. It follows that any real polynomial can be factored into real linear and quadratic factors.

APPLICATIONS TO CODING

Binary codes. Finite fields $GF(2^n)$ of characteristic two are widely used in the design of codes for detecting and correcting errors in the transmission of digital information (see INFORMATION THEORY and COMBINATORICS AND COMBINATORIAL GEOMETRY). Such information is commonly transmitted as a sequence of signals of two kinds; for example, dots and dashes, or timed pulses and non-pulses. These can be interpreted as strings of 0's and 1's; the elements of $GF(2^n)$ can also be represented very compactly as strings of binary digits, or bits, 0 and 1. Thus, in $GF(16)$, the x^k, the remainders of which after division by x^4+x+1 were listed above (see 121) as polynomials, can be expressed in the more compact form of quadruples of binary digits (see 130). Thus, if a message consisting of sequences of binary digits is broken down into segments (words) of length four, each word can be interpreted as an element of $GF(16)$.

Because the transmission of messages consisting of such four-bit (or longer) words is not infallible, it is common practice to add check digits to a given message, so the actual (binary) code word sent is longer than the message word containing the actual information to be transmitted. One speaks of an (m, n)-code $(m<n)$ if the binary code words transmitted all have the same length n, greater than the length m of the message words.

Simple codes

The simplest code consists of one check digit, usually taken to be 1 if the number of digits 1 in the actual message is odd and 0 if it is even. Thus, for the four-bit message words listed above, the code words would be: 00101, 01001, 10001, 00110, 01100, 11000, 10111, 01010, 10100, 01111, 11101, 11110, 11011, 10010, and 00011. The set of all code words clearly consists of just those strings of five 0's and 1's that have an even number of 1's. Note that the set of code words contains with any two code words their sum (all digits being added mod 2); thus $00101+10001=10100$ $(x+x^3=x^9)$.

The preceding code is error-detecting: if there is only one error made in transmitting a given code word, this error can be detected because the received word will have an odd number of 1's. More elaborate codes are error-correcting. A simple error-correcting code would consist in repeating each message word three times, so that $x^8=0101$ would be transmitted as 010101010101. If at most one error is made in transmitting each digit of the given message, then the correct message can be decoded by assigning to each near code word the message that gives each digit correctly two or more times. Thus, 011100010100 would still be correctly deciphered as the message 0101, even though it has three errors of transmittal.

Note that in the preceding triple repetition code, the set of all code words is also closed under addition. This is characteristic of the group codes commonly used in practice.

Optimal binary codes. The triple repetition (4, 12)-code just described will correct all single errors, but not all double errors. A much better code for correcting 4-bit messages can be constructed using $GF(8)=Z_2[x]/(x^3$

$+ x + 1$), as follows. Given a 4-bit message $a_0a_1a_2a_3$, form the product (see 131) of the corresponding binary message polynomial $a(x)$ with the fixed encoding polynomial $g(x) = 1 + x + x^3$, and transmit the 7-digit code word $b_0b_1b_2 \cdots b_6$ corresponding to the product in $Z_2[x]$ (see 132). Because the roots of $x^3 + x + 1 = 0$ are primitive 7th roots of unity in $GF(8)$, all code words differ in at least three digits.

Advantages of the (4, 7)-code

Indeed, the (4, 7)-code just constructed is a perfect code in the sense that each 7-digit word not a code word can be corrected to precisely one code word by changing one digit. Because it transmits as much information as the triple repetition (4, 12)-code in less than 60 percent as much time and has roughly the same error-correcting capabilities, it is clearly superior.

Similarly, one can use the field $GF(16)$ to construct a (7, 15) binary code that will correct two errors with certainty and is far superior to the (5, 15) triple repetition code. Namely, given a 7-bit message word $a_0a_1 \cdots a_6$, the product formed (see 133) is of the corresponding binary message polynomial $a(x)$ with the fixed encoding polynomial $g(x)$ (see 134). The product polynomial with typical coefficient b_k (see 135) is transmitted as the 15-digit code word $b_0b_1b_2 \cdots b_{14}$. Because the roots of both of the factors of $g(x)$ above are primitive 15th roots of unity in $GF(16)$, it can be shown that it takes five changes to convert one code word into another, and that each 15-digit word differs in two or fewer letters from at most one nearest code word. That is, the preceding code corrects all double errors. Moreover, it is not far from optimal.

(131) $$a(x)g(x) = (a_0 + a_1x + a_2x^2 + a_3x^3)(1 + x + x^3)$$

(132) $$a(x)g(x) = b(x) = b_0 + b_1x + b_2x^2 + \cdots + b_6x^6$$

(133) $$a(x)g(x) = (a_0 + a_1x + \cdots + a_6x^6)(1 + x + x^2 + x^4 + x^8)$$

(134) $$g(x) = 1 + x + x^2 + x^4 + x^8 = (1 + x^3 + x^4)(1 + x + x^2 + x^3 + x^4)$$

(135) $$a(x)g(x) = b_0 + b_1x + b_2x^2 + \cdots + b_{14}x^{14}$$

GALOIS THEORY

Algebraic number fields. If E is any extension of a field F—that is, if F is any subfield of the field E—then E can be considered as a vector space over F, which is the field of scalars. A number $x \in E$ is then said to be algebraic over F when it generates with F a finite-dimensional subspace of E. This is the case if, and only if, the powers $1 = x^0, x, x^2, \cdots$ of x are linearly dependent over F, so that x satisfies a polynomial equation with coefficients in F.

In particular, a real or complex number z is called algebraic when it is algebraic over the rational field Q, which is, of course, the prime subfield of $\mathfrak{C}$ (and $\mathfrak{R}$). That is, z is algebraic precisely when it satisfies a polynomial equation with rational coefficients, hence an irreducible polynomial equation $p(z) = 0$, so that the extension of Q by z is the simple algebraic extension $Q[z]/(p(z))$. (Actually, every finite-dimensional extension of Q or any Z_p is just such a simple algebraic extension.)

When E itself is a finite-dimensional vector space over F, its dimensionality (over F) is called its degree as an extension of F, and this degree is written symbolically as $[E : F]$. In such cases, every element of E is algebraic over F, and E may properly be called an algebraic extension of F.

More generally, if E is any extension of a field F, the set of all numbers $x \in E$ that are algebraic over F is a subfield. Thus, the set of all algebraic (complex) numbers is a subfield A of $\mathfrak{C}$, which is algebraically complete (or closed) in the sense that any polynomial $p(z)$ of degree $n > 0$ with coefficients in A has a root z_1 in A. This is a proper subfield of $\mathfrak{C}$, because it fails to contain e, π, and other transcendental numbers (a transcendental number is a number that is not the root of a polynomial with integral coefficients).

Still more generally, if F is any field and $p(x)$ any polynomial with coefficients in F, then x will be a root of $p(x) = 0$ in the simple algebraic extension $F[x]/(p(x))$. By repeated extensions, using transfinite induction if necessary, an algebraically complete algebraic extension of F can always be constructed, in which an analogue of the Fundamental Theorem of Algebra holds.

Splitting fields. If F is assumed to be any field, and $p(t)$ any polynomial with coefficients in F, a splitting field of $p(t)$ over F is defined as an extension E of F over which $p(t)$ can be factored into linear factors. To construct a least such splitting field, the following procedure is followed.

First, if it is supposed that $p(t)$ is factored into factors irreducible over F, it is easy to show that such a factorization exists. If the irreducible factors are all linear—that is, of degree one—there is nothing to prove ($E = F$).

Otherwise, $p(t)$ will have an irreducible factor $p_j(t)$ of degree $n > 1$, in which case the simple algebraic extension $E_1 = F[x]/(p_j(x))$ is formed. In E_1, the polynomial p_j, which was written as $p_j(t)$ to avoid confusion between t and the indeterminate x in E_1, will have the root x and hence the linear factor $(t - x)$. Hence $p_j(t)$, and so $p(t)$, can be factored into more factors over E_1.

Repeating this construction at most n times, an extension will be obtained, $E = E_k$ of F over which $p(t)$ can be factored into linear factors. Because E is generated by F and roots $x_1, \cdots, x_k$ of $p(t) = 0$, it is evident that E is a minimal splitting field (it is called the root field of $f(t)$ over F). To prove this rigorously, however, requires a more careful study of isomorphisms and their extensions.

Galois theory of fields. This study, and many other aspects of the deeper algebraic theory of fields (including that of the algebraic solvability of polynomial equations), involves so-called Galois theory, which in turn depends on the concept of a group (see above *Groups*).

For Galois theory, the concept of a normal field extension is indispensable. An extension E of a field F is called a normal over F when E is a splitting field for every polynomial $p(x)$ that is irreducible over F and has one root in E. Equivalently, the condition is that, if $p(x)$ has one root in E and is irreducible over F, then all its roots are in E. Thus the complex field $\mathfrak{C}$ is a normal extension of the real field $\mathfrak{R}$, and the field $Q[(1 + i)/\sqrt{2}]$ obtained by adjoining the fourth root $\omega = (1 + i)/\sqrt{2}$ of -1 to the rational field Q is a normal extension of Q (it is the root field of $x^4 + 1$ over Q, whose other roots are ω^3, ω^5, ω^7).

If it be assumed that E is a normal extension of a field F, then the Galois group of E over F is defined as the group of those automorphisms γ of the field E that leave every element of F invariant; *i.e.*, satisfy $\gamma(x) = x$ for all $x \in F$.

It is known that if x_1 and x_2 are any two roots in E of a polynomial $p(x)$ that is irreducible over F, then there exists an automorphism in the Galois group of E over F that carries x_1 into x_2. Thus, i and $-i$ are the two roots of $x^2 + 1 = 0$ in the normal extension $\mathfrak{C}$ of $\mathfrak{R}$; the automorphism γ: $z \to z^*$ of $\mathfrak{C}$ carries i into $-i$. Likewise, there exist automorphisms of $Q[(1 + i)/\sqrt{2}]$ over Q that carry $\omega = (1 + i)/\sqrt{2}$ into ω^3, ω^5, and ω^7.

A complex number can be shown to be Pythagorean if and only if the smallest normal extension of Q that contains it has a Galois group (over Q) the order of which is a power of two. (The order of any algebraic system is the cardinal number of its elements.) It follows from this and Lagrange's theorem in group theory that no Pythagorean number can satisfy an irreducible equation the degree of which is not a power of two. On the other hand, using analytic geometry, it can be shown that angle trisection is equivalent to solving a cubic equation that is in general irreducible, as in the case of trisecting an angle of 120° to get an angle of 40°. It follows in general that an angle cannot be trisected by a finite construction with ruler and compass; in particular, an angle of 120° or 60° cannot be so trisected.

One can show that the Galois group of any binomial

$x^n - a$ is solvable in a sense defined above in the section *Groups*. More generally, this is true of any normal field extension obtained by successive adjunctions of roots $a^{1/n}$ of constants a already constructed. It follows that if a polynomial equation $p(x) = 0$ over a field F can be solved by radicals, in the sense that its roots can be expressed in terms of the elements of F using only rational operations and root extractions, then the Galois group of the root field of $p(x)$ over F is solvable. The converse is also true: if $p(x)$ is a polynomial the root field of which has a solvable group, then the roots of $p(x) = 0$ can be expressed in terms of its coefficients using only rational operations and root extractions.

On the other hand, the Galois group of an irreducible polynomial equation of degree n over Q is, in general, the symmetric group of all permutations of its n roots $x_1, \cdots, x_n$. Moreover, this symmetric group is solvable if and only if $n \leq 5$ (the proof of this is quite technical). It follows that any quadratic, cubic, or quartic equation is solvable by radicals, but that quintic and higher degree polynomial equations in general are not.

ORDERED FIELDS

Positivity. To define the real field, to establish the deeper properties of the real and complex fields, and in particular to construct the real field from the rational field, more than the four rational operations of addition, multiplication, subtraction, and division must be considered. For example, these operations fail to distinguish $\sqrt{2}$ from $-\sqrt{2}$: the Galois group of $\mathfrak{C}$ over Q interchanges $\pm\sqrt{2}$.

The simplest and most natural concept with which to construct $\mathfrak{R}$ from Q and to establish its deeper properties (along with those of $\mathfrak{C}$) is the order concept or, equivalently, that of positivity. These concepts are now defined. A field F is said to be ordered when it contains a set P of positive elements having three properties stating that the sum and product of any two elements in P belong to P and that for every x, exactly one of the following is true: $x = 0$, x is in P or $-x$ is in P (see 136).

In an ordered field, $x > y$ is defined to mean that $(x - y) \in P$. Then $x \in P$ is equivalent to $x > 0$, and the properties of P can be used to prove the transitive property of the order relation, its preservation when elements satisfying it are replaced by their sums or products with appropriate field elements, and the threefold alternative of order between any two elements (see 137). Conversely, in any field F having an order relation satisfying these properties (see 137), the set P of all elements x such that $x > 0$ satisfies the definition of an ordered field (see 136), while $u > v$ if, and only if, $(u - v) \in P$; indeed, these results hold more generally in integral domains.

It is a familiar fact that Q and $\mathfrak{R}$ are ordered fields with the usual order. Indeed, in $\mathfrak{R}$, P can be defined as the set of all nonzero a such that $\sqrt{a}$ exists in $\mathfrak{R}$, giving a purely algebraic characterization of the order relation. As a result, there are no nontrivial field automorphisms of $\mathfrak{R}$.

On the other hand, there is no way to make the complex field $\mathfrak{C}$ or any field such as Z_p of finite characteristic into an ordered field. This is because, if x is any nonzero element of an ordered field, then by the second property of the definition of an ordered field (see 136) $x^2 = (-x)^2$ must be positive; hence any sum of squares must be positive. It follows that if a field can be made into an ordered field, then it must satisfy the implication that each of n elements are 0 if the sum of their squares is 0 (see 138). Fields having this property are called formally real; it has been shown that a field cannot be ordered unless it is formally real. Evidently $\mathfrak{C}$ is not formally real, because $1^2 + i^2 = 0$; and no field of characteristic p can be formally real, because in it the sum $1^2 + \cdots + 1^2$ (p summands) is equal to zero.

Using the concept of a formally real field, the mathematician Artin solved in 1926 a famous problem of Hilbert, showing that a real polynomial $p(x_1, \cdots, x_n)$ assumes positive values for all real $x_1, \cdots, x_n$ if, and only if, it is a sum of squares of polynomials.

Archimedean fields. As will be shown, there is a vast variety of ordered fields. The familiar ordered fields Q and $\mathfrak{R}$ have a property that sets them apart from many others, the Archimedean property, as follows.

The definition is that an ordered field (or integral domain or even additive group) is called Archimedean when, given $\delta > 0$ (no matter how small) and $M > 0$ (no matter how large), there exists a positive integer n such that its product with δ, defined in terms of the addition of n summands, is greater than M (see 139).

In fields, but not necessarily in integral domains, the preceding condition is equivalent to either of the following:

(i) The set $\{n\}$ of all positive integers has no upper bound.
(ii) The set $\{1/n\}$ of all reciprocals of positive integers has greatest lower bound 0.

[Explanation: Condition (i) implies the Archimedean property (see 139), because it allows the choice of $n > M/\delta$; by an upper bound of a set S is meant as usual a number λ such that $\lambda \geq x$ for all $x \in S$.]

Condition (ii) can also be written in the form of a vanishing limit of reciprocals of n as n approaches ∞ (see 140), which means simply that, given $\varepsilon > 0$, n can be found such that $n < 1/\varepsilon$; in other words, that there is no $\varepsilon > 0$ such that $1/n \not< 0$ for all positive integers n. Because 0 is obviously a lower bound to $\{1/n\}$, this is equivalent to saying that it is the greatest lower bound.

An interesting non-Archimedean field is provided by the set consisting of 0 and all extended formal power series, whether convergent or not, expressed with a factor x^r common to all terms and a first coefficient a_0 (see 141), in which r can be any integer, the coefficients c_k belong to a specified ordered field F, and $a_0 \neq 0$. The extended formal power series f (see 141) is called positive if, and only if, $a_0 > 0$, while the sum and product of two such series are defined by the usual formal rules (see 142). It is then straightforward to show that the above construction gives a non-Archimedean ordered field. In this ordered field, the set of all ordinary power series—*i.e.*, extended formal power series (141) with $r \geq 0$—is a subring, hence a non-Archimedean ordered domain.

It is evident that any subfield S of an ordered field F is itself an ordered field (positivity having the same meaning), and that S is an Archimedean ordered field if F is.

Limits and convergence. In any ordered set, whether a field or not, one can speak of limits and convergence. Namely (see 140), if $x_1, x_2, x_3, \cdots$ is any infinite sequence

(136) $\begin{cases} x \in P \text{ and } y \in P \text{ imply } x + y \in P \\ x \in P \text{ and } y \in P \text{ imply } xy \in P \\ x = 0,\ x \in P, \text{ or } -x \in P \end{cases}$

(137) $\begin{cases} \text{If } x > y \text{ and } y > z, \text{ then } x > z. \\ \text{If } a \in F \text{ and } x > y, \text{ then } a + x > a + y. \\ \text{If } a > 0 \text{ and } x > y, \text{ then } ax > ay. \\ \text{Given } x \text{ and } y, \text{ precisely one of the following alternatives holds:} \\ x = y, \quad x > y, \quad \text{or} \quad y > x \end{cases}$

(138) $x_1^2 + x_2^2 + \cdots + x_n^2 = 0 \text{ implies } x_1 = x_2 = \cdots = x_n = 0$

(139) $\underbrace{n \text{ summands}}$
$n\delta = \delta + \cdots + \delta > M$

(140) $\lim_{n \to \infty} \frac{1}{n} = 0$

(141) $f(x) = x^r(a_0 + a_1 x + \cdots + a_n x^n + \cdots)$

(142) $[x^r(a_0 + a_1 x + a_2 x^2 + \cdots)] \cdot$
$\cdot [x^s(b_0 + b_1 x + b_2 x^2 + \cdots)]$
$= x^{r+s}[a_0 b_0 + (a_1 b_0 + a_0 b_1)x +$
$+ (a_2 b_0 + a_1 b_1 + a_0 b_2)x^2 + \cdots]$

of elements of an ordered set C, it may be said that the sequence $\{x_n\}$ converges to the limit a (see 143) if, and only if, for every $\varepsilon > 0$ there exists an n so large that $a - \varepsilon < x_m < a + \varepsilon$, whenever m is greater than n.

Monotone sequences

The most familiar examples of convergent sequences are provided by the successive lower and upper n-place decimal approximants to a given number. Thus, ⅓ may be expressed as a limit of decimal fractions composed successively of greater numbers of the digit 3 (see 144) and so on.

The preceding examples are monotone; *i.e.*, increasing or decreasing. Limits of monotone sequences have an especially simple characterization. Thus, if a "less than or equals" relation holds between any pair of elements with adjacent subscripts in a sequence (see 145)—that is, if the sequence is any increasing (or non-decreasing) sequence of numbers from an ordered field F—then $x_n \to a$ if, and only if, a is the least upper bound (l.u.b.) of the set of $\{x_n\}$. By this it is meant that: $a \geq x_n$ for all n, and $b \geq x_n$ for all n implies $b \geq a$. The fact that an increasing sequence $\{x_n\}$ converges to a is often written $x_n \uparrow a$. Likewise, one writes $x_n \downarrow a$ to signify that the sequence $\{x_n\}$ is decreasing (or non-increasing), with greatest lower bound (g.l.b.) a.

Completeness. Now the deepest property of the real field $\mathfrak{R}$ is reached: the property of completeness, which distinguishes it from all other ordered fields. Actually, there are two kinds of completeness, which will be discussed in turn.

Cauchy completeness

First, there is sequential, or Cauchy, completeness. If $\{x_n\}$ is any convergent sequence in an ordered field with limit a, then the doubly infinite set of differences $\{x_m - x_n\}$ has the property of being a Cauchy sequence in the following sense.

A sequence $\{x_n\}$ of numbers of an ordered field F is defined to be a Cauchy sequence when, given $\varepsilon > 0$, there exists N so large that $m, n > N$ imply $-\varepsilon < x_m - x_n < \varepsilon$.

In the real field $\mathfrak{R}$, the converse is also true: any Cauchy sequence has a limit to which it converges. This property defines sequential completeness.

An ordered field (or group) F is complete in the sense of Cauchy when every Cauchy sequence converges to some limit.

The real field is complete in the sense of Cauchy. For, choosing $\varepsilon = 10^{-r}$ ($r = 1, 2, 3, \cdots$), from any Cauchy sequence $\{x_n\}$ can be obtained a subsequence $\{x_{n(r)}\}$ of decimal fractions, such that $x_{n(r)}$ differs from all later terms at most in the rth decimal place.

The real field $\mathfrak{R}$ can be constructed from the rational field Q as follows, using the property of Cauchy completeness. The elements of the field $\mathfrak{R}$ can be constructed as the limits of Cauchy sequences of rational numbers with the identification of two such sequences (see 146) meaning that the limit of the difference of corresponding terms is 0.

Second, there is a more powerful, but less constructive, characterization of completeness, which may be expressed as follows.

Dedekind completeness

An ordered field (or other set) F is defined to be complete in the sense of Dedekind when every bounded set of its elements has a least upper and a greatest lower bound.

It is easy to show that any ordered field that is complete in the sense of Dedekind is Archimedean, and complete in the sense of Cauchy. Conversely, an ordered field F that is Archimedean and complete in the sense of Cauchy is necessarily complete in the sense of Dedekind. What is even more important, however, the following results are true.

Theorem: An ordered field is Archimedean if and only if it is isomorphic with a subgroup of the real field. An ordered field is isomorphic to the real field if and only if it is complete in the sense of Dedekind.

The deepest questions of all are now reached: those of the existence and uniqueness (up to isomorphism) of complete ordered fields. As will be seen, the notion of an absolute value or valuation plays a central role in discussions of these questions.

(143) $$\lim_{n\to\infty} x_n = a \quad \text{or simply} \quad x_n \to a$$

(144) $$\begin{cases} \{.3, .33, .333, .3333, .33333, \cdots\} \to \frac{1}{3} \\ \{.7, .67, .667, .6667, .66667, \cdots\} \to \frac{2}{3} \\ \{3, 1, 3.14, 3.141, 3.1415, 3.14159, \cdots\} \to \pi \end{cases}$$

(145) $$\{x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_n = \cdots\}$$

(146) $$\{x_n\} = \{y_n\} \text{ means that } \lim_{n\to\infty} (x_n - y_n) = 0$$

(147) $$\begin{cases} |0| = 0, \quad |x| > 0 \quad \text{if } x \neq 0 \\ |a+b| \leq |a| + |b| \\ |ab| = |a| \cdot |b| \\ |(|a|)| = |a| \end{cases}$$

(148) $$|z| = \sqrt{x^2 + y^2}, \qquad z = x + iy$$

(149) $$q = (r/s)p^{\nu}$$

(150) $$\begin{cases} \phi_p(0) = 0, \quad \phi_p(q) > 0 \quad \text{if } q \neq 0 \\ \phi_p(a+b) \leq \phi_p(a) + \phi_p(b) \\ \phi_p(ab) = \phi_p(a)\,\phi_p(b) \\ \phi_p(a+b) \leq \text{Max}\{\phi_p(a), \phi_p(b)\} \end{cases}$$

VALUATIONS

Underlying ideas. The reader may be familiar with the distinction between the sign $\pm$ and the magnitude of a real number. More generally, in an ordered field, the absolute value $|x|$ of any number x can be defined as having the value x if x is positive or zero and the value $-x$ otherwise. From this definition such standard relations as those that relate to the absolute value of 0 and of sums, products, and absolute values (see 147) can be easily derived.

Even though the complex field cannot be made into an ordered field, the definition of the square root of the sum of squares of components of a complex number (see 148) defines an absolute value function sharing the formal properties (147). The notion of absolute value with this definition (148) is indispensable in complex analysis, the theory of functions of a complex variable.

There are many analogues and generalizations of the preceding absolute-value function. An interesting analogue is defined, on the rational field Q, for any prime number p, by the following observation. Each rational number $q = m/n$ (m, n in lowest terms) is uniquely expressible in the form of a ratio of r and s times a power of p (see 149), in which r and s are relatively prime to each other and to p, and $s > 0$.

Given the decomposition (see 149), the number $\phi_p(q) = p^{-\nu}$ is called the p-adic valuation of q.

This valuation is counterintuitive (contrary to one's intutitive feeling): for example, $\phi_2(1024)$ is very small, whereas $\phi_2(1/1024)$ is a million times as large. Nevertheless, it shares (see 150) the first three formal properties of the more familiar absolute value function (see 148). Only the fourth property fails. By way of compensation, the p-adic valuation of q has a sharper property (see 150) than that of the second property of absolute value, which one can strengthen to an inequality involving the maximum of two p-adic valuations rather than the sum.

There are many other important generalizations of the absolute value concepts to systems other than fields: the absolute value $(x^2 + y^2 + z^2 + t^2)^{1/2}$ of a real quaternion $q = t + xi + yi + zk$; the norm $N(a + b\sqrt{2}) = a^2 - 2b^2$ of a number $c = a + b\sqrt{2}$ in $Q(\sqrt{2})$; the norm of an element of a Banach space; and so on.

Completing fields. The way will now be described by which any valuation having the first three properties of

the p-adic valuation (see 150) or their equivalents, the first three properties of absolute value (see 147), can be used to complete a given field. The field extension construction to be described gives the real field $\mathfrak{R}$ from the rational field Q; the complex field $\mathfrak{C}$ from the field $Q(\sqrt{-1})$ of all rational complex numbers $r + s\sqrt{-1}$ $(r, s \in Q)$; and a new family of p-adic number fields from the p-adic valuations of Q. It involves the notion of a Cauchy sequence, which is generalized in the following definition:

Let F be any field with a real valuation ϕ satisfying the first three properties of the p-adic valuation (see 150). A sequence $\{x_n\} = \{x_1, x_2, x_3, \cdots\}$ of elements x_n of F is called a Cauchy sequence (for the valuation ϕ) when, given $\varepsilon > 0$, there is an integer N such that the value of ϕ at the difference of two elements with subscripts sufficiently large is less than ε (see 151).

The sum and product of two Cauchy sequences can be defined by obvious formulas (see 152). Under these definitions the Cauchy sequences form a commutative ring F_ϕ (a subring of F^ω), which is, however, not even an integral domain, let alone a field. To make it into a field, the following equivalence relation must be introduced. (In the language of the theory of commutative rings [see below *Rings*], a maximal ideal M of F_ϕ must be constructed; the quotient-ring F_ϕ/M will then be a field.)

Two Cauchy sequences as defined are concurrent when the limit of the valuation ϕ for differences of corresponding terms in the two sequences is 0 (see 153).

In the construction of the real field $\mathfrak{R}$ from the (ordered) rational field Q, typical Cauchy sequences consist of decimal (or binary) approximations to a given unlimited decimal number; two such Cauchy sequences are concurrent when they ultimately agree to any preassigned number of decimal (or binary) digits, leaving aside decimal numbers such as $0.999999\cdots = 1.000000\cdots$ themselves. In this case, the field constructed is not only complete in the sense of Cauchy, it is also an Archimedean ordered field that is complete in the sense of Dedekind. The positive numbers of F_ϕ are the Cauchy sequences $\{x_n\}$ the terms of which are ultimately all positive, and that are not null sequences (do not converge to zero, are not concurrent to $\{0, 0, 0, \cdots\}$).

To recapitulate, the real field $\mathfrak{R}$ is not only Archimedean and complete in the sense of Cauchy; it is also complete in the sense of Dedekind. Conversely, it can be shown that any field that is complete in the sense of Dedekind is Archimedean and complete in the sense of Cauchy. Moreover, in it, one can reconstruct all numbers as limits of decimal fractions in the usual way and derive the usual laws for adding and multiplying the resulting infinite decimals.

This shows that any ordered field that is complete in the sense of Dedekind is isomorphic to the real field: up to isomorphism, there is one and only one such field. On the other, the field $\mathfrak{R}[[x]]$ of all extended formal power series with real coefficients is a non-Archimedean ordered field (see 141), which is complete in the sense of Cauchy.

Finally, it may be remarked that for $\phi = \phi_p$, the previous field extension by Cauchy sequences (modulo null sequences) gives the so-called p-adic numbers of Hensel. These constitute for each prime p a p-adic field, the nonzero elements of which can be thought of as extended formal series in the symbol p with integral coefficients a_k, first coefficient with subscript 0, and with a factor common to each term of the form p^α (see 154), in which α is an arbitrary integer, every a_k is one of the integers 0, 1, $\cdots$, $p - 1$, and $a_0 \neq 0$. Though the p-adic fields have characteristic zero, they are not to be thought of as ordered fields, still less as Archimedean ordered fields.

(G.Bi.)

(151) $\phi(x_m - x_n) < \varepsilon \quad \text{if } m, n > N$

(152) $\begin{cases} \{x_n\} + \{y_n\} = \{x_n + y_n\} \\ \{x_n\} \cdot \{y_n\} = \{x_n y_n\} \end{cases}$

(153) $\lim_{n \to \infty} \phi(x_n - y_n) = 0$

(154) $x = p^\alpha(a_0 + a_1 p + a_2 p^2 + a_3 p^3 + \cdots)$

(155) $u = \cos\frac{2\pi}{n} + i \sin\frac{2\pi}{n}, \qquad u^n = 1$

(156) $a_0 + a_1 u + \cdots + a_{n-1} u^{n-1}$

(157) $x^n = (z - y)(z - uy) \cdots (z - u^{n-1} y)$

(158) $F_s(x_1, \cdots, x_n) = 0, \qquad s = 1, \cdots, r$

IV. Rings

The notion of a ring has been mentioned at the beginning of this article and used in the preceding section *Fields*. It is a set of elements that can be added, subtracted, and multiplied and in which these three operations obey (more or less) the usual rules. A more formal definition will be given later. In particular, if R is a set of real or complex numbers such that the sum, the difference, and the product of any two elements of R is in R, then R is a ring for the usual operations. It is not required that the division by a nonzero element of a ring be always possible; thus the divisibility problem for elements of a ring is important, as it is for the ordinary integers.

The rings introduced by the German arithmeticians of the 19th century for studying problems involving divisibility and factorization were, in fact, mostly sets of complex numbers (more precisely of algebraic integers) as above. The main impetus came from the intensive study of Fermat's equation $x^n + y^n = z^n$, stated without proof by the French mathematician Pierre de Fermat (*q.v.*) to have no solution (x, y, z) in integers greater than 1 for exponents n greater than 2. Fermat's equation

It has been proved that Fermat's statement is true for many exponents n, but the general case is still open. The most fruitful approach to the problem introduces a complex nth root of unity (see 155) and the numbers which are polynomials in such an nth root, coefficients a_i (see 156) being ordinary integers; these numbers form a ring R. Fermat's equation can be written (see 157) in a way that is related with the problem of decomposing x^n in the ring R. If the analogue of the unique decomposition of an integer into prime factors is true in R, then Fermat's statement can be proved. This method (or a variant) works for some values of n (*e.g.*, $n = 3, 5, 14$), but the German mathematicians Peter Gustav Lejeune Dirichlet and Ernst Eduard Kummer noticed that the analogue of the unique decomposition is not true in general. As a remedy, Kummer and another German mathematician, Richard Dedekind, introduced the notion of an ideal, which proved to be of fundamental importance for ring theory.

At the end of the 19th century another impetus came from algebraic geometry. Roughly speaking, this branch of mathematics studies the n-tuples $(x_1, \cdots, x_n)$ of complex numbers that satisfy a given finite system of polynomial equations (see 158); these n-tuples form what is called an algebraic set, and the polynomial functions on such an algebraic set H form a ring, its coordinate ring, for addition and multiplication in a modified sense, called value-wise addition and multiplication. An important paper of Dedekind and Weber, published in 1882, showed that when this algebraic set H is a curve—*i.e.*, has dimension one—the coordinate ring of H has many properties in common with the rings of algebraic integers studied by the arithmeticians, a striking parallelism between geometry and arithmetic, which has been intensively generalized and fruitfully studied since. In 1893 the German mathematician David Hilbert introduced powerful ideal theoretical methods in the study of general polynomial rings and coordinate rings, proving the important finite basis theorem and the "zeroes theorem." In this theory the substitute for prime decomposition is the primary decomposition of ideals, discovered and

studied in 1905 by the German mathematician and chessmaster E. Lasker, and around 1915 by F.S. Macaulay in England. The study of general commutative rings was inaugurated by the German mathematician Emmy Noether around 1925, with a special emphasis on the rings that are now called Noetherian. In the second half of the 20th century, with his theory of schemes, the mathematician A. Grothendieck has introduced a far-reaching generalization of both commutative ring theory and algebraic geometry.

Noncommutative rings also exist, the first examples of them coming from elementary geometry. The study of rotations in ordinary 3-space led the Irish mathematician Sir William Rowan Hamilton (*q.v.*) to discover in 1843 the ring of quaternions, a noncommutative division ring in which every nonzero element has an inverse. At the same time, motivated by the study of linear spaces, the German mathematician Hermann Günther Grassmann introduced what is now called the exterior algebras. The English mathematician Arthur Cayley introduced matrices around 1850 for describing linear transformations, and the fact that the square matrices form a ring was pointed out by the mathematicians B. and C.S. Peirce in 1870. In the second part of the 19th century these rings were studied under the name hypercomplex systems. Around 1930, the tools of the commutative and of the noncommutative theories were unified, giving to the latter one its proper degree of power and generality. Under the impulse of functional analysis, rings of linear transformations in infinite-dimensional spaces and related rings were studied by the Hungarian-born U.S. mathematician John von Neumann (*q.v.*) and his followers.

DEFINITION AND EXAMPLES

Definition. A ring R is defined to be a set endowed with two binary operations satisfying the axioms that are listed below. The first operation, called addition, associates to every pair x, y of elements of R a well-defined element of R, denoted by $x + y$; this addition satisfies the axioms of associativity, commutativity, the existence of a zero element, and the axiom assuring that for every element there is an opposite (or negative) element:

A1. $x + (y + z) = (x + y) + z$ for all x, y, z in R (associativity).
A2. $x + y = y + x$ for all x, y in R (commutativity).
A3. There exists an element 0 of R such that $0 + x = x + 0 = x$ for every x in R (zero element).
A4. For every x in R, there exists an element (denoted by $-x$) such that $x + (-x) = (-x) + x = 0$ (opposite).

In other words, R is a commutative group for the addition.

The second operation, called multiplication, associates to every pair x, y of elements of R a well-defined element of R, denoted by xy or $x \cdot y$; this multiplication satisfies the associativity axiom and two distributive axioms:

(M) $x(yz) = (xy)z$ for all x, y, z in R.
(D) $x(y + z) = xy + xz$, $(y + z)x = yx + zx$ for all x, y, z in R.

It is easily proved that $x \cdot 0 = 0 \cdot x = 0$ for every x in R. If R contains an element e such that $ex = xe = x$ for every x, R is called unitary; the element e (necessarily unique) is called the unit element of R and is usually denoted by 1. Most rings encountered in mathematics are unitary. A ring R such that $xy = yx$ for all x, y in R is called commutative.

Examples of rings. The set Z of ordinary positive and negative integers $(0, 1, 2, \cdots, -1, -2, \cdots)$ is a unitary commutative ring for the usual addition and multiplication; its zero element is 0, its unit element is 1, so that the general notation for rings coincides with the usual notation for integers.

Another example is the set of two elements, "even" and "odd." This set is made into a ring F_2 by using the following natural definitions of addition and multiplication:

A. even + even = even, even + odd = odd + even = odd, odd + odd = even.
B. even · even = even · odd = odd · even = even, odd · odd = odd.

(159) $0+0=0, \quad 0+1=1+0=1, \quad 1+1=0$

(160) $0 \cdot 0=0 \cdot 1=1 \cdot 0=0, \quad 1 \cdot 1=1$

(161)

addition				multiplication		
0	1	a	b	1	a	b
1	0	b	a	a	b	1
a	b	0	1	b	1	a
b	a	1	0			

(162)

0	1	e	f	1	e	f
1	0	f	e	e	e	0
e	f	0	1	f	0	f
f	e	1	0			

(163)

0	1	2	3	1	2	3
1	2	3	0	2	0	2
2	3	0	1	3	2	1
3	0	1	2			

Here "even" is the zero element and "odd" the unit element. In the general notation for rings, if the digits 0, 1, replace the elements "even" and "odd," the elements of F_2 satisfy other formulas (see 159, 160) instead of (A) and (B).

They have a familiar form, except the strange formula $1 + 1 = 0$. Concerning this one, it should be remembered that the symbols 0, 1 used here do not stand for the usual numbers. Mathematicians feel that the risks involved with such unusual formulas are much smaller than the benefit of computing with the symbols 0, 1 according to deeply ingrained habits.

Now three examples of rings with four elements are considered. The elements of the first one are called 0, 1, a, b. Addition and multiplication are given by tables: the value of the sum $x + y$ is obtained by looking, in the addition table, at the element the row of which begins with an x and the column of which begins with a y; similarly for multiplication, except that products of the form $0 \cdot x$ or $y \cdot 0$ are not given, because their values are 0 (see 161). For example, $1 + a = b$, $b + a = 1$, $b \cdot b = a$.

The elements of the second ring are called 0, 1, e, f, with the tables also used to define addition and multiplication (see 162). For the elements of the third one, the symbols 0, 1, 2, 3 are used, and tables are also used (see 163).

It is easy to verify that the axioms for rings hold. These three rings are commutative and unitary. The additions in two of the examples (161 and 162) are the same, provided a corresponds to e and b to f. Their multiplications, however, are distinct because in the first (161) the product of any two nonzero elements is distinct from 0, whereas in the second (162), $ef = 0$; such elements are called zero divisors. Furthermore, in the first, every nonzero element has an inverse; for example, a and b are inverses of one another because $ab = ba = 1$. More generally, a unitary ring R in which every element $x \neq 0$ has an inverse—*i.e.*, an element x' such that $xx' = x'x = 1$—is called a division ring; if R is commutative, it is called a field.

The additions in two of the examples (162 and 163) are not the same, because $x + x = 0$ for every x in one example (162), whereas $3 + 3 = 2 \neq 0$ in the other example (163). Also a new feature occurs in one of the multiplications (163): the product $2 \cdot 2$, which can be written as 2^2 with the usual convention for exponents, is equal to zero. An element x of a ring R such that some power x^n of x is equal to 0 is called a nilpotent element of R. So 2 is nilpotent (see 163). The rings in the first two

examples (161 and 162), however, have no nilpotent elements except 0.

The fourth example involves polynomials and power series with coefficients a_k in R, a commutative unitary ring. A letter X and formal finite sums of coefficients times integer powers of X are considered. These formal sums, called polynomials over R, are to be added coefficientwise (see 164) and multiplied in such a way that exponents of products of powers are added (see 165).

A polynomial over R may be viewed as the infinite sequence $(a_0, a_1, \cdots, a_n, 0, 0, \cdots)$ of its coefficients, only a finite number of terms a_i being unequal to 0. These sequences are to be added termwise and multiplied according to the given rule. Then the letter X corresponds to the sequence $(0, 1, 0, 0, \cdots)$.

If the restriction that only a finite number of terms of the sequences be unequal to 0 is removed, the formal power series over R is obtained (see 166).

Either the polynomials or power series over R form a commutative unitary ring for the above defined addition and multiplication; it is denoted by $R[X]$ for polynomials, $R[[X]]$ for power series. By iterating the process, the polynomial and power series rings in several variables over R are obtained. The former kind of ring is important in algebraic geometry, and the latter in analysis.

Multiplicative groups. If one forgets about multiplication in a ring R, its additive group, a commutative one, is obtained. This group is not much more interesting than a general commutative group; for example, the problem of which commutative groups are additive groups of rings is rather trivial.

Much more interesting is the multiplicative group of a unitary ring R. An element x of R for which there is an element x' of R such that $xx' = x'x = 1$ is called invertible; such an x' is necessarily unique, as a simple calculation shows (see 167). It is called the inverse of x and denoted by x^{-1}. If x and y are invertible, so is xy because it admits $y^{-1}x^{-1}$ as an inverse. Thus the invertible elements of R form a group for the multiplication, called the multiplicative group of R, and sometimes denoted by R^*.

The multiplicative group of the ring Z of integers is $\{+1, -1\}$. In the example of the set with elements "even" and "odd," and the second and third of the four-element rings above (see 160, 162, and 163), it is reduced to 1. For the first four-element ring above (see 161) it is the 3-elements group $\{1, a, b\} = \{1, a, a^2\}$, with $a^3 = 1$. If R has no zero divisors, the invertible polynomials over R are the invertible constants—*i.e.*, 1-term polynomials a_0—so that $R[X]^* = R^*$. The invertible power series $a_0 + a_1X + \cdots$ is the power series with an invertible constant term a_0. In the ring of numbers $a + b\sqrt{3}$ $(a, b \in Z)$, the invertible elements are the powers $(2 + \sqrt{3})^n$ $(n \in Z)$ of $2 + \sqrt{3}$, and their negatives (notice that $(2 + \sqrt{3})(2 - \sqrt{3}) = 4 - 3 = 1$); writing $(2 + \sqrt{3})^n = a_n + b_n\sqrt{3}$ $(a_n, b_n \in Z, n \geq 0)$ the pairs (a_n, b_n) provide all the solutions in positive integers of the equation $x^2 - 3y^2 = 1$. An example is easily constructed (see 168).

Homomorphisms. Modern mathematics has shown the importance not only of sets endowed with some structures, such as groups, rings, and metric spaces, but also of interesting classes of mappings from one of these sets into another one of the same type. To be interesting, such a mapping must be, in some sense, compatible with the given structures. In algebra, it has to preserve the operations. For example, interesting mappings of a ring R into a ring R' are the homomorphisms; that is, the mappings $f\colon R \to R'$ such that sums and products are preserved (see 169). It follows that $f(0) = 0$, $f(-x) = -f(x)$. If R and R' are unitary, it is usually also required that a homomorphism f carry the unit element of R to the unit element of R' (*i.e.*, $f(1) = 1$). For example, for every integer $x \in Z$, $f(x)$ may be called the remainder of division of x by 4. It is one of the integers 0, 1, 2, 3; *e.g.*, $f(7) = 3$, $f(18) = 2$, $f(-3) = 1$, $f(-12) = 0$. Now if R is the set $\{0, 1, 2, 3\}$ endowed with the ring structure defined in the third example of a four-element ring (see 163), it is easily checked that f is a homomorphism of Z into R.

Similarly, for $x \in Z$, $g(x) =$ even, if x is even, and

(164) $$\begin{cases}(a_0 + a_1X + \cdots + a_nX^n) + (b_0 + b_1X + \cdots + b_nX^n) \\ = (a_0 + b_0) + (a_1 + b_1)X + \cdots + (a_n + b_n)X^n\end{cases}$$

(165) $$\begin{cases}(a_0 + a_1X + \cdots + a_pX^p)(b_0 + b_1X + \cdots + b_qX^q) \\ = c_0 + c_1X + \cdots + c_{p+q}X^{p+q} \\ c_j = a_0b_j + a_1b_{j-1} + \cdots + a_{j-1}b_1 + a_jb_0\end{cases}$$

(166) $$a_0 + a_1X + \cdots + a_nX_n + a_{n+1}X^{n+1} + \cdots$$

(167) If $xx'' = x''x = 1$, then $x' = x' \cdot 1 = x'(xx'')$ $= (x'x)x'' = 1 \cdot x'' = x''$.

(168) $$\begin{cases}\text{Because } (2 + \sqrt{3})^2 = 7 + 4\sqrt{3}, \\ \text{it follows that } a_2 = 7, \quad b_2 = 4 \text{ and} \\ 7^2 - 3.4^2 = 49 - 48 = 1.\end{cases}$$

(169) $f(x + y) = f(x) + f(y)$, $\quad f(xy) = f(x)f(y)$ for all x, y in R

$g(x) =$ odd, if x is odd, may be defined; then g is a homomorphism of Z into the two-elements ring described in the second example given earlier (see 160).

For a ring homomorphism $f\colon R \to R'$, the image $f(R)$ —*i.e.*, the set of all $f(x)$ with x ranging over R—is a subring of R'; it is a subset of R' stable under sums, differences, and products. The kernel of f, $\ker f$, is the set of all elements x of R such that $f(x) = 0$. Given two elements y, z of R, $f(y) = f(z)$ if, and only if, $(y - z) \in \ker f$. The kernel I of f satisfies the conditions: (K.1) If $x, y \in I$, then $(x + y) \in I$, and $-x \in I$; *i.e.*, I is a subgroup of the additive group of R. (K.2) For any $x \in I$ and any $a \in R$, $ax \in I$ and $xa \in I$.

A subset I of a ring R that satisfies conditions (K.1) and (K.2) is called a two-sided ideal of R; such a subset is called simply ideal when R is commutative.

Conversely, any two-sided ideal I of a ring R is the kernel of some homomorphism of R. Namely, the binary relation $(x - y) \in I$ between elements of R is an equivalence relation. Given two elements a and b of R, the equivalence classes of $a + b$ and ab depend only on the equivalence classes of a and b. Thus the set of equivalence classes denoted by R/I is endowed with an addition and a multiplication. It is a simple matter to check that R/I is a ring; it is called the factor ring of R by I. The mapping, which associates to every $x \in R$ its equivalence class, is a homomorphism $R \to R/I$ of R into R/I, and its kernel is I. Many properties of R, such as commutativity and unit element, are transmitted to R/I. For example, if $n \geq 1$ is an integer, the set Zn of multiples of n in the ring Z of integers is an ideal of Z. The factor ring Z/Zn has n elements, namely, the equivalence classes of $0, 1, 2, \cdots, n - 1$. The invertible elements of Z/nZ are the classes of the integers prime to n. Thus Z/nZ is a field if, and only if, n is a prime number.

Another example involves a set X and a commutative ring K, such as the field of real or complex numbers, the mappings of X into K, called functions, forming a commutative ring for value-wise addition and multiplication (in which addition and multiplication of functions are defined by addition and multiplication of corresponding function values). A subring R of this ring can be considered, such as the subring of continuous, or analytic, or polynomial functions, if it makes sense. If X' is a subset of X, the elements $g \in R$ such that $g(X) = 0$ for every $x \in X'$ form an ideal I of R. The factor ring R/I may be viewed as a ring of functions on X', namely, the restrictions to X' of the functions $y \in R$.

ALGEBRAS

The hypercomplex systems alluded to earlier will now be defined. Consideration is given to a field K (such as the fields of real numbers $\mathfrak{R}$ or complex numbers $\mathfrak{C}$), an

integer $n \geq 0$, and the set K^n of all n-tuples of elements of K. These n-tuples can be added together and multiplied by elements of K, according to the rules that are expressed in terms of addition and multiplication of components (see 170).

The n-tuples form a commutative group for this addition and, furthermore, a vector space over K (see ALGEBRA, LINEAR AND MULTILINEAR). The n-tuples are called a basis that are written with all components 0 except for a single component 1 (see 171). They are expressed as $e_1, e_2, \cdots, e_n$.

Every element $x = (\xi_1, \cdots, \xi_n)$ of K^n can be written in a unique way as $x = \xi_1 e_1 + \cdots + \xi_n e_n$ ($\xi_i \in K$). Thus, if a multiplication distributive with respect to addition and satisfying a specific condition of associativity (see 172) is to be defined on K^n, it is sufficient to define the n^2 products $e_i e_j$ ($1 \leq i, j \leq n$), because then products of n-tuples are expressed in terms of products of the basis elements (see 173). Because $e_i e_j$ is an element of K^n, it may be written in a unique way as a linear combination of basis elements (see 174).

These formulas (174) are called a multiplication table and the elements γ_{ijk} the structural constants. Endowed with its addition and with the multiplication defined above (see 173 and 174), K^n is called an algebra, or a hypercomplex system over K. Here this algebra will be denoted by A; as a mathematical object A is distinct from K^n because one of the many possible multiplications in K^n is given in A. All the ring axioms are true in A, except possibly the associativity of multiplication. For this associativity to be true, it is necessary and sufficient that it be true for the basis elements e_i; that is, that $e_i(e_j e_k) = (e_i e_j)e_k$ for all indices i, j, k. This equality can be translated in terms of structural constants (see 175) by means of the multiplication table (174) for all i, j, k, m. If this property (175) holds, A is a ring and is often called an associative algebra over K. It is commutative if, and only if, $\gamma_{ijk} = \gamma_{jik}$ for all indices i, j, k.

It is possible to define various structures of algebras on any vector space, finite- or infinite-dimensional over K, instead of K^n. Some examples are now given.

Quadratic algebras. Here $n = 2$ and e_1 is required to be the unit element of A. Then the multiplication table takes the form of four equations between two elements (see 176).

If one writes $e_1 = 1$ and $e_2 = f$, the habits about the use of the symbol 1 make it unnecessary to write the three first equations, and one writes only that the second power of f is linear in f (see 177).

Commutativity and associativity of the multiplication are readily checked on 1 and f. Thus A is an associative and associative algebra, called a quadratic algebra over K.

If for K the field $\mathfrak{R}$ of real numbers is taken and $\alpha = -1$, $\beta = 0$, then $f^2 = -1$. Here f is customarily denoted by i, so that $i^2 = -1$. The algebra A then happens to be a field, and this is the field $\mathfrak{C}$ of complex numbers. The invertibility of any nonzero $z = (\lambda + \mu i) \in \mathfrak{C}$ comes from the formula for the product of z and its conjugate (see 178); in fact, since the real numbers λ and μ are not both zero, $\lambda^2 + \mu^2$ is greater than 0, hence z is invertible in $\mathfrak{C}$. More generally, the quadratic algebra defined over a field K by $f^2 = \alpha + \beta f$ ($\alpha, \beta \in K$) is a field if, and only if, the equation $x^2 = \alpha + \beta x$ has no root x in K.

Quaternions. Here $n = 4$, and 1, i, j, k are the basis elements, the first one being the unit element. The essential part is given in a multiplication table (see 179).

It is an easy, but tedious, exercise to check that this multiplication is associative; it is not commutative in general. The corresponding algebra H is called the algebra of quaternions over K. A simple computation (see 180) proves that if in K a sum of four squares is never 0 unless all its terms vanish, the algebra H is a division ring. This condition is the case if $K = \mathfrak{R}$ and H is then the division ring of quaternions, discovered by Hamilton in 1843.

Group algebras. G denotes a finite group with n elements, multiplicatively written. Instead of indexing the components $\xi_1, \cdots, \xi_n$ of an n-tuple $x \in K^n$ by the integers $1, 2, \cdots, n$, one can index them by the elements of G, so that such an n-tuple is now written $\{\xi_g\}$. The basic n-tuple corresponding to a given $g \in G$ is thus $e_g = \{\delta_{gh}\}$, in which $\delta_{gg} = 1$, $\delta_{gh} = 0$ for $h \neq g$. Hence, $\{\xi_g\}$ for all $g \in G$ equals $\Sigma \xi_g e_g$ the summation extending over all $g \in G$. The multiplication table is given by definition (see 181) and the associativity follows from the associativity in G. The corresponding algebra is called the group algebra of G over K, and is denoted by $K[G]$. In view of the multiplication table, it is harmless to write g instead of e_g.

MATRICES; REPRESENTATIONS OF RINGS

Endomorphisms. If V is a vector space over a field K, an endomorphism of V is a mapping u of V into V such that $u(x + y) = u(x) + u(y)$ and $u(dx) = du(x)$ for all x, y in V and d in K. If the addition of two endomorphisms u and v is defined value-wise as $(u + v)(x) = u(x) + v(x)$ and their multiplication as the composition of mappings $(uv)(x) = u \circ v(x)$, the set of all endomorphisms of V becomes a ring. In general, it is noncommutative.

Suppose now that a finite basis $(e_1, \cdots, e_n)$ is given on V, then each x in V can be written in a unique way as $x = e_1\alpha_1 + \cdots + e_n\alpha_n$ ($\alpha_i \in K$). An endomorphism u of V is then uniquely determined by the values $u(e_i)$ ($i = 1, \cdots, n$); if they are written $u(e_i) = \Sigma e_j \alpha_{ji}$, the sum being on j, it can be seen that u is uniquely determined by the n^2 elements α_{ij} of K. It is customary to arrange these elements in a square (see 182).

(170) $$\begin{cases}(\xi_1, \cdots, \xi_n) + (\eta_1, \cdots, \eta_n) = (\xi_1 + \eta_1, \cdots, \xi_n + \eta_n) \\ \alpha(\xi_1, \cdots, \xi_n) = (\alpha\xi_1, \cdots, \alpha\xi_n)\end{cases}$$

(171) $$e_1 = (1, 0, 0, \cdots, 0), \qquad e_2 = (0, 1, 0, \cdots, 0)$$
$$e_n = (0, 0, \cdots, 0, 1)$$

(172) $$\alpha(xy) = (\alpha x)y = x(\alpha y), \quad x, y \in K^n, \quad \alpha \in K$$

(173) $$(\xi_1 e_1 + \cdots + \xi_n e_n)(\eta_1 e_1 + \cdots + \eta_n e_n) = \sum_{i,j} \xi_i \eta_j (e_i e_j)$$

(174) $$e_i e_j = \sum_{k=1}^{n} \gamma_{ijk} e_k, \quad \gamma_{ijk} \in K$$

(175) $$\sum_{n=1}^{n} \gamma_{ijp} \gamma_{pkm} = \sum_{p=1}^{n} \gamma_{ipm} \gamma_{jkp}$$

(176) $$\begin{cases}e_1^2 = e_1 \qquad e_1 e_2 = e_2 e_1 = e_2 \\ e_2^2 = \alpha e_1 + \beta e_2, \quad \alpha, \beta \in K\end{cases}$$

(177) $$f^2 = \alpha + \beta f$$

(178) $$(\lambda + \mu i)(\lambda - \mu i) = \lambda^2 + \mu^2$$

(179) $$\begin{cases}i^2 = j^2 = k^2 = -1, \qquad ij = k, \qquad ji = -k \\ jk = i, \qquad kj = -i, \qquad ki = j, \qquad ik = -j\end{cases}$$

(180) $$(\alpha + \beta i + \gamma j + \delta k)(\alpha - \beta i - \gamma j - \delta k)$$
$$= \alpha^2 + \beta^2 + \gamma^2 + \delta^2$$

(181) $$e_g e_{g'} = e_{gg'}, \quad g, g' \in G$$

(182) $$\begin{cases}\alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \alpha_{n1} & \alpha_{n2} & & \alpha_{nn}\end{cases}$$

The ith column is formed by the components of $u(e_i)$. Such an array is called a square $n \times n$ matrix over K. The matrix is called the matrix of u with respect to the basis $(e_1, \cdots, e_n)$, denoted $M(u)$. Any square $n \times n$ matrix over K is the matrix of some endomorphism u.

Because endomorphisms and matrices are in one-to-one correspondence, the addition and multiplication of endomorphisms can be translated in terms of matrices. One finds easily that the sum of the matrices (α_{ij}), (β_{ij}) is the matrix $(\alpha_{ij} + \beta_{ij})$, and that their product [in the order $(\alpha_{ij}) \cdot (\beta_{ij})$] is the matrix (γ_{ij}) in which $\gamma_{ij} = \Sigma \alpha_{ik}\beta_{kj}$ the sum being on k. In this last formula, the row $(\alpha_{i1}, \cdots, \alpha_{in})$ is multiplied by the column $(\beta_{1j}, \cdots, \beta_{nj})$.

The ring of $n \times n$ matrices over K is denoted by $M_n(K)$. It has no two-sided ideal except itself and 0. It may be viewed as an algebra over K, with n^2 basis elements, namely, the matrices $E_{pq}(1 \leq p, q \leq n)$, in which E_{pq} has a 1 at the intersection of the pth row and the qth column, and zeroes elsewhere. In other words, the matrix $M = (\alpha_{ij})$ is $M = \Sigma \alpha_{pq} E_{pq}$, with summation extending over p and q. The multiplication table can be given (see 183).

Representations. Because rings of matrices are rather simple and have a fruitful geometric interpretation, it is important to compare a given ring with rings of matrices. This is the purpose of representation theory. A representation of a ring R is a homomorphism h of R into a ring of matrices $M_n(K)$. Thus, for $x \in R$, $h(x)$ may be viewed as an endomorphism of the vector space K^n. The representation h is called irreducible if no proper subspace of K^n is stable under all the endomorphisms $h(x)$ $(x \in R)$. The irreducible representations of rings, and in particular of K-algebras, have been extensively and fruitfully studied.

RINGS AS MODULES

In the same way as a group is often considered as acting on a set—that is, as a group of transformations of this set—it is fruitful to make a ring R act on various commutative groups, then called R-modules. This procedure has the advantage of transforming various algebraic problems into linear ones.

More technically, given a ring R, a left R-module (or a left module over R) is a commutative group M (written additively) endowed with a mapping of $R \times M$ into M satisfying axioms to be listed. This mapping associates to every pair (a, x) formed by an element a of R and an element x of M, a well-defined element of M, denoted by $a \cdot x$ or ax. The axioms are three in number (see 184).

If R is unitary, it is usually required that $1 \cdot x = x$ for all $x \in M$; the module M is then called unitary.

A right module is also defined as the element of M associated to the pair (a, x) and is denoted by xa, and the third axiom (see 184) is replaced by $(xb)a = x(ba)$. This is not a simple change of notation; a right module is not in general a left module. The two notions coincide, however, when R is commutative.

If R is a field, the unitary R-modules are the vector spaces. A module over the ring Z of ordinary integers is nothing else than a commutative group G written additively (for $n \in Z$, $n > 0$ and $x \in G$, nx is $x + x + \cdots + x$, n times).

If R is a subring of a ring S, S is both a left and a right A-module; for $a \in R$ and $x \in S$, ax (or xa) is defined to be the product of a and x computed in S. In particular, for $S = R$, R is a left module and a right module over itself.

Given a left module M over a ring R, a submodule M' of M is an additive subgroup M' of M such that $a \in R$ and $x \in M'$ imply $ax \in M'$. The submodules of R considered as a left module over itself are called the left ideals of R; similar definition holds for the right ideals. The two-sided ideals defined in *Definition and examples*, above, are the subsets of R that are at the same time left and right ideals.

Given a submodule M' of a left module M over a ring R, the binary relation $(x - y) \in M'$ between elements x, y of M is an equivalence relation called the relation of congruence modulo M'. It is written $x \equiv y \bmod M'$. Given elements a of R and x and y of M, the equivalence classes of ax and of $x + y$ depend only on a and on the equivalence classes of x and y. Thus the set of equivalence classes, denoted by M/M', is endowed with a structure of left R-module and is called the factor module of M by M'. In particular, if I and I' are left ideals of R such that $I' \subset I$, then I/I' is an interesting example of a left R-module. Similar definitions and results hold for right modules.

A representation h of a ring R in a matrix ring $M_n(K)$ endows the space K^n with a structure of left R-module. Thus module theory and representation theory are essentially equivalent.

(183) $E_{pq}E_{rs} = 0$ if $q \neq r$, $\quad E_{pq}E_{qs} = E_{ps}$ otherwise

(184)
$$\begin{cases} (a+b)x = ax + bx \text{ for all } a, b \in R \text{ and } x \in M \\ a(x+y) = ax + ay \text{ for all } a \in R \text{ and } x, y \in M \\ a(bx) = (ab)x \text{ for all } a, b \in R \text{ and } x \in M \end{cases}$$

(185) $$\frac{a}{b} + \frac{a'}{b'} = \frac{ab' + ba'}{bb'}, \qquad \frac{a}{b} \cdot \frac{a'}{b'} = \frac{aa'}{bb'}$$

(186) $ax = ay, \quad a \neq 0$ imply $x = y$

SPECIAL CLASSES OF RINGS

Domains. It has been shown that a ring may have the unfamiliar feature of zero divisors; that is, nonzero elements a, b such that $ab = 0$. It is thus interesting to define a class of rings closer to computing habits. A commutative ring with unit element $1 \neq 0$ and without zero divisors—*i.e.*, $a \neq 0$ and $b \neq 0$ imply $ab \neq 0$—is called an integral domain, or a domain or an entire ring. The ring Z of integers is a domain. A field is a domain; if $a \neq 0$ and if $ab \neq 0$, multiplication by the inverse a^{-1} of a gives $0 = a^{-1}ab = 1b = b$. If R is a domain, so are the polynomial ring $R[X]$ and the power series ring $R[[X]]$. A subring of a domain, and in particular a subring of a field, is a domain provided it contains the unit element.

Conversely, every domain R is a subring of some field. The elements of this field are constructed in the same way as fractions are constructed from integers. The set of all pairs (a, b) is considered in which a, $b \in R$ and $b \neq 0$. The binary relation $ab' = ba'$ between two such pairs (a, b) and (a', b') is an equivalence relation in this set. The equivalence class of (a, b) is denoted by a/b and called a fraction. The operations of addition and multiplication (see 185) are well defined on the set of fractions and make a field K out of this set; it is called the field of fractions of R. The fractions $a/1$, $a \in R$, form a subring of K, isomorphic with R. The field of fractions of the ring Z of integers is the field Q of rational numbers. In an integral domain R, the cancellation law holds (see 186).

Rings of integers. It was pointed out earlier that ring theory partly originated from arithmetic. It is thus interesting to study the rings in which important properties of the ring Z of integers are valid. The treatment will be restricted to domains.

In a domain R, the divisibility relation $x \mid y$ (x divides y, x is a divisor of y, y is a multiple of x) is defined to mean that there exists $z \in R$ such that $y = zx$. The set Rx of multiples of x is clearly an ideal in R; it is called the principal ideal generated by x. The divisibility relation $x \mid y$ is equivalent with the inclusion relation $Rx \supset Ry$. As in Z, an element d of R is said to be a greatest common divisor, GCD, of x and y if it is a divisor of x and of y, and if it is the greatest in the following sense: every common divisor of x and y divides d. Similarly, a least common multiple, LCM, m of x and y is a common multiple of x and y that divides every other common multiple of x and y. Two elements x, y of R such that $x \mid y$ and $y \mid x$ are called associates; this means that there exists an invertible element u of R such that $y = ux$; whenever divisibility is concerned, associates have the

same properties. All the GCD's (or LCM's) of two given elements of R, if they exist, are associates. In an arbitrary domain R, LCM's and GCD's need not exist. To say that the elements x and y of R admit an LCM amounts to saying that the ideal $Rx \cap Ry$ is principal. If any two elements x, y of R admit an LCM, say m, they admit a GCD, say d, and conversely; furthermore, xy and md are associates.

Irreducible elements

An element $p \neq 0$ of a domain R is called irreducible if it is not invertible and if the only divisors of p are either invertible or associates of p. In Z, the irreducible elements are the prime numbers and their negatives. In a domain R in which LCM's exist, an irreducible element p enjoys the following property:

(A) If p divides a product ab ($a, b \in R$), it divides one of the factors a, b.

In the ring S formed by the complex numbers including the square root of -5 as a coefficient of one of two elements from Z (see 187), the number 2 is irreducible, but a relation (see 188) shows that (A) does not hold. Thus LCM's do not always exist in S.

In Z the unique decomposition of an integer into prime factors is well-known and quite useful. Thus it is interesting to study the domains R in which every element is a product of irreducible ones in an essentially unique way; *i.e.*, unique up to order and invertible factors. The existence of such a decomposition is usually easy to settle: it is implied by the following property (called by algebraists a finiteness condition):

(B) Any strictly increasing sequence $Ra_1 \subset Ra_2 \subset \cdots \subset Ra_n \subset \cdots$ of principal ideals of R (in which equality between adjacent members of the sequence is excluded) is finite.

When existence is proved, essential uniqueness is implied by property (A).

Gaussian domains

The domains in which the essentially unique decomposition into irreducible elements hold are called Gaussian or factorial domains, or unique factorization domains (UFD). Their definition may be restricted as follows: there exists a subset P of R such that every nonzero element x of R can be uniquely written as: $x = up_1^{n(1)} \cdots p_q^{n(q)}$ with $p_1, \cdots, p_q$ distinct elements of P, u invertible and positive integral exponents $n(1), \cdots, n(q)$.

The elements of P are irreducible; more precisely, P contains one, and only one, representative of each class of irreducible associates. In Z, the set of positive prime numbers may be taken for P. The divisibility theory of a factorial domain R closely resembles that of Z; in particular, LCM's and GCD's exist and can be computed from the decomposition into irreducible factors by the usual rules. The finiteness condition (B) holds. An important theorem, essentially due to Gauss, says that, if a domain R is factorial, so is the polynomial ring $R[X]$; but the power series ring $R[[X]]$ need not be factorial.

An important subclass of factorial domains is formed by the principal domains, or principal ideal domains, PID; *i.e.*, the domains R in which every ideal is principal. In fact, because the ideals $Rx \cap Ry$ are principal, LCM's exist, and the finiteness condition (B) is easy to prove (consider the union of all Ra_j's, which is an ideal). These PID's are very nice rings: many important facts about their ideals and their modules are known. The ring Z, a polynomial ring $K[X]$ over any field K are PID's, but $Z[X]$ is factorial without being a PID.

These facts about Z and $K[X]$ are usually proved by Euclidean division. More generally, a division algorithm on a domain R is a mapping ϕ of R into the nonnegative integers such that, given any a, b in R, $b \neq 0$, there exist q and r in R with $a = bq + r$ and $\phi(r) < \phi(b)$.

Euclidean domains

If a domain admits a division algorithm, it is called Euclidean. The domains Z and $K[X]$ are Euclidean, the algorithms being easily expressed (see 189). Any Euclidean domain R is a PID: if I is any nonzero ideal in R, take a nonzero element a in I such that $\phi(a)$ is minimum; then $I = Ra$ is easily proved.

If $d \in Z$ is any squarefree integer (not divisible by the square of any prime) then for $d \equiv 2$ or 3 modulo 4, R_d is defined as the set of complex numbers $a + b\sqrt{d}$ ($a, b \in Z$). For $d \equiv 1$ modulo 4, R_d is defined as the set of numbers $a + b\frac{1+\sqrt{d}}{2}$ ($a, b \in Z$). In both cases R_d is a domain. An interesting problem is to find out whether R_d is principal or Euclidean. It can be proved that R_d is factorial if, and only if, it is principal. For $d \equiv 1$ modulo 4 the seemingly more natural set R'_d of all $a + b\sqrt{d}$ ($a, b \in Z$), a subring of R_d, is never principal, and less interesting for the purpose (the proof uses the fact that the number $x = \frac{1+\sqrt{d}}{2}$ is a root of an equation with integer coefficients (see 190).

The above problem is solved for $d < 0$. The domain R_d is Euclidean only for $d = -1, -2, -3, -7$, and -11; the proof is technical but not too hard. It is principal without being Euclidean for $d = -19, -43, -67, -163$, and nonprincipal for all other negative values of d; here the proof is very hard and has been completed only after 1960.

Much less is known for $d > 0$. By explicit computations, the values of d for which R_d is principal have been determined at least for $2 \leq d \leq 499$. No general theorem, however, indicates whether R_d is principal for finitely many or infinitely many values of d. The ring R_d is Euclidean for the corresponding algorithm (see 191) for only 16 explicitly known values of d. For other values, however, it might perhaps be Euclidean for another algorithm; the problem is open.

At any rate, many nonprincipal domains R_d are known; for example, $d = -5$, $d = -6$, $d = 10$, $d = 26$. A similar situation arises in the rings of so-called cyclotomic integers connected with Fermat's equation $x^n + y^n = z^n$. This loss of unique factorization led Kummer and Dedekind to introduce the notion of an ideal and to prove that, in a class of domains now called the Dedekind domains that includes the rings R_d and the PID's, every ideal admits a unique decomposition as a product of prime ideals. A well-developed arithmetical theory of Dedekind domains is available.

Noetherian rings. Algebraic geometry motivated the introduction of a larger class of commutative rings, called Noetherian. They are the rings R that satisfy the equivalent conditions:

(N) Every strictly increasing sequence of ideals of R is finite.

(N′) Every nonempty family of ideals of R admits a maximal element (for the inclusion relation).

(N″) Every ideal I of R is finitely generated; that is, there exists a finite number of elements $x_1, \cdots, x_n$ of I such that every element y of I can be written as $y = r_1x_1 + \cdots + r_nx_n$ with $r_1, \cdots, r_n \in R$. It is technically useful to introduce also the Noetherian modules: replace ideal by submodule in (N), (N′), (N″). The class of Noetherian rings is rather large. It includes the Dedekind domains, whence the PID's. Furthermore, it is stable by many ring theoretic operations. For example, the celebrated finite basis theorem of Hilbert states that, if R is Noetherian, so is the polynomial ring $R[X]$; the same holds for power series. By induction it can be shown that a polynomial ring in any finite number of variables over a field is Noetherian, whence the importance of Noetherian rings

(187) $x + y\sqrt{-5}, \quad x, y \in Z$

(188) $(1 + \sqrt{-5})(1 - \sqrt{-5}) = 6 = 2 \cdot 3$

(189)
$$\begin{cases} \phi(n) = |n|, & n \in Z \\ \phi[P(X)] = 1 + \text{degree of } P(X) \text{ and } \phi(0) = 0 \\ P(X) \in K[X], \qquad P(X) \neq 0 \end{cases}$$

(190) $x^2 - x - \frac{d-1}{4} = 0$

(191) $\phi(x + y\sqrt{d}) = |x^2 - dy^2|, \quad x, y \in Q$

(192) $gf: A \to C$

(193) $h(gf) = (hg)f$

(194) $f1_A = f, \quad 1_A g = g$

(195) $\mathfrak{F}(A, B)$

(196) $\mathfrak{F}(A, B) \times \mathfrak{F}(B, C) \to \mathfrak{F}(A, C)$

(197) $f1_A = f, \quad 1_A g = g$

in algebraic geometry. The theory of Noetherian rings is quite extensive.

Division rings. A nonzero ring in which every nonzero element admits an inverse, it may be recalled, is called a division ring; if it is commutative, it is called a field. Division rings are the simplest noncommutative rings, fields the simplest commutative ones: for example, they have no other ideals than the whole ring and the zero ideal; whence their divisibility theory is trivial (see above *Fields*). Here a few words will be said about division rings.

The simplest example of a noncommutative division ring is the ring H of quaternions. It contains a subfield, formed by the quaternions $a \cdot 1 + 0 \cdot i + 0 \cdot j + 0 \cdot k$, that is identified with the field $\mathfrak{R}$ of real numbers. For every $a \in \mathfrak{R}$ and every $x \in H$, $ax = xa$.

This introduces the following notion: given a ring S, the set of all elements a of S such that $ax = xa$ for every $x \in S$ is called the centre of S. It contains 0, also 1 if S is unitary, and is a subring of S, necessarily commutative. If S is a division ring, then $a \in Z$, $a \neq 0$ imply $a^{-1} \in Z$, so that the centre Z is a subfield of S. For the ring H of quaternions, the centre is exactly the field of real numbers.

Thus a division ring S is a vector space over its centre Z. The dimension of this vector space may be finite or infinite, the first case being better known than the other. In this case the dimension is the square n^2 of an integer n; *e.g.*, $4 = 2^2$ for the quaternions. Given a field Z, the set of all division rings D (up to isomorphism) admitting Z as centre and finite-dimensional over Z can be organized by a rather technical process into a commutative group $\mathrm{Br}(Z)$, called the Brauer group of Z. This group is a powerful instrument for studying the field Z. It is trivial for the field of complex numbers and is a two-elements group for the field of real numbers.

A beautiful theorem of the Scottish mathematician Joseph Henry Maclagan Wedderburn says that every finite division ring is commutative; in other words, the Brauer group of a finite field is trivial. Because every element x of a finite division ring D with q elements satisfies $x = x^q$, Wedderburn's theorem is a particular case of the following generalization: if, for every element x of a ring R, there exists an exponent $n(x) \geq 2$ such that $x^{n(x)} = x$, then R is commutative. (P.S.)

V. Categories

Since the mid-1940s mathematicians have found it valuable to formalize the notions of areas of mathematical discourse and interrelations between such areas; the formalization used is the language of categories and functors. A rapidly developing mathematical theory accompanied this formalization so that the theory of categories and functors has become an autonomous part of mathematics. This section deals with the role of this formalization in providing an appropriate language for the expression of mathematical ideas.

The study of mathematics involves certain domains of mathematical discourse, their structural properties, and their interrelations. For example, there is the idea of a set (see SET THEORY) and of functions that are transformations of sets. There is arithmetic, the study of the domain of natural numbers, rational numbers, and integers. Geometry involves subsets of Euclidean space of three dimensions together with the appropriate transformations of such subsets—for example, translations, rotations, and reflections. Algebra is concerned with rational numbers, real numbers, and complex numbers; and abstract algebra involves further mathematical domains such as groups, rings, fields, and lattices. The calculus is again concerned in the first instance with subsets of the real numbers, or of Euclidean space of higher dimension, and with particular transformations of such subsets; for example, differential operators.

Algebraic topology relates domains of interest in geometry to domains of interest in algebra. Algebraic geometry, on the other hand, goes in the opposite direction, associating, for example, with each commutative ring its spectrum of prime ideals.

Set theory is concerned with a class of objects $A, B, C, \cdots$, called sets, and a class of transformations $f, g, h, \cdots$, called functions. With each function f is associated a domain, which is a set A, and a range or codomain, which is a set B. The notation $f: A \to B$, or $A \xrightarrow{f} B$, indicates that f is a function from the domain A to the codomain B. There is a strict distinction between the codomain of the function f and the image of the function f, which is simply the set of values taken by the function f. Thus, in particular, two functions can only be identical if not only their domains but also their codomains coincide. They are then, of course, identical if, and only if, they each take the same value at each element of the domain.

Further, functions may be composed under certain conditions. To be precise, the functions $f: A \to B$ and $g: B' \to C$ may be composed to yield a function, written gf or $g \circ f$ (see 192), if, and only if, $B = B'$. Further, the law of composition is associative (see 193), provided, of course, that the relevant compositions are defined. With each set A may be associated its identity function 1_A: $A \to A$. This function 1_A has basic properties (see 194) provided the compositions are defined. Thus, 1_A behaves very much like the integer 1 in the ring of integers, a fact that leads to the habit of dropping the subscript A and simply writing $1: A \to A$. It is an easy exercise to see that the identity function is entirely characterized by the properties of 1_A (see 194). That is to say, if the function $u: A \to A$ also satisfies the conditions $fu = f$, $ug = g$ for all appropriate f, g, then indeed $u = 1_A$.

Of course set theory has more structure than this, but if the ideas described above are abstracted from set theory, a model for the notion of a category is obtained.

THE NOTION OF A CATEGORY

A formal definition can be given in the following way: A category $\mathfrak{F}$ consists of three sets of data:

- C1. There is a class of objects $A, B, C, \cdots$.
- C2. To each ordered pair of objects A, B in $\mathfrak{F}$, there is associated a set (see 195) called the set of morphisms, or maps or transformations, from A to B in $\mathfrak{F}$.
- C3. To each ordered triple of objects A, B, C in $\mathfrak{F}$, there is associated a law of composition, or composition function (see 196), the image of (f, g), $f \in \mathfrak{F}(A, B)$, $g \in \mathfrak{F}(B, C)$, under this law of composition being written gf or $g \circ f$, so that gf is a morphism from A to C.

These data satisfy the following three axioms, of which the first is in the nature of a convention, while the remaining two are more substantial:

- C4. $\mathfrak{F}(A_1, B_1)$ and $\mathfrak{F}(A_2, B_2)$ are disjoint unless $A_1 = A_2$, $B_1 = B_2$.
- C5. (Associative law) $h(gf) = (hg)f$, provided the compositions are defined.
- C6. (Existence of identities). To each object A of $\mathfrak{F}$ there is associated a morphism $1_A \in \mathfrak{F}(A, A)$ such that two equations hold (see 197) provided the compositions are defined.

As a first example, the category $\mathfrak{S}$ is considered, called the category of sets and functions, or simply the category of sets. Precisely, the objects of $\mathfrak{S}$ are sets $A, B, C, \cdots$. A morphism in $\mathfrak{S}$ from the set A to the set B is merely a

function with domain A and codomain (range) B. Thus either of two notations (see 198) can be adopted in an arbitrary category as a convenient representation of the statement $f \in \mathfrak{F}(A, B)$. The law of composition in $\mathfrak{S}$ is simply the familiar composition of functions. The axioms are then certainly satisfied. Moreover, this example serves to explain the notation gf for the composition of f: $A \to B$ with g: $B \to C$. It is usual to write functions on the left of their arguments, so that the image of $a \in A$ under the composite of f and g appears as $g(f(a))$. A less conservative attitude, defying this tradition and leading to a happier notation, would be to write fg for the composite of f and g, especially in view of the natural notational convention (see 199). In this article, however, tradition is not defied.

Just as in the special case of the category $\mathfrak{S}$, the equations relating to the identities (197) entirely determine the morphism 1_A in $\mathfrak{F}(A, A)$; 1 is often written for 1_A.

Some further examples of categories are now listed, although the concept is so general that the list could be continued indefinitely: finite sets and functions; groups and homomorphisms; Abelian groups and homomorphisms; rings and homomorphisms; subsets of Euclidean space of 3 dimensions and Euclidean movements; subsets of Euclidean space of n dimensions and continuous functions; topological spaces and continuous functions.

The law of composition is not specified explicitly in describing these categories. This is the custom when the objects have underlying set-structure, the morphisms are functions of the underlying sets (transporting the additional structure), and the law of composition is merely ordinary function-composition. Indeed, sometimes even the specification of the morphisms is suppressed if no confusion would arise—thus one speaks of the category of groups.

The examples given suggest a conceptual framework. For example, the concept of group may be regarded as constituting a first-order abstraction or generalization from various concrete, familiar realizations such as the additive group of integers, the multiplicative group of nonzero rationals, groups of permutations, symmetry groups, groups of Euclidean motions, and so on. Then, again, the notion of a category constitutes a second-order abstraction, the concrete realizations of which consist of such first-order abstractions as the category of groups, the category of rings, the category of topological spaces, and so on.

It should not be supposed that, in every category, the objects are sets (probably with additional structure) and that the morphisms are certain preferred functions. One example serves to dispel this misconception. X is a set with a pre-ordering relation $\geq$. Thus the relation $a \geq b$ holds for certain elements a, b of X and the following axioms are satisfied: $a \geq a$; if $a \geq b$, $b \geq c$, then $a \geq c$. For example, the integers may be ordered by size or they may be preordered by the divisibility condition: $a \mid b$ if a is a factor of b. Then if $(X, \geq)$ is a set with a pre-order, a category $\mathfrak{C}_X$ is formed, the objects of which are the elements of X and such that $\mathfrak{C}_X(a, b)$ is the single element $a \to b$ if $a \geq b$ and is empty otherwise. There is evidently a unique law of composition and the category axioms obviously hold.

A morphism f: $A \to B$ in the category $\mathfrak{F}$ is said to be invertible, or a unit, or an equivalence, if there is a morphism g: $B \to A$ in $\mathfrak{F}$ with $gf = 1_A$, $fg = 1_B$. It is easy to prove that g is then invertible and is determined by f (g is the inverse of f, written $g = f^{-1}$), and that if A is isomorphic to B, there existing an invertible f: $A \to B$, then the relation of isomorphism is an equivalence relation on the objects of $\mathfrak{F}$. Thus every category carries automatically with it a notion of the isomorphism of objects, proper to that category. The concept gives a great gain in universality compared with traditional procedures, which involve the definition, as distinct concepts, of one-one correspondence between sets, isomorphism between groups, isomorphism between rings, homeomorphism between topological spaces, bi-continuous isomorphism between topological groups, order type of ordered sets, and so on. It is particularly offensive, from the categorical point of view, to define an isomorphism of groups ϕ: $G \to H$ to be a homomorphism that is one-one and onto H; for an isomorphism of groups should be just an invertible morphism of the category of groups. It is a theorem that, in this category, a morphism that is one-one and onto its codomain is an isomorphism. This theorem is false in the category of topological spaces, so that the definition masks the universality of the concept of isomorphism.

$$f: A \to B \quad \text{or} \quad A \overset{f}{\to} B \tag{198}$$

$$A \overset{f}{\to} B \overset{g}{\to} C \tag{199}$$

$$F(fg) = (Ff)(Fg) \tag{200}$$

$$F(1_A) = 1_{FA} \tag{201}$$

FUNCTORS

The notion of category being established as that which gives precision to the concept of domain of mathematical discourse, the formalization of the precise notion corresponding to the intuitive idea of the interrelation or connection between different domains is now considered. Within any particular category not only the objects must be specified but also the morphisms. Furthermore, these morphisms are, in the principal examples, required in some sense to transport the structure characteristic of the objects of the category; *e.g.*, group structure, topological structure. In the same way a functor F from the category $\mathfrak{F}$ to the category $\mathfrak{G}$ is defined as a rule that associates with every object A of $\mathfrak{F}$ an object FA of $\mathfrak{G}$, and with every morphism f: $A \to B$ in $\mathfrak{F}$ a morphism Ff: $FA \to FB$ in $\mathfrak{G}$, subject to the conditions of transport of structure (see 200, 201).

In expressing the relation (200), there is a tacit assumption that the composition fg is defined. This is standard practice, to avoid unnecessarily complicated statements.

A functor F always sends isomorphisms to isomorphisms, a very crucial property. Otherwise expressed, if Ff is not an isomorphism, then f cannot be an isomorphism. This yields a basic methodological procedure; to demonstrate that A and B, two objects of $\mathfrak{F}$, really are different—*i.e.*, non-isomorphic—it may be possible to find a functor F from $\mathfrak{F}$ to a category $\mathfrak{G}$ in which it is far easier to make the comparison. This procedure may be regarded as a massive generalization of the classical arithmetical test of casting out nines.

Some examples of functors are now given:

Examples of functors

(a) For any group G it is possible to render G commutative by adjoining the relations $xy = yx$ for all x, y in G. The resulting group may be written G_{ab}. If ϕ: $G \to H$ is a homomorphism, it is easy to see that ϕ induces a uniquely determined homomorphism ϕ_{ab}: $G_{ab} \to H_{ab}$. Thus there is a functor from the category of groups to the category of Abelian groups. This should be the first functor tried in an attempt to prove two groups non-isomorphic.

(b) For any pointed topological space (X, x)—that is, a topological space X, together with a distinguished point x in X—it is possible to construct the Poincaré fundamental group $\pi(X, x)$. Then π is a functor from the category of pointed topological spaces and pointed continuous functions to the category of groups. Because, for example, the sphere and the torus (a surface in the shape of a doughnut) have non-isomorphic fundamental groups, it follows that they are not homeomorphic.

(c) If X, Y are two pre-ordered sets regarded as categories, as above, then a functor from X to Y is merely an order-preserving function.

(d) There are many important functors, called forgetful or underlying functors, that simply forget part of the structure present in the objects and transported by the morphisms. Thus, the objects in the examples in *The*

notion of a category all have underlying set structures and the morphisms are all functions; moreover, the law of composition is simply that of functions. Thus, in all these cases, there is an underlying functor to $\mathfrak{S}$. It is also possible to take a ring, forget the multiplication in it, and thereby obtain an Abelian group; this yields an underlying functor. Underlying functors may seem trivial things; actually, they are fundamental to categorical language.

NATURAL TRANSFORMATIONS

One further basic notion in the theory of categories, or, as it may be said, a basic item of categorical language, will now be introduced. This is the notion of a natural transformation of functors from one category to another. Indeed, the whole language and apparatus of categories and functors were developed initially by the U.S. mathematicians Samuel Eilenberg and Saunders MacLane in order to render precise the intuitive concept of naturality. First an example will be given, the example that may be said to have motivated the definition.

Let V be a vector space over some field K, and let V^* be the dual space of V; that is, the space of linear functionals on V. There is then a linear transformation T: $V \to V^{**}$ that is given (see 202). There is an intuitive feeling that the linear transformation T is natural because its description only involves the terms v and ϕ. Now if V is finite-dimensional, then it is known that T is an isomorphism in the category of vector spaces over K and linear transformations. One way of proving that T is then an isomorphism is to show that V and V^{**} are isomorphic and then to observe that T is one-one. The usual proof that V and V^{**} are isomorphic would be to proceed by establishing an isomorphism between V and V^*, in the case when V is finite-dimensional. Now if a base $(e_1, e_2, \cdots, e_k)$ for V is given, then a basis for V^* may be set up, called the dual basis, by defining e_i^* to be that linear functional on V given by certain rules (see 203). Then the correspondence $e_i \leftrightarrow e_i^*$ sets up an isomorphism between V and V^*.

On the other hand, this isomorphism does not look natural, because it depends on the choice of bases. Of course, the argument above could be generalized to set up a linear transformation from V to V^* even if V is not finite-dimensional over K, but, again, this transformation would not appear to be natural. What is required is a formal and precise expression of the feeling that, for finite-dimensional vector spaces V over the field K, V and V^{**} are naturally isomorphic, while V and V^* are isomorphic in some unnatural way. Eilenberg and MacLane solved this problem in their seminal article, "General Theory of Natural Equivalences," published in 1945, which may be said to have laid the foundation of the theory of categories. The precise definition is:

If F, G: $\mathfrak{F} \to \mathfrak{G}$ are two functors from the category $\mathfrak{F}$ to the category $\mathfrak{G}$, then a natural transformation T, from F to G, is a rule assigning to each object A of $\mathfrak{F}$ a morphism T_A in the set $\mathfrak{G}(FA, GA)$, subject to the condition that a diagram (see 204) should be commutative for every f: $A \to B$ in $\mathfrak{F}$; that is, $Gf \circ T_A = T_B \circ Ff$. Further, if each T_A is invertible, T is said to be a natural equivalence. It is clear that if T is a natural equivalence from F to G, then T^{-1}, given by an equation (see 205), is a natural equivalence from G to F. Thus the term equivalence used here is fully justified. Indeed, the functors from $\mathfrak{F}$ to $\mathfrak{G}$ may be collected into equivalence classes according to the existence of a natural equivalence between them.

This definition can be tested against the example. There are two functors from $\mathfrak{V}$ to $\mathfrak{V}$, in which $\mathfrak{V}$ is the category of vector spaces over the field K and linear transformations. One functor is the identity functor. The other functor is the double dual functor **: $\mathfrak{V} \to \mathfrak{V}$ that associates with every vector space V its double dual V^{**} and with every linear transformation f: $U \to V$ in $\mathfrak{V}$ the linear transformation f^{**}: $U^{**} \to V^{**}$ (see 206). A linear transformation T: $V \to V^{**}$ was described above. If it is now written as T_V, it is easy to check that T is a natural transformation from the identity functor to the functor **. If the subcategory $\mathfrak{V}_f$ of $\mathfrak{V}$ that consists of finite-dimensional vector spaces over K and their linear transformations is considered, then it turns out that the functor ** transforms $\mathfrak{V}_f$ into itself; and the natural transformation T, restricted to $\mathfrak{V}_f$, is then a natural equivalence. Further examples of natural transformations of functors can be given:

(a) The category $\mathfrak{A}$ of Abelian groups and homomorphisms is considered. With every Abelian group may be associated its torsion subgroup. The torsion subgroup A_t of the Abelian group A consists of those elements of A that are of finite order. A homomorphism from A to B must necessarily send A_t to B_t. Thus a functor F is obtained from $\mathfrak{A}$ to $\mathfrak{A}$ (or to $\mathfrak{A}_t$, the category of torsion Abelian groups and their homomorphisms), by associating with every Abelian group A the Abelian group $FA = A_t$. Now A_t is a subgroup of A. Thus there is always an embedding i_A of A_t in A. It is easy to see that i is a natural transformation from the torsion functor F to the identity functor. Further, the quotient group $A_{fr} = A/A_t$ may be considered. It is a torsion-free Abelian group. This gives a functor G from $\mathfrak{A}$ to $\mathfrak{A}$ (or from $\mathfrak{A}$ to $\mathfrak{A}_{fr}$, the category of torsion-free Abelian groups) by associating with the Abelian group A the Abelian group $GA = A_{fr}$. Then the projection of A onto A_{fr} yields a natural transformation from the identity functor to the torsion-free functor G.

(b) With every group may be associated its commutator subgroup. It is then not difficult to see that the embedding of the commutator subgroup in the group is a natural transformation from the commutator subgroup functor to the identity functor. On the other hand, the centre of every group may be associated with the group. Here, however, there is not a functor because a homomorphism from one group to another does not necessarily map the centre of the first group to the centre of the second. On the other hand, if the category of groups and surjective homomorphisms (a surjective homomorphism is one in which the image coincides with the codomain) is considered, then in this category the centre is a functor. It is a functor, however, from the category of groups and surjective homomorphisms to the category of groups and all homomorphisms, because a surjective homomorphism does not necessarily map the centre surjectively. Then the embedding of the centre of a group in the group may be regarded as a natural transformation from the centre functor to the inclusion functor, both of which are functors from the category of groups and surjective homomorphisms to the category of groups and homomorphisms.

(c) In algebraic topology, the singular homology groups and the homotopy groups of a pointed topological space (X, x) are considered. A Hurewicz homomorphism (see 207) exists, from the homotopy groups to the homology groups. Then π_n and H_n, $n \geq 2$, are functors from the category of pointed spaces and pointed continuous func-

(202) $T(v)(\phi) = \phi(v), \quad v \in V, \quad \phi \in V^*$

(203) $e_i^*(e_i) = 1, \qquad e_i^*(e_j) = 0, \qquad i \neq j$

(204)
$$\begin{array}{ccc} FA & \xrightarrow{T_A} & GA \\ {\scriptstyle Ff}\downarrow & & \downarrow{\scriptstyle Gf} \\ FB & \xrightarrow{T_B} & GB \end{array}$$

(205) $(T^{-1})_A = T_A^{-1}$

(206)
$$\begin{cases} f^{**}(u^{**})(v^*) = u^{**}(f^*(v^*)) \\ u^{**} \in U^{**}, \quad v^* \in V^* \\ f^*(v^*)(u) = v^*(f(u)), \quad u \in U \end{cases}$$

(207) θ_n: $\pi_n(X, x) \to H_n(X, x)$

tions to the category of Abelian groups, and the Hurewicz homomorphism is a natural transformation of functors.

UNIVERSAL CONSTRUCTIONS

An important procedure that is made possible by the use of categorical language is the relation of apparently distinct concepts in various categories. Here an example is given, but it should be pointed out that the technique described is of extremely broad application.

The Cartesian product

In the category of sets, the notion of Cartesian product is vital. Its description, however, involves mention of the elements of the sets: the Cartesian product of the sets X and Y consists of the ordered pairs (x, y), $x \in X$, $y \in Y$. In a definition of the product of two objects in an arbitrary category, this definition cannot be imitated because, in the definition of a category, it is not required that the objects should be sets with elements. Therefore, a property of the Cartesian product $X \times Y$ must be found that characterizes the Cartesian product and that can be expressed in the language of categories. Such a property must therefore be entirely expressible in terms of the data of a category; that is, in terms of objects, morphisms, and composition of morphisms.

(208) $p_1 h = f, \qquad p_2 h = g$

(209) $p_1' \omega = p_1, \qquad p_2' \omega = p_2$

(210) $(X_1 \times X_2) \times X_3 \cong X_1 \times (X_2 \times X_3)$

(211) $gcd(gcd(a, b), c) = gcd(a, gcd(b, c))$

(212) $\mathfrak{F}^0(A, B) = \mathfrak{F}(B, A)$

For any given set A and functions $f\colon A \to X$, $g\colon A \to Y$, there is a unique function $h\colon A \to X \times Y$ such that equations involving two projections hold (see 208), in which $p_1\colon X \times Y \to X$, $p_2\colon X \times Y \to Y$ are the projections. This is the clue to the generalization. For two objects X and Y of the category $\mathfrak{F}$, the triple $(Z; p_1, p_2)$ is taken to be a product of X and Y if $p_1\colon Z \to X$ and $p_2\colon Z \to Y$ are morphisms of $\mathfrak{F}$, and if the universal property holds that, given any object A of $\mathfrak{F}$ and morphisms $f\colon A \to X$, $g\colon A \to Y$, there exists a unique morphism $h\colon A \to Z$ such that certain relations hold (see 208).

It is easy to prove that the triple $(Z; p_1, p_2)$ is, essentially, uniquely determined by the universal property. This means that if $(Z'; p_1', p_2')$ is also a product, then there exists a unique isomorphism $\omega\colon Z \to Z'$ in $\mathfrak{F}$ such that two equalities hold (see 209). Therefore, it is permissible to speak of *the* product of X and Y in $\mathfrak{F}$. Of course, X and Y do not necessarily have a product in $\mathfrak{F}$. If they do have a product in $\mathfrak{F}$, in the sense defined above, however, then that product is unique. Now the product may be sought in various other categories. For example, in the category of Abelian groups the product, as defined, characterizes the direct sum; in the category of topological spaces it characterizes the topological product. In any category corresponding to a pre-ordered set, the product characterizes the greatest lower bound. Thus, in particular, in the category corresponding to the natural numbers ordered by divisibility, the product is precisely the gcd. It is possible to prove that the product in any category in which it exists is associative in the following sense. Given three objects X_1, X_2, and X_3, then there is an equivalence (see 210). This is, of course, a completely familiar—and trivial—fact for sets: in considering the Cartesian product of three sets, the way in which the sets are to be associated is not relevant. If a proof has been given that is valid in any category, however, it is immediately possible to deduce, for example, that the greatest common divisor satisfies associativity. That is, if a, b, c are any three natural numbers, then a relation involving the gcd holds (see 211). Here there are two points to be made. First, there was no apparent connection between this statement (211) and the statement that the Cartesian product of sets is associative. Second, by proving the statement in sufficient generality—*i.e.*, at the categorical level—it is possible to obtain many more instances of the theorem than could have been obtained by simply being confined to the original category in which the definition of product was carried out; that is, the category of sets.

THE DUALITY PRINCIPLE

For any category $\mathfrak{F}$ a new category $\mathfrak{F}^0$ can be formed by interchanging the domains and codomains of the morphisms of $\mathfrak{F}$. More precisely, in the category $\mathfrak{F}^0$ the objects are simply those of $\mathfrak{F}$ and the effect of interchange of domains is expressed in an equation (see 212). Moreover, the composition in $\mathfrak{F}^0$ is simply that of $\mathfrak{F}$, suitably interpreted. $\mathfrak{F}^0$ is called the category opposite to $\mathfrak{F}$; notice that $(\mathfrak{F}^0)^0 = \mathfrak{F}$. This apparently trivial operation leads to highly significant results when specific categories are used. In the general setting it enables any concept in the language of categories to be dualized. For example, the coproduct in $\mathfrak{F}$ is simply the product in $\mathfrak{F}^0$. Any theorem that holds in an arbitrary category has a dual form. For example, the theorem asserting that the product in an arbitrary category is associative may be effectively restated as asserting that the coproduct in an arbitrary category is associative. In the special cases, however, the second statement looks very different from the first. For example, in the category of sets, the coproduct becomes the disjoint union; in the category of groups it is the free product; and in a pre-ordered set regarded as a category, the coproduct is the least upper bound. In particular, for the set of natural numbers, ordered by divisibility, the coproduct is the lcm. Thus, the same universal argument that led to the deduction that the gcd is associative also indicates that the lcm is associative. The duality principle has very wide ramifications indeed. Here it is merely noted that the important concept of a contravariant functor from $\mathfrak{F}$ to $\mathfrak{G}$ may be most simply defined as a functor from $\mathfrak{F}^0$ to $\mathfrak{G}$. Thus the association of the dual vector space V^* with V yields a contravariant functor from $\mathfrak{V}$ to $\mathfrak{V}$.

CONCLUSIONS

Because the concept of a category is so general, it is to be expected that theorems provable for all categories will not usually be very deep. Consequently, many theorems of category theory are stated and proved for particular classes of categories. For example, homological algebra is concerned with Abelian categories, which exhibit features suggested by the category of Abelian groups. In mathematics one seeks generality and depth, but there is a certain conflict between these two aims because the more general the setting, the less likely that the results will be profound. The art of mathematics might be said to be that of finding the best possible compromise between these two aims. Further, a not unimportant purpose of the language of categories and categorical reasoning is to identify within a given argument that part which is trivial and separate it from the part which is deep and proper to the particular context. For example, in the study of the theory of the gcd, the fact that it is essentially unique simply follows from the uniqueness of the product in any category and is thus really trivial. Similarly, as has been demonstrated, the fact that it is associative is trivial. On the other hand, the fact that the gcd of the integers a and b can be expressed as a linear combination of a and b with integer coefficients—$gcd(a, b) = ma + nb$, m, n integers—is a much deeper fact that is special to the particular situation, or, at least, to a highly restricted class of situations.

VI. Homological algebra

Development of homological algebra

The origins of homological algebra are to be found in algebraic topology. The mathematician W. Hurewicz pointed out in 1936 that if X is a connected polyhedron that is aspherical in all dimensions—that is, any map of sphere S^n into X, $n \geq 2$, is contractible—then the homol-

ogy groups, H_nX, of the space X are entirely determined by $\pi = \pi_1 X$, the fundamental group of X. Indeed, the homotopy type of X is determined by π. More generally, consider the Hurewicz homomorphism (207) from the kth homotopy group of X to the kth homology group. If a condition holds (see 213), π entirely determines the quotient of H_nX by the subgroup $\theta_n\pi_nX$, called the subgroup of spherical cycles. The mathematician H. Hopf investigated the nature of this dependence (see 214) and sought a purely algebraic description. In 1941, he was led in his work on the influence of the curvature of a closed Riemannian manifold on its topology to observe that the phenomenon under consideration could be expressed as a connection between π_1X and H_2X, valid for any connected polyhedron X; in fact, he produced a description of $H_2X/\theta_2\pi_2X$. For simplicity it is supposed here that X is aspherical in all dimensions; following Eilenberg–MacLane, one writes $X = K(\pi, 1)$, in which $\pi = \pi_1X$. This notation is justified because the homotopy type of X is determined by π. $K(\pi, 1)$ is called the Eilenberg–MacLane space of type $(\pi, 1)$. Note that a set of conditions holds (see 215). Then Hopf produced a formula for $H_2(K(\pi, 1))$ in terms of a free presentation of the group π. It is presented as the quotient of a free group F by a normal subgroup R; thus $\pi = F/R$. Writing $H_2\pi$ for $H_2(K(\pi, 1))$, the Hopf formula takes a specific form (see 216). Here $[F, F]$ is the commutator subgroup of F; and $[F, R]$ is the (normal) subgroup of F generated by all commutators $f^{-1}r^{-1}fr, f \in F, r \in R$.

This relation (216) is of considerable intrinsic algebraic interest, apart from the topological significance that it attaches to the group $(R \cap [F, F])/[F, R]$; for it shows that this group actually depends only on π, although it is expressed in terms of an arbitrary free presentation of π. This feature is characteristic of the procedures of homological algebra.

In 1944 Hopf gave his definitive solution of the problem raised by Hurewicz' result. Because World War II had virtually destroyed communication between European and American mathematicians, further developments of Hopf's theory, by B. Eckmann and H. Freudenthal in Europe and by Eilenberg and MacLane in the U.S., proceeded quite independently.

HOMOLOGY GROUPS OF GROUPS

Hopf's description of the homology groups of $K = K(\pi, 1)$ was based on a geometrical idea. If L is the universal cover of K, then π acts freely as a permutation group on the simplexes of L lying above a given simplex of K. It follows that, for each n, the chain group $C_n(L)$, with integer coefficients, is a free π-module. Moreover, the boundary homomorphisms are maps of π-modules. Further, because L is contractible, the homology groups of L are given by a set of equations (see 217) in which Z is the group of integers. Thus it may be said that the chain complex $C(L)$ is π-free and acyclic. If $C(L)_\pi$ arises by killing the π-action, then $C(L)_\pi = C(K)$, so that the homology of K, with any coefficients, may be computed from the chain complex $C(L)_\pi$. Hopf observed that if P is any π-free acyclic chain complex, then P_π is chain equivalent to $C(L)_\pi$ and so could be used to calculate the homology of K. He described such a chain complex P purely in terms of the group π itself.

Eilenberg–MacLane space

A particular chain complex $C(\pi)$ will now be described. Let $C_n(\pi)$ be the free Abelian group on $(n+1)$-tuples $(x_0, x_1, \cdots, x_n)$ of elements of π. The Abelian group $C_n(\pi)$ can be turned into a π-module by defining $(x_0, x_1, \cdots x_n)x = (x_0x, x_1x, \cdots, x_nx)$, and $C_n(\pi)$ is then plainly a free π-module. A differential $\partial\colon C_n(\pi) \to C_{n-1}(\pi)$ can be defined by the usual simplicial boundary formula (see 218), in which $x_i{}^*$ indicates that x_i is to be omitted. Obviously ∂ is a module map and $\partial\partial = 0$. An augmentation $\epsilon\colon C_0(\pi) \to Z$ is defined by $\epsilon(x) = 1$ and thus is obtained a chain complex of π-modules (see 219).

It can be shown that this chain complex (219) is acyclic by defining a contracting homotopy $s\colon C_{n-1}(\pi) \to C_n(\pi)$, which is a homomorphism of Abelian groups, by a certain rule (see 220).

Thus $C(\pi)$ may be used to replace $C(L)$, so that the homology groups of K are given (see 221) by an equation in which the group on the right is defined entirely in terms of π.

A little more generally, a local system of coefficients on K may be considered; that is, a π-module A of coefficients. Then $C(L) \otimes_\pi A$, the tensor product over π of the (right) π-complex $C(L)$ and the (left) π-module A, is a chain complex yielding $H_*(K; A)$, the homology of K with local coefficients A; and $C(L) \otimes_\pi A$ is chain equivalent to $C(\pi) \otimes_\pi A$, whence a relation follows (see 222). It is natural to write $H_n(\pi; A)$ for the group on the right of this relation (222), and $H_n(\pi; A)$ is the nth homology group of π with coefficients in the π-module A.

(213) $\pi_k X = 0, \qquad 2 \leqslant k \leqslant n-1$

(214) The dependence of $\dfrac{H_nX}{\theta_n\pi_nX}$ on $\pi = \pi_1X$

(215) $\pi_1[K(\pi, 1)] = \pi, \qquad \pi_n[K(\pi, 1)] = 0, \qquad n \geqslant 2$

(216) $H_2\pi \cong (R \cap [F, F])/[F, R]$

(217) $H_0L = Z, \qquad H_nL = 0, \qquad n \geqslant 1$

(218) $\partial(x_0, x_1, \cdots, x_n) = \sum_{i=0}^{n} (-1)^i(x_0, x_1, \cdots, x_i^*, \cdots, x_n)$

(219) $\cdots \to C_n(\pi) \xrightarrow{\partial} C_{n-1}(\pi) \to \cdots \xrightarrow{\partial} C_0(\pi) \xrightarrow{\epsilon} Z$

(220) $s(x_0, x_1, \cdots, x_{n-1}) = (e, x_0, x_1, \cdots, x_{n-1})$

(221) $H_nK = H_n(C(\pi)_\pi)$

(222) $H_n(K; A) \cong H_n(C(\pi) \otimes_\pi A)$

(223) $H^n(K; B) \cong H^n(\mathrm{Hom}_\pi(C(\pi), B))$

(224) $F(\phi + \psi) = F\phi + F\psi$

(225) $\begin{cases} P\colon \cdots \to P_n \xrightarrow{\partial} P_{n-1} \xrightarrow{\partial} \cdots \xrightarrow{\partial} P_0 \\ H_0(P) = M, \qquad H_n(P) = 0, \qquad n \geqslant 1 \end{cases}$

(226) $FP\colon \cdots \to FP_n \to FP_{n-1} \to \cdots \to FP_0$

Similarly, as observed by Eckmann and Eilenberg–MacLane, the cochain complex $\mathrm{Hom}_\pi(C(\pi), B)$ may be formed and thus, in cohomology, a similar relation holds (see 223) in which it is natural to write $H^n(\pi; B)$ for the group on the right of (223), so that definitions have been stated for the homology and cohomology groups of the group π (with coefficients in an arbitrary π-module).

It will now be shown how the definitions that have been stated above constitute special cases of the notion of a derived functor, which is basic to homological algebra.

DERIVED FUNCTORS

Let Λ be a fixed unitary ring and consider the category of right Λ-modules, for definiteness. Let F be a functor from this category to the category of Abelian groups that is additive in the sense that, if ϕ, ψ are Λ-module homomorphisms from the module M_1 to the module M_2, then an equality involving F holds (see 224). For a given module M, let there be given a chain complex of Λ-modules such that each is Λ-free (see 225).

Applying F leads to a chain complex of Abelian groups (see 226), and it may be shown that the homology groups of FP depend only on M and are independent of the choice of P [subject to the conditions described above and in (225)]. P is called a Λ-free resolution of M and a new functor L_nF can be defined because of this independence (see 227). It turns out that L_nF is again an

additive functor, called the nth left derived functor of F. For the re-expression of $H_n(FP)$ to be valid, it suffices that each P_n be Λ-projective. This means that if $\eta\colon R \to S$ is a surjection of Λ-modules and $\phi\colon P_n \to S$ is a Λ-homomorphism, there is a Λ-homomorphism $\psi\colon P_n \to R$ that lifts ϕ; *i.e.*, such that $\eta\psi = \phi$. A Λ-free module is certainly Λ-projective. If each P_n is Λ-projective, P is spoken of as a Λ-projective resolution of M.

The special case of the homology of groups is obtained by taking for Λ the integer group ring, $Z(\pi)$, of π, and for M the additive group of integers regarded as a trivial π-module. Then the functor F, given by the tensor product (see 228) in which A is a π-module, is an additive functor to the category of Abelian groups. The complex $C(\pi)$ is then a π-free resolution of Z, the additive group of integers, so that a relation holds (see 229) in which F is given above (see 228).

There are also right derived functors, obtained by taking Λ-injective resolutions, in which injective is dual, in the categorical sense, to projective; and it is possible to consider contravariant functors. The domain and codomain categories of the additive functor F can also be generalized; the essential requirement is that they be Abelian categories. The cohomology groups of the group π are obtained by taking, for a π-module B, the contravariant functor (see 230); GM is the group of π-homomorphisms from M to B; then, taking right derived functors of G, a relation is obtained (see 231).

The homology and cohomology groups of groups thus arise by taking left derived functors of the tensor product functor and right derived functors of the Hom functor. While there is an extensive general theory of derived functors, the theory of derived functors of these particular functors, but for arbitrary unitary rings Λ, is of particular importance. Thus, generalizing the definitions above (see 228 and 230), two sets of derived functors are defined (see 232 and 234) based on certain functors F, G (see 233 and 235), respectively. These are called the Tor and Ext functors, respectively; they are really bifunctors —that is, functors of two variables—and they enjoy the property of being balanced between their two arguments M, N in a sense that two equations express (see 236). Comparing past results (see 229, 232 and see 231, 234), it can be seen that two further relations hold (see 237, 238).

Tor and Ext functors

HOMOLOGY OF LIE ALGEBRAS

Many other homology theories have been found useful in algebra and arise by taking suitable derived functors. Here one will be described in which the derived functors are again Tor and Ext functors.

If g is a Lie algebra over the field K, Ug denotes its universal enveloping algebra. By a g-module, it is understood that A is a K-vector space on which the algebra Ug acts. Definitions are then made (see 239, 240), and these are called the homology groups of g (with coefficients in A) and the cohomology groups of g (with coefficients in B). Just as the (co)homology groups of a group may be interpreted as the (co)homology groups of a suitable space, so may the (co)homology groups of g be interpreted as the (co)homology groups of the underlying topological space of the compact, connected Lie group G the associated Lie algebra of which is g.

APPLICATIONS TO ALGEBRA

The topological significance of the homology theories of groups and Lie algebras has already been explained; these and other homology theories, however, also have important applications in pure algebra. For example, in the homology theory of groups, the group $H^1(\pi; A)$ is isomorphic to the quotient group of the group of derivations from π to A, modulo the inner derivations. The group $H^2(\pi; A)$ is isomorphic to the group $E(\pi; A)$ of equivalence classes of Abelian extensions of the quotient group π by the π-module A; such an extension can be written (see 241) to show that A is embedded in E and E projects onto π with kernel A. Here the group structure in the set $E(\pi; A)$ is by no means obvious, whereas the group structure in $H^2(\pi; A)$ follows immediately as in any homology or cohomology group. Similar interpretations of H^1 and H^2 are available in the homology theory of Lie algebras.

Uses of homological algebra

Many theorems of pure algebra may be proved by homological methods. For example, the homology theory of groups may be used to prove I. Schur's theorem that, if the quotient of a group G by its centre is finite, then the commutator subgroup of G is finite. It may also be used to prove Maschke's theorem that if the order of the finite group G is not divisible by the characteristic of the field K, then the (finite-dimensional) K-representations of G are completely reducible. Similarly, the homology theory of Lie algebras may be used to prove theorems that make no mention of homology; for example, the Levi-Malcev theorem that every (finite-dimensional) Lie algebra g, over a field K, is the split extension of a semisimple Lie algebra by the radical of g. Hilbert's celebrated chain-of-syzygies theorem may also be proved by homological methods—a notable early triumph for homological algebra; and there is the Auslander–Buchsbaum proof (the first given) that a regular local ring is a unique factorization domain. The claim that homological algebra is a useful tool in mathematics is not invalidated by the fact that these theorems can also be proved without the use of homological methods. For the homological methods are extremely systematic and serve to throw into relief the connections between apparently dissimilar areas of mathematics.

OTHER ASPECTS OF HOMOLOGICAL ALGEBRA

The scope of homological algebra is extremely broad. Only two homology theories (groups, Lie algebras) have been described from among many. There are also, however, aspects of homological algebra in which homological ideas are further refined and developed or in which homological techniques lead to new algebraic constructs. Two examples are now described.

Attention may be given to the category of Abelian groups that may well be graded by the integers, or even bigraded by pairs of integers. If A is such an Abelian group it is said to be differential if it is equipped with an

(227) $H_n(FP) = (L_nF)(M)$

(228) $FM = M \otimes_\pi A$

(229) $H_n(\pi; A) = (L_nF)(Z)$

(230) $GM = \mathrm{Hom}_\pi(M, B)$

(231) $H^n(\pi; B) = (R_nG)(Z)$

(232) $\mathrm{Tor}_n^\Lambda(M, N) = (L_nF)(M)$

(233) $FM = M \otimes_\Lambda N$

(234) $\mathrm{Ext}^n_\Lambda(M, N) = (R_nG)(M)$

(235) $GM = \mathrm{Hom}_\Lambda(M, N)$

(236) $\mathrm{Tor}_n^\Lambda(M, N) = (L_n\overline{F})(N), \quad \overline{F}N = M \otimes {}_\Lambda N$
$\mathrm{Ext}^n(M, N) = (R_n\overline{G})(N), \quad \overline{G}N = \mathrm{Hom}_\Lambda(M, N)$

(237) $H_n(\pi; A) = \mathrm{Tor}_n^{Z(\pi)}(Z, A)$

(238) $H^n(\pi; B) = \mathrm{Ext}^n_{Z(\pi)}(Z, B)$

(239) $H_n(g; A) = \mathrm{Tor}_n^{Ug}(K, A), \qquad n \geqslant 0$

(240) $H^n(g; B) = \mathrm{Ext}^n_{Ug}(K, B), \qquad n \geqslant 0$

(241) $A \to E \twoheadrightarrow \pi$

endomorphism d: $A \to A$, called the boundary, or differential, such that $d^2 = 0$. Notice that a chain complex of Abelian groups may be thought of as a differential graded Abelian group, in which d is the boundary homomorphism and so of degree -1; whenever A is (bi)graded a degree is supposed for d. If A is differential, the group $ZA = \ker d$ may be defined, and ZA is called the cycle group of A; one writes $BA = \operatorname{im} d$, and calls BA the boundary group of A. Because $d^2 = 0$, it follows that $BA \subset ZA$; HA is written for ZA/BA and HA is called the homology group of A. This is, of course, familiar from topology. A spectral sequence of differential Abelian groups is a sequence $\{E_n, d_n\}$, $n \geq 0$, in which E_n is an Abelian group with differential d_n and $HE_n = E_{n+1}$.

Associated with every spectral sequence there is a limit term, which may be described as follows. If $z \in ZE_n$, then its homology class $\{z\}$ belongs to E_{n+1} and the boundary d_{n+1} may be applied to $\{z\}$. It is convenient to talk of applying d_{n+1} to z itself in this case. Thus one may talk of successively applying the differentials $d_n, d_{n+1}, \cdots$ to an element c of E_n, each differential becoming applicable when its predecessor produces the value 0. In particular E^*_n may be defined as the subgroup of E_n consisting of those elements of E_n that are cycles for all d_r, $r \geq n$. Now the passage from cycle to homology class plainly induces a surjection of E^*_n onto E^*_{n+1} and E_∞, the limit of the spectral sequence, is defined to be the direct limit of the resulting sequence (see 242). Thus E_∞ is the quotient of E^*_0 by those cycles that eventually become boundaries; *i.e.*, that go to zero in some E^*_n. It is also said that the spectral sequence converges to E_∞.

Often in applications of spectral sequence theory the terms E_n are graded (or bigraded), and an early term of the spectral sequence is known and the limit term E_∞ is related to an Abelian group to be calculated, or, at least, about which information is sought. The relation of E_∞ to this group is frequently that of the graded group associated to a filtration of that group. A filtration (strictly, an increasing filtration) of the Abelian group G is a sequence of subgroups (see 243) and the associated graded group $\mathrm{Gr}(G)$ is the graded group, the pth component of which is given (see 244).

It is plain that if the filtration (see 243) is to be such that $\mathrm{Gr}(G)$ yields useful information about G, it must have certain properties. Usually it is asked that $F_pG = 0$ for $p \leq$ some p_0, and that $\cup_p F_pG = G$. Then every nonzero element of G appears as a nonzero element in precisely one component of $\mathrm{Gr}(G)$. From the given knowledge of E_2, say, conclusions can be drawn about E_∞ by passing through the spectral sequence and then inferences made about the filtered group G of which E_∞ is the associated graded group; sometimes this procedure is reversed.

The concept of a spectral sequence was invented by the French mathematician Jean Leray. Topologists first became generally aware of its importance, however, when the mathematician Serre applied it to the study of the relations between the homology groups of fibre, total space, and base space in a fibration, in 1951. He was thus able to obtain deep relations between the homology groups and homotopy groups of a space. Precisely, Serre showed the following (his result is not stated in its fullest generality): If F is the fibre, E the total space, B the base space, all spaces being connected, B simply-connected, and if G is an Abelian coefficient group, then there is a spectral sequence of bigraded Abelian groups $\{E_n, d_n\}$ with specific properties (see 245) and E_∞ is the graded group associated with $H_*(E; G)$, suitably filtered. Moreover, the filtration has the good properties mentioned earlier; and the convergence to the limit is finite in the sense that, given any (p, q), there exists n such that a sequence of equations holds (see 246). The differential d_n on E_n has degree $(-n, n-1)$. Thus knowledge of the homology groups of B and F leads to conclusions about the homology groups of E; indeed, knowledge of the homology groups of any two of F, E, B leads to information about the third.

Serre's theorem

Serre's theorem revolutionized homotopy theory; but spectral sequences then began to appear all over algebraic topology and homological algebra. For example, in the homology theory of groups, let a short exact sequence of groups be given (see 247); G is a group with normal subgroup N, such that $G/N \cong Q$. Let A be a G-module. There is then a Lyndon-Hochschild-Serre spectral sequence such that an equation holds (see 248), which converges to the graded group associated with $H_*(G; A)$, suitably filtered. There is a similar sequence in the homology theory of Lie algebras. Indeed, all the sequences described are special cases of a procedure due to the mathematician A. Grothendieck. Additive functors F: $\mathfrak{A} \to \mathfrak{B}$, G: $\mathfrak{B} \to \mathfrak{C}$ of suitable Abelian categories (for example, categories of modules) may be considered. Suppose that, for every projective object P of $\mathfrak{A}$, FP is (left) G-acyclic, that is to say, the left derived functors of G, evaluated on FP, give a set of equations (see 249). Then, for each A in $\mathfrak{A}$, there is a spectral sequence such that a relation holds (see 250) and it converges to the graded group associated with $L_*(GF)(A)$, suitably filtered. Even this formulation by no means exhausts the scope of useful generalization.

$$E_0^* \to E_1^* \to \cdots \to E_n^* \to E_{n+1}^* \to \cdots \tag{242}$$

$$\cdots \subseteq F_{p-1}G \subseteq F_pG \subseteq \cdots \subseteq G \tag{243}$$

$$Gr(G)_p = \frac{F_pG}{F_{p-1}G} \tag{244}$$

$$E_2^{pq} = H_p(B; H_q(F; G)) \tag{245}$$

$$E_n^{pq} = E_{n+1}^{pq} = \cdots = E_\infty^{pq} \tag{246}$$

$$N \to G \twoheadrightarrow Q \tag{247}$$

$$E_2^{pq} = H_p(Q; H_q(N; A)) \tag{248}$$

$$L_0G(FP) = GFP, \qquad L_nG(FP) = 0, \qquad n \geqslant 1 \tag{249}$$

$$E_2^{pq} = (L_pG)(L_qF)(A) \tag{250}$$

$$\mu^*: H^*(M; Q) \to H^*(M \times M; Q) = H^*(M; Q) \otimes_Q H^*(M; Q) \tag{251}$$

$$\psi: A \to A \otimes_Q A \tag{252}$$

$$(1 \otimes \eta)\psi = (\eta \otimes 1)\psi = 1: A \to A \tag{253}$$

$$(1 \otimes \theta)\psi = \psi(1 \otimes \theta) = \eta \tag{254}$$

Hopf algebras

Turning now to another topic of homological algebra, namely, that of Hopf algebras, again the source lies in algebraic topology. Suppose there is given a connected Lie group M; for example, the rotation group of Euclidean n-space. The cohomology ring $H^*(M; Q)$ of the underlying space of M with coefficients in the ring Q of rational numbers may then be formed; this is a graded commutative Q-algebra. Moreover, the multiplication map μ: $M \times M \to M$ induces an algebra map (see 251). Further, the fact that there is a two-sided unity for μ shows that if μ^* is followed by either projection onto $H^*(M; Q)$, the composite is the identity. This leads to a study of graded Q-algebras $A = \bigoplus_{n \geq 0} A_n$ admitting a diagonal map ψ, which is a map of Q-algebras (see 252), and a co-unit η: $A \to Q$ such that a consequence follows (see 253). (Note that $A \otimes_Q Q \cong A \cong Q \otimes_Q A$ so that $1 \otimes \eta$ and $\eta \otimes 1$ correspond to the projections mentioned above.) This is now called a graded Hopf algebra. Of course, the example $A = H^*(M; Q)$ has other properties that may qualify the Hopf algebra. It is commutative; it is connected in the sense that $A_0 = Q$ and η is simply the obvious projection onto A_0; it is co-associative in the sense that ψ is associative; and it admits an involution θ: $A \to A$ such that equations result (see 254).

In a pioneer work in 1941, Hopf proved that, in the terminology introduced above, if A is a commutative graded connected Hopf algebra over Q, then it is, as an algebra, an exterior algebra on odd-dimensional generators; moreover, if it is co-associative, then these generators are primitive, meaning that $\psi(x) = x \otimes 1 + 1 \otimes x$. Hopf thus explained a phenomenon relating to the four main series of simple Lie groups that had been observed by geometers, algebraists, and topologists in the 1930s. Because Hopf's theorem only required that M admit a continuous multiplication with two-sided unity, it encompassed a broader class of topological space than the Lie groups. Just how broad is now beginning to be glimpsed, as more and more Hopf manifolds are being discovered.

The study of Hopf algebras is now a part of algebra; moreover, there is a theory of differential Hopf algebras that is securely placed in the domain of homological algebra. Hopf's is seen as the first of many structure theorems that have proved of great value in both algebra and topology. A particularly interesting connection between the theory of Hopf algebras and what has gone before is provided by the group ring, KG, of a (finite group) G, over the field K. The diagonal map $\Delta : G \to G \times G$, yields a diagonal map ψ: $KG \to KG \otimes KG$ and so makes of KG a co-associative, co-commutative Hopf algebra over K with involution. (P.J.H.)

VII. Universal algebra

Historical background

Algebra began as the manipulation of numbers, using the four operations of arithmetic: addition, subtraction, multiplication and division. As the subject developed, the operations have changed very little, if at all, but the notion of number has been progressively enlarged. It was already recognized by the school of Pythagoras that rational numbers do not suffice to describe the diagonal of a unit square (*i.e.*, $\sqrt{2}$). In order to take account of such irrational numbers as $\sqrt{2}$, a geometric theory of proportion was developed by Eudoxus of Cnidus, which forms the substance of Book V of Euclid's *Elements*.

In the 16th century, when it was found that square roots of negative numbers may be needed to solve cubic equations, even when all their roots are real numbers, complex numbers were introduced. For a long time these numbers were known as impossible or imaginary, and their use was put on a firm basis only at the end of the 18th century, by Gauss.

The 19th century saw a series of discoveries, resulting in the creation of modern function theory. Some extensions of the complex number system were also made. Although these did not rival the complex numbers in importance, they formed a beginning of the more abstract outlook adopted today.

The first of these generalized number systems was Hamilton's quaternions. They were intended to describe points in space just as complex numbers describe points in the plane, and in fact Hamilton had mainly geometric applications in mind. From a somewhat different point of view, Grassmann, in his calculus of extension, treated geometric entities as objects of an algebraic system with addition and multiplication. The ideas of Hamilton and Grassmann gave rise, at the hands of Heaviside and Gibbs, to the modern development of vector analysis, not of concern here, except in so far as both the quaternions and Grassmann algebra may be regarded as an algebraic system with vectors instead of numbers, or a linear associative algebra. A system of quite a different sort, but still formally fitting into the framework of linear associative algebras, was described by Boole, who defined certain operations on statements: If two statements A, B are combined by conjunction (*i.e.*, the new statement "A and B" is formed), disjunction (forming "A or B") and negation (forming "not A"), a system satisfying many of the rules of the ordinary number system is obtained.

Another branch of algebra that began to take shape in the early days of the 19th century was the theory of groups, first as groups of permutations by Abel and Galois and later more generally as groups of transformations by Lie and Klein. The mere existence of all these different types of algebraic systems suggested a comparative study. Such a study was begun by the English mathematician Alfred North Whitehead, who in his 'Universal Algebra' I, described the systems of Boole, Hamilton, and Grassmann, as a preparation to a comparative study, which was to form volume II, but was never written. The name 'Universal Algebra,' according to Whitehead, was taken from a paper by the British mathematician James Joseph Sylvester: "Lectures on the Principles of Universal Algebra," published in the *American Journal of Mathematics*, vol. 6, 1884.

Development of axiomatic foundations

One of the characteristics of 20th-century mathematics has been the development of axiomatic foundations. Instead of referring new developments back to some part of the subject regarded as known, one would base everything on a small set of axioms. The method was used especially in algebra; the whole of group theory could be built up on three or four simple axioms. It also helped to crystallize such general notions as linear associative algebra and ring. Emmy Noether showed how the notion of group with operators allowed a number of different notions, such as group, vector space, representation module, to be treated by the same method. From here it was only a small step to drop the group structure. Universal algebra is nowadays understood to be the study of such general algebraic systems, by methods uniformly applicable to all of them.

BASIC NOTIONS OF UNIVERSAL ALGEBRA

An operation on a set A is a mapping or function defined on A with values in A. For example, if A is mankind, then "mother of x" is an operation in one argument, or a unary operation (here the fact that the mapping is undefined for Adam and Eve is ignored). On the set of integers, the operation of taking the sum of the numbers x and y, written $x + y$, is a binary operation, because it depends on two arguments: x and y. In the plane, the operation of taking the centroid of a triangle ABC is a ternary operation, depending on three arguments, A, B, C. Most operations depend on a finite number of arguments; that is, they are said to be finitary. If an operation depends on n arguments, in which n may be 1, or 2, or any positive integer, then the operation is said to be n-ary, and the number n is called its arity. Occasionally, 0-ary operations are used; such an operation depends on no arguments at all; it simply picks out a particular element of the set. For this reason it is also called a constant operator. For example, in a group the neutral element may be regarded as the value of a 0-ary operation.

Operations that are n-ary

An algebraic structure on a set A consists in specifying a number of operations on A; these operations are not restricted in any way except that the arity of each is usually a preassigned integer. In this way A becomes an algebraic system, or simply an algebra. For example, groups may be regarded as algebras with a binary operation, multiplication: xy; a unary operation, the inverse: x^{-1}; and a 0-ary operation, the neutral: e, together with certain laws (such as the associative law). As an example of an algebra with infinitely many operations, a real vector space is considered. This has the group operations (because it is an additive group) as well as a unary operation for each real number λ: scalar multiplication by λ. To facilitate the comparison of different algebras, a set Ω is usually chosen as operator set; with each member ω of Ω an integer $n(\omega)$ is associated, to indicate the arity of the resulting operation, and an Ω-algebra structure is obtained by prescribing for each ω in Ω an operation $\omega(x_1, \cdots, x_n)$ of arity $n = n(\omega)$.

The notion of a subgroup of a group or a subspace of a vector space is generalized by defining a subalgebra of an algebra A as a subset admitting all the operations of A. The result is an algebraic structure of the same kind as A, just as every subgroup of a group is a group in its own right. To give an example, the complex numbers form a real vector space, with the real numbers as subspace. But the complex numbers may also be regarded as a complex vector space, and now the real numbers do not form a subspace because they do not allow multiplication by

$\sqrt{-1}$. This illustrates the need to specify the operations as well as the sets in describing algebras.

If in an algebra A a subset X is taken, it will in general not be a subalgebra; *i.e.*, by applying the algebra operations to elements of X new elements are generally obtained that do not lie in X. $<X>$ denotes the set of all elements obtainable from X by repeated application of the algebra operations. When any of these operations is applied to elements of $<X>$, again an element of $<X>$ is obtained; in other words, $<X>$ admits all the algebra operations and so is a subalgebra of A. It is actually the least subalgebra containing X and is usually called the subalgebra generated by X.

Quotient groups

The notion of a quotient group can be generalized as follows: If an Ω-algebra A is partitioned into classes, it is possible to define an operation ω of A on the set of classes, provided that varying the arguments of ω within a given class does not change the class in which the result lies; that is, if x_1, x'_1 lie in the same class and likewise x_2, x'_2, etc., then $\omega(x_1, \cdots, x_n)$ and $\omega(x'_1, \cdots, x'_n)$ must lie in the same class. When this condition holds for all ω in Ω, the given partition is called a congruence, and denoting it by q say, A/q is written for the set of classes. The above condition ensures that each ω in Ω can be defined unambiguously as an operation on A/q. In this way A/q is turned into an Ω-algebra, called the quotient algebra of A by q. For example, if $\mathbf{Z}$ is the ring of integers and $\mathbf{Z}$ is partitioned into classes by putting two numbers in the same class precisely when their difference is divisible by 7, then there are just 7 classes, described by their least remainders after division by 7. If two numbers a, a' leave the same remainder and b, b' do likewise, then so do their respective sums $a + b$ and $a' + b'$; similarly for the other ring operations (subtraction and multiplication), and therefore a congruence is obtained and the residue classes form a quotient ring.

A mapping f from one Ω-algebra, A, to another, B, is called a homomorphism and is written as $f: A \to B$ if it preserves the Ω-algebra structure in the sense that the result of applying f to the value of an operation ω is the same as applying f to the arguments and then taking ω;

$$f(\omega(x_1, \cdots, x_n)) = \omega(f(x_1), \cdots, f(x_n)).$$

For example, if q is a congruence on an algebra A, then there is a natural mapping ϕ that associates with each element of A the class of q in which it lies; now the condition for q to be a congruence ensures that ϕ is a homomorphism. A homomorphism from A to B that is a one-one correspondence (bijection) between these algebras is called an isomorphism between A and B, and A is then said to be isomorphic to B.

Given two Ω-algebras A and B, a new Ω-algebra $A \times B$ can be defined, called their direct product, by taking the set of all pairs with the first component a member of A and the second a member of B, and performing the operations of Ω componentwise. More generally, such a direct product can be formed from any collection (even infinite) of Ω-algebras.

These notions of subalgebra, quotient algebra, homomorphism, and direct product arose by generalizing the familiar situation in group theory. There the notion of a quotient group looks simpler than in the general case because the classes of a congruence on a group are just the cosets of a subgroup. For other algebraic systems having no underlying group structure, such as lattices or semigroups, the situation is more involved; nevertheless the notion of congruence has been useful there too, in analysing homomorphisms. Thus with each homomorphism $f: A \to B$ between Ω-algebras, there can be associated a quotient algebra A/q of A and a subalgebra B' of B, together with an isomorphism $f_1: A/q \to B'$ such that the original homomorphism f is obtained by taking in turn the natural mapping $A \to A/q$, the isomorphism f_1, and the inclusion mapping $B' \to B$, which leaves each element of B' fixed but considers it as an element of B. The congruence q of A that arises in this way is called the kernel of the mapping f, and it has the following interpretation: two elements of A are mapped to the same element of B by f precisely when they belong to the same class of the kernel of f. The set of values of f, namely B', is also called the image of A under the mapping f, and the above analysis of homomorphisms shows that the images of an algebra under homomorphisms are precisely its quotient algebras. For this reason the quotient algebras of an algebra are also called its homomorphic images.

Now the usual isomorphism theorems of group theory, which allow a comparison of subgroups of quotients, and of quotients of quotients, to be made, can be taken over with hardly any change. One can also prove, under suitable hypotheses, an analogue of the Jordan-Hölder-Schreier theorem, which compares different chains of factor algebras; *i.e.*, quotients of subalgebras. There is a close connection with lattices here: as already observed, the results can be applied to lattices, as particular algebraic systems. At the same time the subalgebras of an algebra form a lattice, as do the congruences on an algebra, and in this way the results of lattice theory can be used in a general analysis of algebraic systems.

THE CONNECTION WITH MATHEMATICAL LOGIC

The results described in the last section are quite general in scope, and apply even to systems with infinitary operations; that is, operations depending on infinitely many arguments. There is, however, another group of results for which finitarity of the operations is essential; these results are often regarded as characteristic of algebraic systems. The basic property is the following algebraicity condition: if A is any algebra and X a subset of A, then an element of A lies in $<X>$, the subalgebra generated by X, precisely when it lies in the subalgebra generated by some finite subset of X. This is so because any member of $<X>$ is obtained by applying the algebra operations repeatedly to elements of X, and, because all operations are finitary, the total number of arguments needed is also finite. This principle forms the basis for the application of Zorn's lemma (an algebraic formulation of the axiom of choice in set theory). It applies not only to algebraic systems—that is, sets with operations—but also to relational systems; that is, sets with operations and relations defined on them, such as ordered groups. It also has a remarkable connection with logic: any mathematical theory may be regarded as an algebraic system the elements of which are the theorems (the provable statements in the theory) and the operations of which are the usual rules of deduction in logic (modus ponens and the rule of substitution); therefore it satisfies the algebraicity condition mentioned earlier. Now a basic theorem of logic due to the logician Gödel, the completeness theorem, states that any theory (composed of statements of the lower predicate calculus) admitting the rules of deduction and not including a contradictory statement has a model. On the other hand one can show, by purely algebraic means, that the theories that have models satisfy the algebraicity condition. The practical consequence is often stated as the compactness theorem: A set of statements within a theory (of the lower predicate calculus) is consistent whenever every finite subset of statements is consistent.

The compactness theorem

A typical application of this result to the algebraic theory of fields states that a sentence holds in every field of characteristic zero, provided that it holds in every field of sufficiently high characteristic. This follows because fields of characteristic greater than a given number n can be defined by the conjunction of the sentences E_2, E_3, $\cdots$, E_n, in which E_2 is: "for each $x \neq 0$, $x + x \neq 0$"; *i.e.*, no nonzero element when added to itself gives zero (to exclude fields of characteristic 2), E_3 is: "for each $x \neq 0$, $x + x + x \neq 0$," and so on, while fields of characteristic zero are defined by the conjunction of all the sentences E_2, E_3, $\cdots$. If a sentence holds for all fields of characteristic exceeding some fixed number N, then it must be consistent with any finite subset of the collection E_2, E_3, $\cdots$, and hence (by the compactness theorem) with all of them; *i.e.*, it holds in characteristic zero.

Besides being a field of application of universal algebra, mathematical logic is also an important tool. Thus to study a particular class of algebraic systems, it is not

enough to specify the operations. It is also necessary to lay down which algebras are to be admitted. This is usually done by means of sentences formed from such basic statements as $x + y = z$, in accordance with the rules prescribed by logic; for example, if the system to be specified has a multiplication that is to have an inverse, a sentence is required of the form: For all x, y, there exists u such that $xu = y$. The sentences needed to define a class of algebras can be classified according to logical principles, and conclusions can then be drawn about the class of algebras defined. Thus, if the sentences defining the class of algebras contain only universal quantifiers ("For all . . .") and no existential quantifiers ("There exist . . ."), in other words, if they are universal sentences, then the class of algebras so defined is closed under subalgebras: with an algebra, all its subalgebras belong to the class. This is fairly easy to prove; more difficult (and more interesting) is a result in the opposite direction: A class of algebras that is closed under subalgebras and that is definable by axioms can be defined by universal sentences. The condition of being definable by axioms can also be expressed in terms of explicit conditions on the algebras of the class, but is a little more complicated to state and so is omitted here.

Classes of algebraic systems

As an example, the class of all groups is considered. This can be defined by the following laws: (a) the associative law: $x(yz) = (xy)z$; (b) the law expressing that e is neutral: $xe = ex = x$; and (c) the law expressing that x^{-1} is inverse to x: $xx^{-1} = x^{-1}x = e$. These laws are understood to hold for all group elements x, y, z, and so they are universal sentences, and in fact any subalgebra, within the general sense of the word (that is, any subgroup), of a group is again a group. The same is true if a sentence such as "for all x, $x^3 = e$" is added. This picks out the class of groups in which the cube of any element is the neutral element, and any subgroup of such a group has the same property. Next consider the class of periodic groups; *i.e.*, groups in which every element has some power equal to e: $x^n = e$, in which the value of the exponent n depends generally on x. Any subgroup of a periodic group is periodic, but the class of periodic groups cannot be defined by axioms. If one attempts this, one finds oneself having to say that for each x, either $x^2 = e$ or $x^3 = e$ or $x^4 = e$ or $\cdots$, a sentence of infinite length that is not admitted. It might seem as if periodic groups could be defined by the finite sentence: For every x there exists a positive integer n such that $x^n = e$. This is not, however, allowed by the rules of the game, which permit only sentences of the lower predicate calculus, because a quantifier has been applied to an integer (. . . there exists a positive integer n . . .). Next examine the class of all groups that are not periodic; it is easy to find a group that is not periodic but that has a periodic subgroup. So this class cannot be expected to be defined by universal sentences. Finally, the class of aperiodic groups is considered; that is, groups in which no element other than e has any positive power equal to e. This class is closed under subgroups and can in fact be defined by universal sentences, expressing the fact that no power of an element other than e can be *e:* For all x, if $x \neq e$, then $x^2 \neq e$; for all x, if $x \neq e$, then $x^3 \neq e \cdots$. These sentences form the multiplicative analogue of the sentences E_2, E_3, $\cdots$ used earlier to define fields of characteristic zero.

VARIETIES OF ALGEBRAS

Many important classes of algebras can be defined entirely by identities; that is, equations that hold for all values of the arguments. Thus groups were defined by the identities expressing the associative law and the properties of neutral and inverse elements. Such a class is called a variety. Groups form a variety, likewise Abelian groups, but not all finite groups, because any law satisfied by all finite groups is also satisfied by all infinite groups. Testing for varieties is facilitated by the following criterion due to Garrett Birkhoff: A class of algebras is a variety if, and only if, it is closed under taking subalgebras, homomorphic images, and direct products. Using this test it is easy to check that finite groups do not form a variety: a direct product of finite groups, with infinitely many factors, is no longer finite. In the same way one finds that periodic groups do not form a variety; nor do the aperiodic groups; the groups satisfying $x^3 = e$, however, do form a variety.

Testing for varieties

Varieties have several properties that make them more amenable to treatment. A useful property, actually characteristic of varieties, is the following: for each integer n, there is an algebra F_n the quotients of which constitute precisely the algebras in the variety that can be generated by n elements; a corresponding algebra exists when n is replaced by an infinite cardinal number. This algebra F_n is called the free algebra of the variety, on n free generators; it is rather like the algebra of polynomials. In elementary algebra a polynomial in n variables $x_1, \cdots, x_n$ with real coefficients has the property that any substitution of real numbers for the variables $x_1, \cdots, x_n$ gives a real number. Moreover, these polynomials can be added, subtracted and multiplied; *i.e.*, they form a ring, the polynomial ring in the variables $x_1, \cdots, x_n$ over the real numbers, and it is not hard to see that any substitution of real numbers for the x's preserves the ring operations; *i.e.*, it is a homomorphism. In the same way group elements can be formed in n variables $x_1, \cdots, x_n$, using only the general group laws to simplify the expressions. Thus a variable that is followed or preceded by its formal inverse can always be omitted: $xyy^{-1}z = xz$, $ux^{-1}xu = u^2$, but no other simplification is possible. For example, one may not replace xy by yx because the commutative law $xy = yx$ does not hold in all groups. In this way the free group on $x_1, \cdots x_n$ is obtained; its elements are all products in $x_1, \cdots, x_n$ and their inverses; that is, all group words in these variables. By the associative law all parentheses can be omitted, and two group words represent the same group element if it is possible to pass from one to the other by inserting or omitting occurrences of a variable together with its inverse. Given any group G, any elements of G can be substituted for the variables in words of the free group. Thus x_1 can be replaced by a_1, x_2 by a_2, etc., taking care to replace x_1^{-1} by a_1^{-1}, etc. In this way each word corresponds to an element of G, and words representing the same element of the free group will certainly correspond to the same element of G. Therefore, a mapping is obtained from the free group into G, and this mapping can easily be shown to be a homomorphism. By taking enough variables (infinitely many if necessary) any group can be expressed in this way as a homomorphic image of a free group. A similar construction applies in any variety of algebras, such as the variety of Abelian groups, or the variety of lattices, or of semigroups with a neutral element.

Algebraic systems often arise in a concrete way; for example, groups may arise as groups of transformations. There is another important method of defining algebras within a given variety, however, which consists in writing down a set of generators and defining relations. This means that a set of symbols is written as the generating set, and a set of relations is written that are to hold between these symbols, such that all other relations can be deduced from them. Such an algebra then consists of all expressions formed from these symbols by the algebraic operations, in which two expressions are to represent the same element whenever it is possible to pass from one to the other using only the given relations and the defining laws of the variety. To give a simple example, the group of complex cube roots of unity has the following presentation: All expressions obtained by applying group operations to a single group variable x are taken; *i.e.*, all group words in x, namely e, x, x^2, x^3, $\cdots$, x^{-1}, x^{-2}, $\cdots$. Two of them are declared equal if and only if it is possible to pass from one to the other using the relation expressing that the cube of x is the neutral element: $x^3 = e$. This is abbreviated by writing the group as $gp\{x \mid x^3 = e\}$, called a presentation of the group. In this particular case it is easy to see the structure of the group: two powers of the generator x^m and x^n, say, represent the same element of the group if, and only if, the difference of their exponents $m - n$ is divisible by 3; so there are just three elements, represented by the words e, x, x^2. Of course,

Defining algebras by means of generators and relations

few groups have a presentation as simple as this one, and for more complicated presentations it may be quite difficult to decide when two words represent the same group element. This problem is known as the word problem for groups; a given presentation of a group is said to have a soluble word problem if there exists a method enabling one to decide for each pair of words in a finite number of steps whether or not they represent the same element. The word problem has been solved for some simple classes of groups, but more interestingly, it has been shown to be insoluble in general. Group representations have been found that have an insoluble word problem. Similarly, presentations can be obtained for the algebras in any variety, and it will be necessary to solve the word problem. For many algebras encountered in practice, a presentation can be found with soluble word problem, but generally there will also be algebras with insoluble word problem in all but the simplest varieties.

The word problem

The notion of a free algebra is useful in many ways. If it is required to establish a property of groups that is preserved by homomorphisms, the first step is to check whether the free group has it. If so, then the desired conclusion follows because every group is a homomorphic image of a free group; if the free group does not have the property, it is a counter-example. This method is particularly effective for groups (as well as for semigroups and rings) because the word problem for free groups (and semigroups and rings) is rather easy to solve. This means that the elements of free groups can be represented in a very perspicuous form. It is no longer true of such varieties as alternative algebras or Jordan algebras, and as a consequence some apparently quite simple problems—such as the existence of zero-divisors in free Jordan algebras (nonzero elements the product of which is zero)—are still open.

Free algebras

Free algebras (in different varieties) have been the subject of intensive investigation, partly as an end in itself and partly on account of applications. Thus in commutative ring theory, the free algebras are polynomial rings, and the theorem that every commutative ring is the homomorphic image of a free commutative ring is the algebraic counterpart of the geometric fact that every geometric figure lies in n-dimensional space (for a suitable integer n). This accounts for the importance of polynomial rings in the study of algebraic geometry.

A free semigroup in the variables $x_1, \cdots, x_n$ may be realized as the set of all finite sequences of letters from $x_1, \cdots, x_n$, allowing repetitions; that is, all words in the alphabet $x_1, \cdots, x_n$. The study of such words has led to applications in coding theory as well as in the algebraic theory of languages, which has provided much of the impetus for studying free semigroups. As an example, the free semigroup on x, y, z and a set of words is sought such that no spaces or commas are needed to divide any message into its constituent words (this is known as a comma-free code). The set of words x, xy, z, yz clearly will not do, for the expression xyz can be parsed in two ways: as $x \cdot yz$ or $xy \cdot z$. But there is an infinite set of words with the required property: $x, xy, xy^2, xy^3, \cdots$. From the algebraic point of view a comma-free code is a set that forms a free generating set for a free subsemigroup, and such sets have been characterized by simple criteria. Another application is to elementary arithmetic: the positive integers may be regarded as the free semigroup on one free generator. Now the basic property, that every mapping of the single free generator defines a unique homomorphism, is just another way of expressing the principle of induction.

Besides applications in algebra itself, free algebras have many uses in algebraic geometry and topology, automata theory, and the statistical theory of queues.

UNIVERSAL ALGEBRA AND CATEGORY THEORY

There are close links between universal algebra and category theory. Although the precise degree of useful interaction is still open to discussion, it is widely agreed that category theory provides a convenient language for stating many of the results of universal algebra. Often there are several ways of expressing the same phenomenon; thus the characteristic property of free algebras is a particular instance of a universal construction, which appears in many different contexts. It may also be described as the left adjoint of the forgetful functor (which passes from an algebra to its underlying set and so forgets the algebra structure).

The definition of an Ω-algebra involved prescribing, for each member of Ω, an operation of the right arity. If n elements of the algebra are taken and m n-ary operations are performed on them, m elements are obtained on which an m-ary operation can be performed. In this way an n-ary operation is obtained by composition. There are also the identity operations, which pick out the first argument, or the second, or any other of the n arguments. Any collection of operations on a set, closed under the composition described above and containing all the identity operations (one unary, two binary ones, three ternary ones, etc.) is called a clone, and in studying a particular Ω-algebra the operations corresponding to members of Ω may be replaced by the clone they generate. Now it is possible to define clones abstractly, and to define algebraic systems one may take a particular abstract clone C and define C-algebras as sets on which the clone C is realized by operations. This definition is due to P. Hall, and is very close to the categorical definition, due to Lawvere. The latter defines a theory as a category $\mathfrak{T}$ the objects of which are nonnegative integers, in which the number n is regarded as the coproduct of n copies of the number 1. Any contravariant functor from $\mathfrak{T}$ to the category of sets and mappings, in which coproducts map to products, is called a $\mathfrak{T}$-algebra. If A is the set corresponding to the number 1, then n corresponds to the nth direct power A^n, and the various morphisms in $\mathfrak{T}$ from 1 to n correspond to mappings from A^n to A; *i.e.*, n-ary operations on A. Any morphism in $\mathfrak{T}$ from m to n corresponds to m morphisms from 1 to n; this means that the mappings from A^n to A^m each correspond to m n-ary operations on A. In other words, the $\mathfrak{T}$-algebra structure so defined amounts to picking out a clone of operations on A.

Although many of the results of universal algebra can be proved within the framework of category theory, they can be proved just as easily directly. The most significant progress in universal algebra has been in the areas adjacent to logic, and there category theory has been of little help. On the other hand, in applications to other parts of mathematics (especially in algebraic geometry) category theory has to some extent replaced universal algebra. (P.M.C.)

BIBLIOGRAPHY. JOHN E. WHITESITT, *Boolean Algebra and Its Applications* (1961), gives a very elementary introduction to Boolean algebra, with some applications to logic and switching circuits; GARRETT BIRKHOFF and THOMAS C. BARTREE, *Modern Applied Algebra* (1970), gives a more comprehensive and more recent discussion from a similar standpoint. JAMES C. ABBOTT, *Sets, Lattices and Boolean Algebras* (1969), is a very readable introductory text on sets and lattices from a purely mathematical standpoint; *Trends in Lattice Theory*, ed. by Abbott (1970), discusses informally some recent trends. GARRETT BIRKHOFF, *Lattice Theory*, 3rd ed. (1967), is the standard treatise on the subject. PAUL R. HALMOS, *Algebraic Logic* (1962), discusses in depth some of the Boolean algebras arising in logic; JOHN B. ROSSER, *Simplified Independence Proofs* (1969), shows how these have been recently applied to prove fundamental theorems in set theory and logic.

CAMILLE JORDAN, *Traité des substitutions et des équations algébriques* (1870); and WILLIAM BURNSIDE, *Theory of Groups of Finite Order*, 2nd ed. (1911, reprinted 1955), are the great classical works on group theory. For a period the subject was relatively dormant, but in the 1930s a revival began signaled by PHILIP HALL's celebrated paper "A Contribution to the Theory of Groups of Prime Power Orders," *Proc. Lond. Math. Soc.*, Second Series, 36:29–95 (1932). In 1923 there appeared the first edition of ANDREAS SPEISER, *Die Theorie der Gruppen von endlicher Ordnung*, 4th ed. (1956); HANS ZASSENHAUS, *Lehrbuch der Gruppen theorie* (1937; Eng. trans., *The Theory of Groups*, 2nd ed., 1958); B.L. VAN DER WAERDEN, *Gruppen von linearen Transformationen* (1935); and ALEXANDER G. KUROSCH, *The Theory of Groups*, 2 vol. (1955–56; orig. pub. in Russian, 1953), which developed the theory of free groups and free products ex-

tensively. IRVING KAPLANSKY, *Infinite Abelian Groups*, rev. ed. (1969), treats a theory that is now a major subject in its own right. The importance of group theory in physics is recognized by VOLKER HEINE, *Group Theory in Quantum Mechanics* (1960); and G.Y. LYUBARSKII, *The Application of Group Theory in Physics* (1960; orig. pub. in Russian, 1957). A turning point was the new approach in the paper by PHILIP HALL and GRAHAM HIGMAN, "On the p-Length of p-Soluble Groups and Reduction Theorems for Burnside's Problem," *Proc. Lond. Math Soc.*, Third Series, 6:1–42 (1956). This more modern approach and the revival of representation theory, omitted by Kurosch and Zassenhaus, is shown in MARSHALL HALL, JR., *The Theory of Groups* (1959); and WILLIAM R. SCOTT, *Group Theory* (1964). GILBERT DE BEAUREGARD ROBINSON, *Representation Theory of the Symmetric Groups* (1961), deals with a special but important topic. WILLIAM MAGNUS, ABRAHAM KARRASS, and DONALD SOLITAR, *Combinatorial Group Theory* (1966), deals mostly with generators and relations. The book-length paper by WALTER FEIT and JOHN THOMPSON, *Solvability of Groups of Odd Order* (1962), brought activity to its present high pitch; and the development and extension of this theory is the subject of the advanced treatise by DANIEL GORENSTEIN, *Finite Groups* (1968). BERTRAM HUPPERT, *Endliche Gruppen I* (1967), is the first volume of an encyclopaedic reference work.

Many works on algebra contain material on fields. Of these, GARRETT BIRKHOFF and SAUNDERS MACLANE, *A Survey of Modern Algebra*, 3rd ed. (1965); SOLOMON FEFERMAN, *The Number Systems: Foundations of Algebra and Analysis* (1964); I.N. HERSTEIN, *Topics in Algebra* (1964); and CYRUS C. MACDUFFEE, *An Introduction to Abstract Algebra*, ch. 1–6, (1940), are four standard undergraduate works.

Most treatises on algebra have chapters on rings—*e.g.*, BIRKHOFF-MACLANE (*op. cit.*); and SERGE LANG, *Algebra*, ch. 2–6 and 17 (1965). A deeper study of commutative rings may be found in OSCAR ZARISKI and PIERRE SAMUEL, *Commutative Algebra*, 2 vol. (1958–60); MASAYOSHI NAGATA, *Local Rings* (1962); IRVING KAPLANSKY, *Commutative Rings* (1970); and MICHAEL F. ATIYAH and I.G. MACDONALD, *Introduction to Commutative Algebra* (1969). For noncommutative rings, see NATHAN JACOBSON, *The Theory of Rings* (1943, reprinted 1966); and I.N. HERSTEIN, *Noncommutative Rings* (1968). Applications of ring theory to arithmetic may be found in PIERRE SAMUEL, *Théorie algébriques des nombres* (1967; Eng. trans., *Algebraic Theory of Numbers*, 1970).

Standard texts on the theory of categories include BARRY MITCHELL, *Theory of Categories* (1965); and BODO PAREIGIS, *Categories and Functors* (1970). ION BUCUR and ARISTIDE DELEANU, *Introduction to the Theory of Categories and Functors* (1968), is an excellent account of the theory of categories and functors with special reference to its applications in algebraic topology. SAUNDERS MACLANE, *Categories for the Working Mathematician* (1971), is a recent work in which the author, a distinguished categorist, sets out his view of the material of category theory basic to the professional mathematician. PETER J. FREYD, *Abelian Categories* (1964), is devoted to a special and very important class of categories. The first mention of the notions of category and functor in their explicit form is contained in the paper by SAMUEL EILENBERG and SAUNDERS MACLANE, "General Theory of Natural Equivalences," *Trans. Am. Math. Soc.*, 58:231–294 (1945). The classic text in homological algebra is HENRI CARTAN and SAMUEL EILENBERG, *Homological Algebra* (1956). More recent texts are SAUNDERS MACLANE, *Homology* (1963); and PETER J. HILTON and U. STAMMBACH, *A Course in Homological Algebra* (1971)—an abbreviated version of this work may be found in Hilton's *Lectures in Homological Algebra* (1971). A natural context for homological algebra is described in PETER J. FREYD, *Abelian Categories* (1964). The role played by homological algebra in algebraic topology may be inferred from PETER J. HILTON and SHAUN WYLIE, *Homology Theory* (1960); and EDWIN H. SPANIER, *Algebraic Topology* (1966). Particular aspects of homological algebra are to be found in JOHN W. MILNOR and JOHN C. MOORE, "On the Structure of Hopf Algebras," *Ann. Math.*, 81:211–264 (1965); and EDWIN WEISS, *Cohomology of Groups* (1969). The background for the study of homological algebra in the theory of Lie groups and Lie algebras is contained in GERHARD P. HOCHSCHILD, *The Structure of Lie Groups* (1965); and NATHAN JACOBSON, *Lie Algebras* (1962). The pioneer work of Hopf that led to the emergence of homological algebra as a separate discipline is contained in the papers of HEINZ HOPF: "Fundamentalgruppe und zweite Bettische Gruppe," *Comment. math. helvet.*, 14:257–313 (1941–42); and "Über die Bettischen Gruppen die zu einer beliebigen Gruppe gehören," *ibid.*, 17:39–79 (1944–45). His work on Lie groups that led to a Hopf algebra is described in "Über die Topologie der Gruppen-Mannigfaltigkeiten und ihrer Verallgemeinerungen," *Ann. Math.*, 42:22–52 (1941).

Some works on universal algebra are: ALFRED NORTH WHITEHEAD, *A Treatise on Universal Algebra, with Applications*, vol. 1 (1898, reprinted 1960), a description of the algebras of Boole, Hamilton, and Grassmann, with emphasis on geometrical applications; ABRAHAM ROBINSON, *Introduction to Model Theory and the Metamathematics of Algebra* (1963), a treatment of model theory with many applications to algebra; PAUL M. COHN, *Universal Algebra* (1965), a general introduction to the subject, stressing the connections with logic and giving applications to algebra; GEORGE GRATZER, *Universal Algebra* (1968), a comprehensive treatment of the subject, including much current research in the field; J.F.C. KINGMAN, "On the Algebra of Queues," *Methuen's Review Series in Applied Probability*, vol. 6, pp. 1–44 (1966), a description of the algebraic background to the problem of queues (waiting lines) with a single server; GIAN-CARLO ROTA, "Baxter Algebras and Combinatorial Identities I, II," *Bull. Am. Math Soc.*, 75:325–334 (1969), in which an operator identity from probability theory is proved by first solving the word problem for a class of algebras; and GARRETT BIRKHOFF, (*op. cit.*), which has a long chapter on universal algebra.

(G.Bi./Ma.H./P.S./P.J.H./P.M.C.)

Algeria

Algeria (in French, Algérie) is an independent republic in the Maghrib, the Arabic term for Northwest Africa. With an area of 896,588 square miles (2,332,164 square kilometres), it is the second largest country in Africa and the tenth largest country in the world; 780,617 square miles (2,021,800 square kilometres) of the national territory, however, consist of two Saharan *wilāyāt* (*départements* or provinces) in the south. In the 1966 census, Algeria's population was numbered at about 11,833,000; later estimates counted (1970) 13,547,000 and (1971) 14,643,700. The country is bounded to the east by Tunisia and Libya; to the south by Niger, Mali, and Mauritania; to the west by Morocco and Spanish Sahara; and to the north by the Mediterranean. Under the Évian Agreements—which were concluded with France in 1962 and which led to independence for Algeria—France, which formerly ruled in Algeria, has retained certain economic concessions in the Saharan *wilāyāt* in the south.

Algeria is a country of vastness; the Mediterranean to the north separates it from Europe but at the same time opens a thousand doors to it, while in the south the Sahara links it to Africa. History, language, customs, and the Muslim religion make it an integral part of the Arab world. Algeria's struggle for liberation, which lasted from 1954 to 1962, has gained it a significant place within the Third World (composed of those countries that are uncommitted to either the Western or Communist groups of states), while the internal as well as external choices that its leaders have made place it among the ranks of socialist nations. The problems that confront the nation, among them economic development, literacy programs, health, housing, and cultural rebirth, are to be dealt with in the framework of some basic concepts:

1. Algeria is a nation with a socialist and anti-imperialist vocation.
2. Algeria seeks to build a modern state and has at its disposal the organization and the appropriate national institutions to do so.
3. A socialist economy is to be built in which private enterprise and planning coexist.
4. While agriculture, in the main, will continue to be the dominant activity, Algeria's nascent industry—above all the petroleum industry, from prospecting to refining—is to continue to expand.
5. Algeria's citizens are to be encouraged to recover their past traditions and to control their own future; this is to be done by the provision of national education and by emphasizing such themes as public health, justice, social work, and recreation. (For associated physical features, see ATLAS MOUNTAINS; MEDITERRANEAN SEA; and SAHARA; for the capital city, see ALGIERS; for Algerian history, see NORTH AFRICA, HISTORY OF.)

THE LANDSCAPE

Relief. Algeria's inhabitants are unequally distributed throughout her territory; almost the entire population lives in the northern part of the country—the Tell—

which is the richer and the better watered. The remainder of the country, the Saharan region, is uninhabited desert, with the exception of several "islands" of verdure and of life, the oases. The Saharan region conceals rich mineral resources and, above all, sources of energy. Each of the two physical regions into which the country is divided results from a different geological evolution. The main features of structure and relief are produced by the collision of the rocks now forming the geologically young and crumpled mountain chains in the north with the solid and ancient platform of rock, horizontal and uniform, that forms the south. The first of these areas is the mountainous Atlas region, consisting of two large mountain chains, running roughly parallel from east to west and separated by high plateaus. The second region consists of the Sahara, which also extends into the adjacent countries to the south.

The northern region. In the northern part of the country there are, from north to south, five different zones, all running from east to west. The first of these is formed by the folded littoral (coastal zone); the second by the plains which lie to the south of the littoral; the third zone consists of the first chain of mountains, the Atlas Tellien (Tell Atlas); the fourth zone consists of the Haut Plateaux (High Plateaus); and the fifth consists of the second chain of mountains—the Atlas Saharien (Saharan Atlas)—which extends in some areas of broken terrain into the Sahara itself.

On the coast the folds and ridges usually terminate in high cliffs, which are indented with numerous bays; these indentations are spaced out along the coast and the principal ports are located in them. The coastal massifs (mountainous masses of rock) are often separated from each other by plains—such as the plains of Oran and of Annaba—which extend from the littoral into the region behind it. In the same way, the Atlas Tellien is not continuous but forms two distinct ranges to the west, separated by interior plains. Thus the plains of Maghria and of Ghossels separate the Monts de Tlemcen to the south from the Monts des Traras to the northwest, and the Monts du Tessala to the northeast. The plains of Sidi bel Abbès and of Mascara are interpolated between the Monts des Beni-Chougran to the north and the mountains of Daïa and of Saïda to the south. The Dahra Massif forms a long range, extending from the mouth of the Chelif River in the west to the Djebel Chenoua in the east; it is separated from the Massif de l'Ouarsenis to the south by the plains of the Chelif Valley.

The relief as a whole, therefore, does not constitute a barrier to communication between the different areas that form the western Tell region. This is not so, however, in the central Tell, where the previous regular pattern permitting communication progressively diminishes and disappears. The Atlas de Blida is not easily distinguished from the Monts du Titeri, while the mountainous and compact ensemble of the Grande Kabylie joins the Biban and the Hodna mountains. Only the valley of the Oued (Wadi) Soummam permits communication with the port of Bejaïa on the coast.

Mountain barriers

More to the east, from Bejaïa to Edough, one mountain barrier follows another, surrounding the Plains of Constantine, which are dominated to the south by the Hodna, Aurès, and Nementcha mountains. The Plains of Constantine themselves, which have a distinct local topography, and since ancient antiquity have been used for growing cereals, do not present the same features as the Algero-Oranian Haut Plateaux. These extend from the Monts du Hodna into Morocco; they are broken by sebkhas (lake beds encrusted with salt) and are much less favourable to agriculture because of the scarcity of rain.

To the south of the Algero-Oranian Haut Plateaux and the Plains of Constantine runs the second mountain chain, the Atlas Saharien, which is formed of a series of small ranges that run generally southwest to northeast. They decline in altitude from the west (where Djebel Aïssa attains a height of 7,336 feet) to the east. Their structure is such that interregional communications from north to south are facilitated. More to the east, between the Plains of Constantine to the north and the eastern Atlas Saharien to the south, rises the Massif de l'Aurès, whose summits reach altitudes higher than 6,000 feet.

The southern region. The Algerian Sahara may be roughly divided into two depressions of different altitude, separated from one another by a central north-to-south rise called the Mzab (M'Zab). Each zone is covered by a vast sheet of sand dunes, called an erg. The Grand Erg Oriental (Great Eastern Erg), which has an average altitude of from 1,300 to 2,000 feet, and the Grand Erg Occidental (Great Western Erg) are made of drifts of sand, declining in height from the foot of the Ahaggar (Hoggar) Mountains in the south to the foot of the Massif de l'Aurès, below sea level, in the north. A notable feature is the Ahaggar Massif, which rises to majestic summits, the highest more than 9,500 feet in altitude.

Climate, vegetation, and animal life. Climate, more than relief, is the determining geographical factor. The amount of rainfall, and above all its distribution throughout the agricultural year, as well as the nature of certain winds, constitute the principal elements upon which pastoral activities depend.

Regional variations in rainfall

The amount of rainfall increases as one proceeds from west to east but diminishes from the north to the south. The greatest amount of rainfall is experienced in the mountainous regions of the eastern littoral, which are directly exposed to the humid winds that blow inland from the Mediterranean. From the meridian that runs through the port of Cherchell, about 50 miles west of Algiers, to the Tunisian frontier, the rainfall exceeds 23 inches, and in certain places—for example in the Grande Kabylie, Petite Kabylie, and Edough regions—reaches about 40 inches. To the west of the Cherchell meridian, however, a considerable part of the Chelif Plains, and the plains of the littoral and the region immediately to the south of it in the vicinity of Oran, are insufficiently watered, receiving less than 23 inches. Only the Atlas Tellien receives from 16 to 23 inches. As soon as one crosses the Atlas ranges to the south, however, rainfall rapidly diminishes, except in the Massif de l'Aurès and in a section of the Djebel Amour, which still receives about 16 inches.

This boundary roughly divides the two principal agricultural zones. In the first, dry farming is generally possible and commercially profitable. Fine forests and an abundant vegetation are also found. In the second zone the cultivation of cereal crops is only possible under exceptional circumstances. Instead, pastoral activities are predominant, and the forests disappear.

There is a regular annual rainy season from December to March; at intervals during the rest of the year heavy storms occur, accompanied by rain. Rainfall is often torrential and may be spread out over several days, whereas at other times the weather is dry and hot for months. The amount of rainfall may vary drastically, sometimes increasing by 200 percent from one year to another. Algiers, where the average is almost 30 inches, received about 16 inches in 1913 and 50 inches in 1847. Djelfa, which receives an average of about 12 inches, received less than 4 inches in 1913, and about 30 inches in 1847. This irregularity, which often is of considerable consequence for agriculture in the interior, is aggravated by temperature variations and, locally, by dry winds. If the annual temperature range in the coastal area is comparatively low, elsewhere it is high—111° F (44° C) at Miliana and 129° F (54° C) at el-Asnam. In inverse relationship to the proximity to the sea, the cold becomes sharper on the southern fringes of the Tell, with extreme minimum temperatures reaching 9° F (−13° C) at Berrouaghia and 12° F (−11° C) at Bouïra, while on the Haut Plateaux, extreme minimum temperatures reach 16° F (−9° C) at Sétif, 10° F (−12° C) at Batna, and 5° F (−15° C) at el-Bayadh in the Oran region.

Only the littoral, the Plains of Oran, the lower Chelif River Valley, and the Mitidja Plain behind Algiers are spared the frost. While in the Haut Plateaux the frost occurs 20 days a year, it may occur for 50 days in the Sétif and Bordj Bou Arreridj regions.

The coastal regions, which are subjected directly to the influence of the Mediterranean Sea, are well watered and

Cultivation on the coastal plains

experience relatively little temperature variation. Permanently under intensive cultivation, they are especially used for fruit growing and market gardening. The plains and the hills in the region immediately to the south of the littoral still receive sufficient rain but have a much drier atmosphere; the temperatures show more variation in range. The natural vegetation cover, preserved only on a few massifs, includes some fine trees, such as the cork oaks that grow around Constantine, or the thugas (trees resembling the cypress) that are found in the Oran region of the Atlas Tellien.

The Haut Plateaux, on the other hand, are characterized by great daily and annual extremes of temperature, hot summers and cold winters, and insufficient rainfall.

NORTHERN ALGERIA

MAP INDEX

Political subdivisions

Adrar 26·00n 1·00w
Alger 36·42n 3·08w
Annaba 36·54n 7·46e
Batna 35·15n 6·00e
Béchar 30·00n 3 30w
Bejaïa 36·30n 5·00e
Biskra 34·15n 5 40e
Blida 36·30n 2·35e
Bouira 36·15n 3·50e
Constantine 36·25n 6·42e
Djelfa 34·45n 3·00e
el-Asnam 36·10n 1·20e
Guelma 36·25n 7·45e
Jijel 36·45n 5·45e
Laghouat 32·00n 3 00e
Mascara 35·30n 0·10e
Médéa 36·12n 2·50e
Mostaganem 35·51n 0·07e
M'Sila 35·00n 3·50e
Oran 35·43n 0·43w
Ouargla 29·00n 8·30e
Oum el-Bouaghi 35·53n 7·07e
Saïda 33·00n 0·15e
Sétif 36·10n 5·05w
Sidi bel Abbes 35·00n 0·45w
Skikda 36·50n 7·00e
*Tamanrasset 25 00n 6·00e
Tébessa 35·20n 8·05e
Tíaret 35·28n 1·21e
Tizi-Ouzou 36·48n 4·02e
Tlemcen 34·52n 1·15w

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

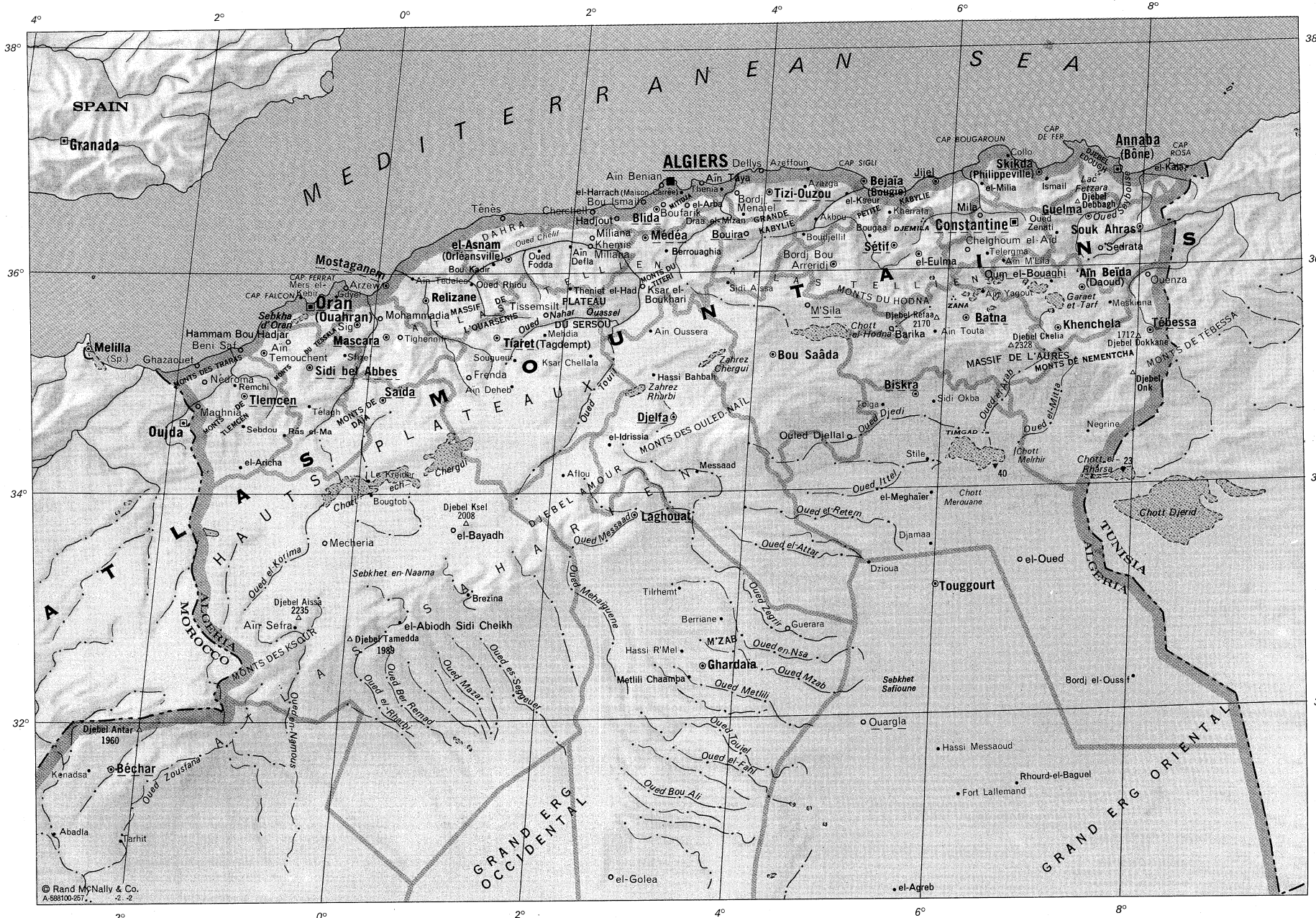
Cities and towns

Abadla 31·02n 2·45w
Aflou 34·01n 2·03e
'Aïn Beïda (Daoud) 35·50n 7·29e
Aïn Benian 36·45n 2·51e
Aïn Defla 36·16n 1·58e
Aïn Deheb 34·51n 1·31e
Aïn M'lila 36·02n 6·30e
Aïn Oussera 35·24n 2·54e
Aïn Sefra 32·47n 0·37w
Aïn Taya 36·50n 3·10e
Aïn Tedeles 36·00n 0·21e
Aïn Temouchent 35·16n 1·04w
Aïn Touta 35·26n 5·59e
Aïn Yagout 35·50n 6·28e
Akbou 36·27n 4·31e
Algiers (Alger) 36·42n 3·08e
Annaba (Bône) 36·54n 7·46e
Arzew 35·50n 0·23w
Azazga 36·48n 4·30e
Azeffoun 36·51n 4·25e
Barika 35·23n 5·17e
Batna 35·34n 6·15e
Béchar 31·37n 2·13w
Bejaïa (Bougie) 36·45n 5·05e
Beni Saf 35·17n 1·15w
Berriane 32·49n 3·44e
Berrouaghia 36·07n 2·52e
Biskra 34·50n 5·52e
Blida 36·30n 2·49e
Bône, see Annaba
Bordj Bou Arreridj 36·02n 4·49e
Bordj el-Oussif 32·20n 8·02e
Bordj Menaïel 36·44n 3·45e
Boudjellil 36·19n 4·22e
Boufarik 36·37n 2·52e
Bougaa 36·21n 5·05e
Bougie, see Bejaïa
Bougtob 34·03n 0·06e
Bouira 36·23n 3·54e
Bou Ismail 36·40n 2·44e
Bou Kadir 36·04n 1·08e
Bou Saâda 35·11n 4·09e
Brezina 33·04n 1·14e
Chelghoum el-Aïd 36·10n 6·08e
Cherchell 36·35n 2·16e
Collo 36·58n 6·37e
Colomb-Béchar, see Béchar
Constantine 36·25n 6·43e
Daoud, see 'Aïn Beïda
Dellys 36·50n 3·55e
Djamaa 33·32n 5·59e
Djelfa 34·30n 3·20e
Draa el-Mizan 36·35n 3·47e
Dzioua 33·14n 5·14e
el-Agreb 30·48n 5·30e
el-Abiodh Sidi Cheikh 32·56n 0·42e
el-Arba 36·37n 3·13e
el-Aricha 34·09n 1·10w
el-Asnam (Orléansville) 36·10n 1·20e
el-Bayadh 33·40n 1·01e
el-Eulma 36·08n 5·40e
el-Golea 30·30n 2·50e
el-Harrach (Maison Carrée) 36·45n 3·07e
el-Idrissia 34·30n 2·37e
el-Kala 36·50n 8·30e
el-Kseur 36·46n 4·49e
el-Meghaïer 33·55n 5·58e
el-Milia 36·48n 6·14e
el-Oued 33·20n 6·58e
Fort Lallemand 31·13n 6·17e
Frenda 35·02n 1·01e
Gdyel 35·48n 0·26w
Ghardaïa 32·31n 3·37e
Ghazaouet 35·06n 1·51w
Guelma 36·28n 7·26e
Guerara 32·46n 4·34e
Hadjout 36·31n 2·25e
Hammam Bou Hadjar 35·20n 1·05w
Hassi Bahbah 35·05n 2·55e
Hassi Messaoud 31·43n 5·59e
Hassi R'Mel 32·35n 3·24e
Igli 30·25n 2·12w
Ismail 36·50n 7·21e
Jijel 36·52n 5 50e
Kenadsa 31·30n 2·30w
Khemis Miliana 36·16n 2·13e
Khenchela 35·28n 7·11e
Kherrata 36·31n 5·26e
Ksar Chellala 35·13n 2·18e
Ksar el-Boukhari 35·51n 2·52e
Laghouat 33·50n 2·59e
Le Kreïder 34·06n 0·02e
Maghnia 34·50n 1·50w
Maison Carrée, see el-Harrach
Marnia, see Maghnia
Mascara 35·20n 0·09e
Mecheria 33·35n 0·18w
Médéa 36·12n 2·50e
Mehdia 35·26n 1·40e
Mers el-Kebir 35·45n 0·43w
Meskiana 35·39n 7·41e
Messaad 34·12n 3·25e
Metlili Chaamba 32·18n 3·40e
Mila 36·27n 6·16e
Miliana 36·15n 2·15e
Mohammadia 35·33n 0·03e
Mostaganem 35·51n 0·07e
M'Sila 35·46n 4·31e
Nédroma 35·01n 1·45w
Negrine 34·30n 7·30e
Nemours, see Ghazaouet
Oran (Ouahran) 35·43n 0·43w
Orléansville, see el-Asnam
Ouahran, see Oran
Ouargla 31·59n 5·25e
Oued Fodda 36·11n 1·32e
Oued Rhiou 35·58n 0·55e
Oued Zenati 36·18n 7·05e
Ouenza 35·57n 8·04e
Ouled Djellal 34·28n 5·02e
Oum el-Bouaghi 35·53n 7·07e
Philippeville, see Skikda
Rás el-Ma 34·31n 0·46w
Relizane 35·44n 0·30e
Remchi 35·04n 1·26w
Rhourd-el-Baguel 31·15n 6·55e
Saïda 34·50n 0·11e
Sainte-Denis-du-Sig, see Sig
Sebdou 34·35n 1·20w
Sedrata 36·08n 7·32e
Sétif 36·09n 5·26e
Sfizef 35·14n 0·15w
Sidi Aïssa 35·53n 3·48e
Sidi bel Abbes 35·13n 0·10w
Sidi Okba 34·48n 5·54e
Sig 35·32n 0·11w
Skikda (Philippeville) 36·50n 6·58e
Sougueur 35·12n 1·30e
Souk Ahras 36·23n 8·00e
Stile 34·20n 5·51e
Tagdempt, see Tíaret
Tarhit 30·55n 2·02w
Tébessa 35·28n 8·09e
Télagh 34·47n 0·34w
Telergma 36·07n 6·21e
Ténès 36·31n 1·14e
Thenia 36·43n 3·34e
Theniet el-Had 35·47n 2·01e
Tíaret (Tagdempt) 35·28n 1·21e
Tighennif 35·20n 0·21e
Tilrhemt 33·10n 3·21e
Tissemsilt 35·35n 1·50e
Tizi-Ouzou 36·48n 4·02e
Tlemcen 34·52n 1·15w
Tolga 34·46n 5·22e
Touggourt 33·10n 6·00e

Physical features and points of interest

Aïssa, Djebel, *mountain* 32·56n 0·38w
Amour, Djebel, *mountains* 34·00n 2·15e
Antar, Djebel, *mountain* 31·59n 1·57w
Arab, Oued el-, *watercourse* 34·41n 6·31e
Atlas Mountains 35·30n 3·00e
Atlas Saharien (Saharan Atlas), *mountains* 33·25n 1·20e
Atlas Tellien (Tell Atlas), *mountains* 35·00n 3·00e
Attar, Oued el-, *watercourse* 33·27n 5·26e
Aurès, Massif de l', *mountains* 35·08n 6·30e
Ber Remad, Oued, *watercourse* 31·45n 1·10e
Bou Ali, Oued, *watercourse* 31·14n 4·16e
Bougaroun, Cap, *cape* 37·10n 6·30e
Chélia, Djebel, *mountain* 35·19n 6·42e
Chelif, Oued, *river* 36·01n 0·07e
Chergui, Chott ech-, *lake* 34·15n 0·30e
Chergui, Zahrez, *lake* 35·12n 3·32e
Dahra, *mountains* 36·20n 1·21e
Daïa, Monts de, *mountains* 34·40n 0·14w
Debbagh, Djebel, *mountain* 36·33n 7·18e
Djedi, Oued, *watercourse* 34·28n 6·05e
Djemila, *ruins* 36·25n 5·44e
Dokkane, Djebel, *mountain* 35·23n 8·00e
Edough, Djebel, *mountains* 36·56n 7·41e
Fahl, Oued el-, *watercourse* 31·15n 4·41e
Falcon, Cap, *cape* 35·45n 1·00w
Fer, Cap de, *cape* 37·08n 7·10e
Ferrat, Cap, *cape* 35·55n 0·30w
Fetzara, Lac, *lake* 36·52n 7·22e
Grande Kabylie, *mountains* 36·34n 4·18e
Grand Erg Occidental, *desert* 30·30n 0·30e
Grand Erg Oriental, *desert* 30·30n 7·00e
Hauts Plateaux, *plateau* 34·20n 0·15e
Hodna, Chott el-, *lake* 35·30n 4·45e
Hodna, Monts du, *mountains* 35·50n 4·50e
Ittel, Oued, *watercourse* 34·19n 6·01e
Korima, Oued el-, *watercourse* 33·51n 0·23w
Ksel, Djebel, *mountain* 33·43n 1·10e
Ksour, Monts des, *mountains* 32·45n 0·30w
Mazar, Oued, *watercourse* 31·50n 1·36e
Mediterranean Sea 37·40n 3·00e
Mehaïguene, Oued, *watercourse* 32·15n 2·59e
Melrhir, Chott, *lake* 34·17n 6·24e
Merouane, Chott, *lake* 34·02n 6·10e
Messaad, Oued, *watercourse* 33·45n 2·51e
Metlili, Oued, *watercourse* 31·54n 4·53e
Mitidja, *plain* 36·39n 3·00e
Mitta, Oued el-, *watercourse* 34·20n 6·44e
M'zab, *physical region* 32·35n 3·55e
Mzab, Oued, *watercourse* 32·19n 5·24e
Naama, Sebkhet en-, *salt marsh* 33·20n 0·16w
Nahar Ouassel, Oued, *river* 35·42n 2·48e
Namous, Oued en-, *watercourse* 31·00n 0·15w
Nementcha, Monts de, *mountains* 34·59n 6·51e
Nsa, Oued en-, *watercourse* 32·28n 5·24e
Onk, Djebel, *mountain* 34·52n 7·57e
Oran, Sebkha d', *lake* 35·30n 1·00w
Ouarsenis, Massif de l', *mountains* 35·42n 1·32e
Ouled-Naïl, Monts des, *mountains* 34·33n 3·28e
Petite Kabylie, *mountains* 36·30n 5·29e
Refaa, Djebel, *mountain* 35·34n 5·53e
Retem, Oued er-, *watercourse* 33·30n 5·45e
Rharbi, Oued el-, *watercourse* 31·50n 0·51e
Rharbi, Zahrez, *lake* 34·50n 2·50e
Rharsa, Chott el-, *salt lake* 34·06n 7·50e
Rosa, Cap, *cape* 36·58n 8·15e
Safioune, Sebkhet, *salt marsh* 32·16n 5·27e
Saharan Atlas, see Atlas Saharien
Seggeuer, Oued es-, *watercourse* 31·39n 2·26e
Sersou, *plateau* 35·30n 2·00e
Seybouse, Oued, *river* 36·54n 7·47e
Sigli, Cap, *cape* 36·56n 4·45e
Tamedda, Djebel, *mountain* 32·47n 0·05e
Tarf, Garaet et-, *lake* 35·40n 7·10e
Tébessa, Monts de, *mountains* 35·20n 8·20e
Tell Atlas, see Atlas Tellien
Tessala, Monts du, *mountains* 35·08n 0·53w
Timgad, *ruins* 35·23n 6·30e
Titeri, Monts du, *mountains* 35·57n 3·17e
Tlemcen, Monts de, *mountains* 34·40n 1·25w
Touiel, Oued, *watercourse* 31·30n 4·46e
Touil, Oued, *watercourse* 35·30n 2·33e
Traras, Monts des, *mountains* 34·57n 1·46w
Zana, *ruins* 35·45n 6·05e
Zegrir, Oued, *watercourse* 32·43n 5·00e
Zousfana, Oued, *watercourse* 30·29n 2·17w

* See map of Algeria



NORTHERN ALGERIA

Size of symbol indicates relative size of town • ○ ◉ ▣ ■ ■

Elevations in metres

The cultivation of cereals is not always or everywhere possible. Natural vegetation is represented by the very characteristic association of wild olive and the mastic tree accompanied by gorse and dwarf palms. This brush type of vegetation has, however, almost everywhere disappeared as a result of land clearance; usually only the dwarf palms and such hardy plants as the asphodel remain.

Where less than 12 inches of rain are received, the vegetation of the steppe type (characterized by extensive treeless plains) occurs, although discontinuously; steppes of esparto grass or wormwood are the most widespread.

On the southern border of the Atlas Saharien, the Sahara proper begins. The demarcation coincides with a diminution of the rainfall to less than four inches a year. The landscape and the type of plant life differ greatly from those in the north. Life and activity are limited to a few privileged locations. The rains, more than in the north, are marked by their irregularity; Tamanrasset one year received about six inches, and another year only about a quarter of an inch, while Ouargla received about three inches one year and nothing the following year. Three years may pass without rainfall in the Tademaït region, and as much as five years in the Ahaggar.

These circumstances produce a very rare vegetation, which appears rapidly after rain and disappears almost immediately. Plant life is very dispersed and consists of tufts of several kinds of grass, various robust species that need almost no water such as drinn and cram-cram, several shrubs (which are always stunted and are sometimes spiny), ethels, acacia, jujube trees, as well as some more varied species which are found in the beds of oueds (watercourses which are dry except during periods of rain) with underground water, or in mountainous regions.

Such plant life can only sustain a scarce animal life. Few species are found—some antelope, gazelles, hare, jerboas, and wild boars, as well as a few birds and some reptiles and insects.

Orchards and cereal crops

The *sahels* (coastal plains) of Oran, Algiers, and Annaba specialize in early vegetables, while the plains of the interior permit a varied cultivation. The hillsides are usually covered with orchards, while the plateaus and the Haut Plateaux are sufficiently well watered to be used for the cultivation of cereals. Elsewhere, pastoralism is dominant, except for a few irrigated tracts.

In the Saharan regions, agriculture is confined to a few selected regions with underground water resources, such as the *souf* (intermittent stream). Despite the extreme hostility of nature, however, man occasionally succeeds in establishing a basis for life—for example in the Mzab region—by using the rare runoff water found after rain.

Elsewhere, in the Tell, much greater possibilities exist, though runoff water after showers is lost, and its infiltration into the soil remains partial or reduced. Even so, the natural conditions that prevail, although they deteriorate toward the south, present many advantages from which man has benefitted. Resources are generally varied and abundant, as has been shown by numerous surveys.

Traditional regions. The geographic distribution of the population shows an imbalance in the population density that corresponds, approximately, to differences in soil fertility and to variations in the amount of rainfall. Different types of life style virtually coincide with the principal vegetation zones. In the Tell, man predominates on such plains as are to be found and without irrigation has been able to use them for growing orchards and cereals. Where the slopes are too steep for cultivation, continuous enclosures for livestock are established. Human settlement has occurred throughout this region, and the population density is everywhere greater than 13 persons per square mile.

The cultivation of cereals is uneconomic in the steppe region of the south, except where irrigation has been well planned and where the soil is also favourable. A certain amount of local nomadism occurs, but this is declining as a way of life. The population density is between 1 and 13 persons per square mile.

Beyond this lies the Saharan *bled*, an area where true nomads are to be found and where groves of palm trees dot the landscape. The population density falls below one person per square mile. Inequality of population distribution is caused not only by the establishment of towns but also by density variations in purely rural areas, varying from 420 persons per square mile in the *wilāya* (*département;* province) of Tizi-Ouzou to 41 in the Tiaret *wilāya*, and to about 12 in the *wilāya* of Saïda. The town versus country division is nevertheless important. During Algeria's War of Independence, which ended in 1962, the French colonial authorities moved more than 2,000,000 Algerians to 1,800 centres in an attempt to remove them from the control of the Algerian nationalists' Army of National Liberation. After independence these centres were abandoned, and it is believed that about 500,000 Algerians returned to the towns. This sudden upheaval accelerated the pace of the continual exodus from the country to the towns that is occurring. The reasons for this exodus are principally to be found in the agrarian system, which does not allow for demographic growth. The movement is facilitated by improvements in transportation and further accelerated by natural disasters, such as droughts or floods. The consequences are many—sociopolitical, for example, with the Arabization of the Berber-speaking people in the large towns; and demographic, with the imbalance that results from the departure of single men from the countryside. Economic consequences also follow—changes in social and administrative organization, the destruction of pastureland by neglect, and a diminution of the *khamesat* (tenant farmer) system and of nomadism.

The exodus to the towns

PEOPLE AND POPULATION

Groups historically associated with the contemporary country. The peoples of Algeria represent diversity in unity—a diversity and a unity formed of many different cultural zones that are subtly interconnected. Among these are the Tell and the Sahara; the mountains, the plains, and the plateaus; the world of the nomad as opposed to that of the sedentary farmer—not to speak of the intermediate world of the seminomad. The diversity is also represented not only by the people themselves but by the kinds of houses that they live in—the terraced houses of the Aurès and the Mzab regions; the tiled roofs of the Kabylie dwellings; the cupolas on houses in the Souf, a cluster of oases towns northeast of Touggourt; the Moorish style of architecture in the towns. Still further marks of division are to be found between the closely grouped houses of those who have long been accustomed to a settled way of life and the dispersed habitations of those who have only recently adopted a sedentary life style. Other distinctions may be drawn between the original Berber population and the Arab "newcomers" from the east. Yet Algeria's long and checkered history has resulted in so thorough an intermingling of the population that it is extremely difficult to distinguish perfectly pure ethnic types. Although the Berbers constitute the main element in the population of the country and the Arabs the minority that has strongly influenced them, no clear ethnic frontier between the two groups can now be drawn. Berber-speakers as well as Arabic-speakers can, of course, be found; there are many cultural differences—represented for example by the antithesis between Berber law and Muslim law. Most of the Berbers, however, are Arabic-speaking, and other differences are usually obscured by a host of nuances and intermediate transitional zones. In art, however, the two cultural poles are represented by the opposition between the stark rectilinear style of Berber decoration and the more slender and supple lines of the Arab style.

The Arab and Berber cultures

Among the intermingling of cultures, certain relatively clear-cut cultural zones can nevertheless be distinguished. Thus, wherever the Berber language has been preserved—that is to say, essentially in the mountainous massifs of the Kabylie and the Aurès—not only particular cultural traits have been maintained but also an original life style. Among other things, a certain independence from Islām may be observed. In the Mzab region a similar independence is also manifest, above all in the juridical domain, in an attachment to the soil, in the

heavy labour used to make it productive, and in the egalitarian social structure. If it is true, on the one hand, that the Arab nomads have introduced their own system of values marked by a detachment so far as ownership of land is concerned, by an "aristocratic" social orientation, and by a pastoral way of life, it would nevertheless be unwise for either group to challenge the other. The Arab tribe is not separable from its territorial patrimony, which is, moreover, clearly defined to prevent encroachments from rivals; likewise, yet differently, the social structure of Berber society is patterned on genealogy. Between the two systems a constant dialogue is in progress, a dialogue that is founded upon a deep affinity and that is dominated by the temptation to seek identification with the other as well as by the will to affirm a distinct identity.

Seen as a whole, from its populated plains to its desert extremities, the country appears divided into such an apparently haphazard variety of groups that it might be imagined that it would disintegrate into individual units. To this tendency, however, is opposed the intense interpenetrating circulation that characterizes the whole. This takes such forms as the migrations of shepherds, the cycle of markets (which provides occasions for juridical and cultural exchanges), and the expansion of the urban areas. The Muslim faith also provides unity, and religious centres are the focusses where orthodoxy and Eastern civilization intermingle. Diversity of speech exists, but general reference is made to the same sacred language—Arabic. In effect the two antithetical aspects of Algerian life—unity and plurality, compartmentalization and continuity—can be understood only in relationship to each other. In Algeria, as in the Maghrib as a whole, there is no closed society, no group so isolated or so involved with its own affairs that it does not to some extent relate itself to other entities or to outside ways. All these cultural considerations, moreover, do not include the factors brought into play by the contacts between European civilization and Maghribian civilization or the upheavals experienced as a result of the colonial era.

Demography. At the census conducted in 1966 the population of Algeria was numbered at about 11,833,000. The demographic survey of 1971 gave a total of 14,643,700, and an estimate in 1974 gave 16,275,000. The population is characterized by its youth: in the mid-1970s about 58 percent of Algerians were less than 20 years old, 38 percent were between the ages of 20 and 64, and only the remaining 3 percent were over 65 years old. The birth rate, 4.8 percent, is one of the highest in the world; if it is maintained, the country will have a population of 24,000,000 by the late 1980s, an increase of more than 1,000,000 every three years.

A number of factors are thought to contribute to the continuing high birth rate. In addition to the very high proportion of young persons in the total population, there is a very high marriage rate among young women (60 percent of Algerian women marry before the age of 20); furthermore, some 74 percent of the population has not attended school, which causes difficulties for any program of family planning. Nevertheless, the problem of such a high birth rate is recognized by the central government, which is trying, though so far not very successfully, to reduce it.

Linguistically, speakers of Arabic comprise about 81.5 percent of the population and Berber-speakers 17.9 percent; there are a few native French-speakers and others. The Berbers are concentrated in the eastern portions of Algeria, particularly the more mountainous areas. The census of 1966 indicated that these were the portions of the country that displayed the strongest patterns of emigration between the censuses of 1954 and 1966. One inference drawn from these facts was the possibility of a substantial migration (perhaps 110,000 to 170,000 persons a year) within the country toward areas of greater economic opportunity, the cities and mining areas in particular. Socio-economic patterns after 1966 did little to diminish the volume or direction of such migration.

So far as juridical nationality is concerned, in the late 1960s more than 98 percent of the inhabitants of the country were Algerian, about 0.5 percent—about 68,000, of whom 32,000 lived in the city of Algiers—were French, and about 1 percent were foreigners of other nationalities. The *wilāya* of Alger is the most populous; the least populous *wilāyāt* are those of the southern desert areas. The urban and semi-urban population is provisionally estimated at (1970) 43 percent of the total. In 1966 Algiers, the capital, had a population of about 944,000; Oran had about 327,000 inhabitants and Constantine about 255,000.

Between World War II and April 1970, 630,000 Algerians emigrated to France; almost 100,000 Algerians settled in France between 1955 and 1970 alone.

The illiteracy rate is still high, but the government is making considerable efforts to reduce it. As a result of the Four-Year Plan, which covered the years from 1970 to 1973, it was hoped that about three-quarters of the population would become literate. (A.K.Ch./Ed.)

THE NATIONAL ECONOMY

Algeria is one of the most important oil-producing countries in the world and is the largest exporter of liquefied natural gas and the second largest exporter of bulk wine. It is the third most important producer of cork, after Spain and Portugal, and ranks fifth in date production. Until 1962 the economy was largely rural-based and complementary to that of France. Since independence, however, production of oil and natural gas has come into its own, and industrialization has proceeded rapidly. In the early 1970s, nevertheless, almost half the country's trade was still with France and another quarter with other Common Market countries, including West Germany, Italy, Belgium, The Netherlands, and Luxembourg. At the same time, Algeria was seeking to formalize its relations with the Common Market and to acquire associate status in it. During the mid-1970s the United States also became a major trading partner.

Since the mid-1960s there have been several abortive attempts to establish a Maghrib economic community with Algeria's neighbouring North African states, Tunisia and Morocco. If the plan ever becomes a reality, Algeria would certainly be the best placed to use the larger market to fortify its industrial and economic growth.

The extent and distribution of resources. *Mineral resources.* Extensive deposits of sulfur-free, light crude oil were discovered in the Algerian Sahara in the mid-1950s. Production began in 1958, concentrated in three main fields: Hassi Messaoud North, Hassi Messaoud South, and Zarzaitine-Edjeleh. These fields still accounted for over three-fifths of production in the late 1960s, even though several important fields had been discovered elsewhere. Established reserves in the mid-1970s were estimated at around 9,000,000,000 barrels, but more than 1,500,000 of an estimated total of 2,560,000 square miles of potential oil-bearing land remained to be explored.

Oil and natural gas

Deposits of natural gas were first discovered at Hassi R'Mel in 1956, and since then discoveries have also been made at several other fields. Proved reserves are put at more than 100,000,000,000,000 cubic feet. The gas has a methane content of almost 85 percent and also contains ethane, propane, and helium. Liquefaction plants at Arzew and Skikda are in operation.

The country has extensive deposits of high-grade iron ore, mainly at Ouenza. In the future it is hoped to begin exploiting the 14,000,000,000 tons of poorer quality ore at Gara Djebilet, near Tindouf; half of this could be mined in surface workings, but there are serious transport problems. One solution would be to create a new industrial complex, using the country's bituminous coal deposits, at Béchar, 470 miles to the northeast.

Deposits of nonferrous metal ores are small and widely scattered, and in most cases the existing transport facilities prevent economic mining. There are, however, sizable deposits of zinc and lead at el-Abed on the Moroccan border near Tlemcen, and some 250,000 tons of mercury ore were being mined annually in the mid-1970s at Ismail. Various studies have been undertaken to see how best to exploit the other unmined deposits of these minerals, as well as the known reserves of copper and

pyrites, and intensive prospecting for minerals in the Ahaggar has been carried out with Soviet assistance. Zinc ore is mined at el-Abed. Traces of tin, nickel, cobalt, chrome, and uranium have been found, and there are also high hopes of finding tungsten, gold, platinum, and diamonds. Even if sizable deposits are discovered, however, it will still be necessary to improve transport and washing facilities before it will be feasible to exploit the Sahara's undoubted mineral wealth.

Phosphate deposits of relatively inferior grade are located south of Tébessa at Djebel Onk on the Tunisian border; reserves in 1970 were estimated at 1,000,000,000 tons, of which about one-fifth can be mined by surface works. Algeria has sizable kaolin deposits at Djebel Debbagh and large reserves of marble and onyx.

Biological resources. The country's intensively cultivated land is restricted almost entirely to the narrow coastal plain that was colonized under French rule, when the settlers established vineyards, orchards, citrus groves, and farms. The best farms were located in the well-watered and fertile plains around Bejaïa and Annaba in the east, in the Mitidja Plain south of Algiers, and beyond Oran from Sidi bel Abbes to Tlemcen. The Médéa and Mascara plateaus are also ideal for grapes. The mountains that isolate the coastal plain from the south are heavily covered with dense forests of cork oak and cedar in the east and of pine in the west. Behind the mountain chains lie dry plateaus—notably the Plains of Constantine in the east and the Sersou Plateau to the west—where cereals are grown. In the Aurès region to the east, cereals are grown on the mountain slopes, and there are orchards in the valleys. In the west there is a large natural crop of esparto grass, between 150,000 and 200,000 tons a year, grown on some 7,500,000 acres (3,000,000 hectares).

Shortage of arable land

More than four-fifths of the country's surface land area, 588,700,000 acres (238,200,000 hectares), is uncultivable, and erosion adds another 100,000 to the total each year. An estimated 72,200,000 acres in the Sahara can be classed as semidesert pasture. Just under 6,000,000 acres are covered by forests. In 1970 there were some 109,300,000 acres of agricultural land, of which 85 percent was pasture or scrubland, about 15 percent was tilled, and the rest was devoted to vineyards and orchards. Winter grains (wheat, barley, and oats) are grown on the largest area of arable land. Cotton, tobacco, olives, and dates are important, and sorghum, millet, maize, rye, and rice are also grown.

The climate is ill suited to extensive stock raising. In 1974 there were an estimated 8,100,000 sheep, about 2,400,000 goats, and about 950,000 head of cattle.

Water resources. Irregular rainfall has long been a threat to agriculture and has limited hydroelectric development. At independence (1962), Algeria possessed some 20 sizable dams, altogether holding only 53,000,000,000 cubic feet (1,500,000,000 cubic metres) of water a year. In the early 1970s, only one-fifteenth of the river water had been trapped for irrigation; it was estimated that it would take 30 years to satisfy water requirements. Construction of dams was under way, however.

Sources of national income. *Agriculture, forestry, and fishing.* In the early 1970s between 10 and 15 percent of Algeria's gross domestic product was accounted for by agricultural output, although more than 55 percent of the population lived off the land.

There are more than 600,000 privately owned farms covering 14,400,000 acres, all but 5 percent being used to grow cereals. Yields per acre, however, are only three-fifths of those recorded for state-owned farms. More than half of the farms occupy less than 13 acres (five hectares) each, and 98 percent are less than 125 acres (50 hectares) each.

State-owned farms

About 2,000 farms are owned by the state, covering only 5,700,000 acres and employing only one-fifth of the total agricultural work force. They account for 75 percent of output, however, including 85 percent of the wine production, 60 percent of the citrus fruits, and three-quarters of the vegetable crops. These modern efficient farms are those that were vacated by the French settlers in 1962. They were run by independent workers' committees until more direct centralized control was instituted in the late 1960s.

The annual cereal production of about 1,300,000 tons in the mid-1970s was substantially short of domestic requirements. Efforts are being made, however, to increase the output of cereals and other crops so that imports can be reduced. Annual production of wine in the mid-1970s was well over 740,000 tons. Following the French government's policy of reducing wine imports to protect its own peasant growers, however, the Algerian authorities decided to convert the vineyards to other uses. It was expected that the land covered by vines would

ALGERIA

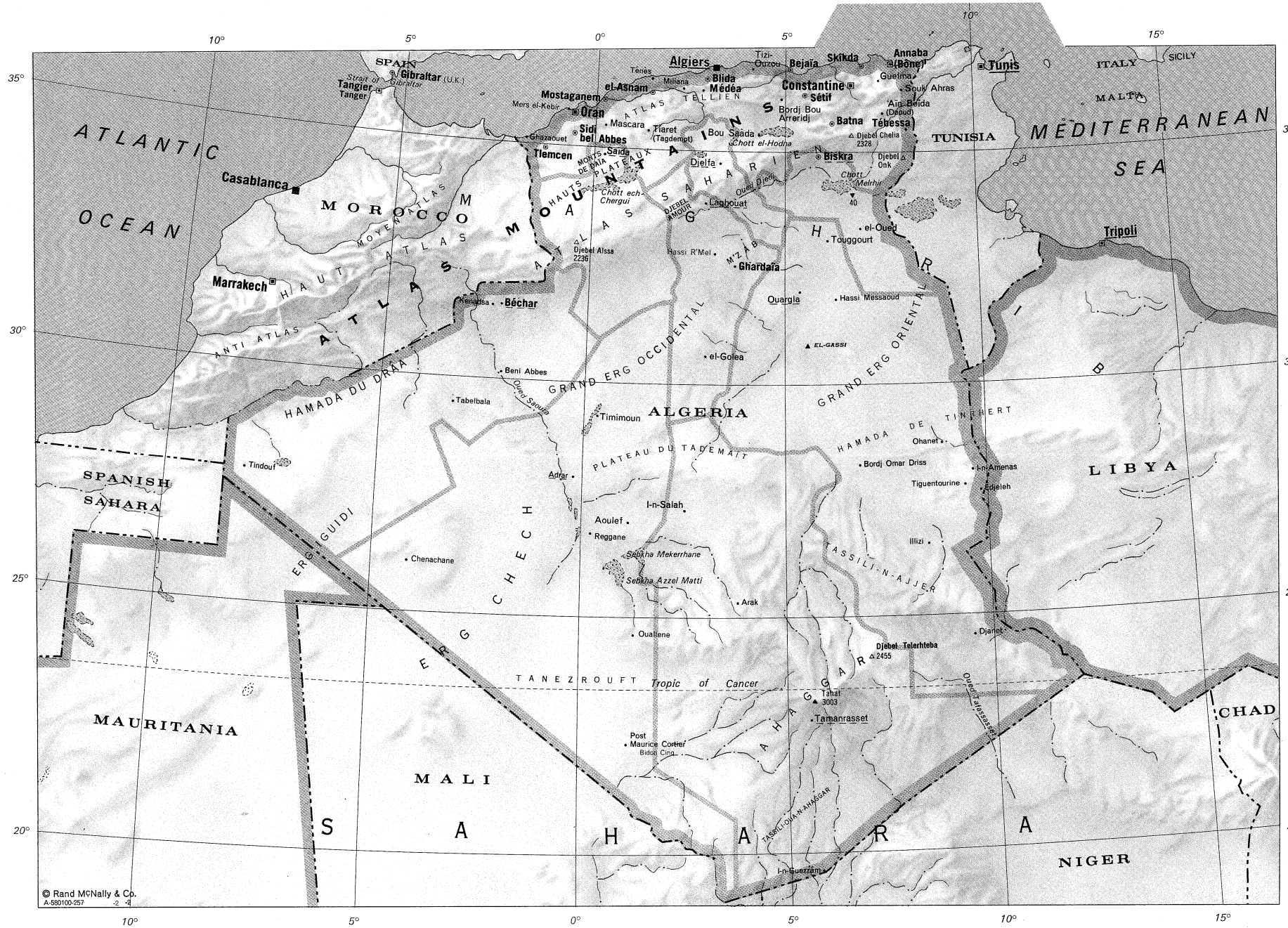
MAP INDEX

Cities and towns

Adrar 27·51n 0·19w
'Aïn Beïda (Daoud) 35·48n 7·23e
Aïn Salah, see I-n-Salah
Algiers 36·42n 3·08e
Annaba (Bône) 36·54n 7·46e
Aoulef 26·55n 1·02e
Arak 25·18n 3·42e
Batna 35·34n 6·11e
Béchar 31·37n 2·13w
Bejaïa (Bougie) 36·45n 5·05e
Beni Abbes 30·05n 2·10w
Bidon Cinq, see Post Maurice Cortier
Biskra 34·50n 5·43e
Blida 36·30n 2·49e
Bône, see Annaba
Bordj Bou Arreridj 36·02n 4·46e
Bordj Omar Driss 28·09n 6·49e
Bougie, see Bejaïa
Bou Saâda 35·11n 4·09e
Chenachane 26·04n 4·15w
Constantine 36·25n 6·38e
Daoud, see 'Aïn Beïda
Djanet 24·33n 9·29e
Djelfa 34·41n 3·15e
Edjeleh 27·49n 9·57e
el-Asnam (Orléansville) 36·10n 1·20e
el-Golea 30·35n 2·53e
el-Oued 33·20n 6·52e
Ghardaïa 32·31n 3·41e
Ghazaouet 35·06n 1·51w
Guelma 36·28n 7·26e
Hassi Messaoud 31·40n 6·08e
Hassi R'mel 32·56n 3·16e
Illizi 26·29n 8·28e
I-n-Amenas 28·05n 9·30e
I-n-Guezzâm 19·32n 5·42e
I-n-Salah (Aïn Salah) 27·12n 2·28e
Kenadsa 31·34n 2·27w
Laghouat 33·50n 2·53e
Mascara 35·24n 0·08e
Médéa 36·16n 2·45e
Mers el-Kebir 35·45n 0·43w
Messaoud, North and South, see Hassi Messaoud
Miliana 36·16n 2·13e
Mostaganem 35·56n 0·07e
Ohanet 28·32n 8·50e
Oran (Ouahran) 35·43n 0·38w
Orléansville, see el-Asnam
Ouallene 24·37n 1·14e
Ouargla 31·59n 5·20e
Philippeville, see Skikda
Post Maurice Cortier (Bidon Cinq) 22·18n 1·05e
Reggane 26·42n 0·10e
Saïda 34·50n 0·11e
Sétif 36·09n 5·26e
Sidi bel Abbes 35·13n 0·38w
Skikda (Philippeville) 36·50n 6·58e
Souk Ahras 36·17n 7·57e
Tabelbala 29·23n 3·15w
Tagdempt, see Tîaret
Tamanrasset 22·50n 5·25e
Tébessa 35·24n 8·09e
Ténès 36·31n 1·19e
Tîaret (Tagdempt) 35·23n 1·21e
Tiguentourine 27·50n 8·43e
Timimoun 29·14n 0·16e
Tindouf 27·40n 8·08w
Tizi-Ouzou 36·43n 4·02e
Tlemcen 34·52n 1·15w
Touggourt 33·06n 6·04e

Physical features and points of interest

Ahaggar, *mountains* 23·00n 6·30e
Ahaggar, Tassili-oua-n-, *plateau* 21·00n 5·00e
Aïssa, Djebel, *mountain* 32·56n 0·38w
Ajjer, Tassili-n-, *mountains* 25·54n 70·48e
Amour, Djebel, *mountains* 33·50n 2·05e
Atlas Mountains 33·00n 2·00w
Atlas Saharien, *mountains* 33·25n 1·20e
Atlas Tellien, *mountains* 35·00n 3·00e
Azzel Matti, Sebkha, *salt flat* 25·55n 0·56e
Chech, Erg, *desert* 25·00n 2·15w
Chelia, Djebel, *mountain* 35·19n 6·42e
Chergui, Chott ech-, *lake* 34·15n 0·30e
Daïa, Monts de, *mountains* 34·39n 0·15w
Drâa, Hamada du, *plateau* 29·00n 6·45w
el-Gassi, *gasfield* 31·05n 5·40e
Grand Erg Occidental, *desert* 30·30n 0·30e
Grand Erg Oriental, *desert* 30·30n 7·00e
Hodna, Chott el-, *lake* 35·30n 4·45e
Hoggar, see Ahaggar, Tassili du
Iguidi, Erg, *dunes* 26·35n 5·40w
Maghrib, *physical region* 33·00n 5·00e
Mekerrhane, Sebkha, *salt flat* 26·19n 1·20e
Melrhir, Chott, *lake* 34·17n 6·24e
M'zab, *physical region* 32·43n 3·50e
Onk, Djebel, *mountain* 34·52n 7·57e
Saoura, Oued, *watercourse* 29·00n 0·55w
Tademaït, Plateau du 28·30n 2·00e
Tafassasset, Oued, *watercourse* 25·15n 10·15e
Tahat, *mountain* 23·18n 5·47e
Tanezrouft, *desert* 24·00n 0·45w
Telerhteba, Djebel, *mountain* 24·09n 6·58e
Tinrhert, Hamada de, *plateau* 29·00n 9·00e



ALGERIA

Size of symbol indicates relative size of town • ⊙ ⊡ ▣ ■

Elevations in metres

be reduced to 655,000 acres and there were plans to convert a further 160,000 acres before the end of the decade. In the early 1970s about 450,000 tons of citrus, 150,000 tons of dates (of which 70 percent was exported), 18,000 tons of olive oil, and 570,000 tons of vegetables were produced annually. Thirty percent of the milk, 40 percent of the sugar, 75 percent of the cooking oils, and 95 percent of the butter, however, have to be imported each year, and strenuous efforts were being made in the early 1970s to increase production of these commodities.

Production capacity of cork is about 80,000 tons a year, and exports earn approximately 40,000,000 dinars annually. Some cork is also processed domestically. About 160,000 acres of forest were lost in the War of Independence, adding to the serious problem of soil erosion. Schemes for reforestation were introduced in the 1960s and continued in the 1970s. Over 60,000,000 trees, mainly pines, were planted in 1967 and 1968.

The fishing industry is underdeveloped, but the state fishing authority planned to acquire new trawlers and equipment in the early 1970s, which would almost double the catch.

Mining and quarrying. The country's main mining centres are at Ouenza and Djebel Onk on the eastern border with Tunisia and at el-Abed and Béchar on the western boundary with Morocco. An electrified railway links the Ouenza and Djebel Onk workings to Annaba, while both el-Abed and Béchar have rail links with Oran.

The opencut works at Ouenza produces over 2,500,000 tons of high-grade iron ore each year—well over 80 percent of national production. Output of phosphates from the Djebel Onk mines was under 450,000 tons in 1971 and it was planned to treble this figure by 1973. Most of the iron ore and phosphates are exported, but following the completion of the steel and fertilizer complexes at Annaba in 1972, some production was diverted to them. There are also plans to supply the steelworks with coke from the Béchar coal mines, where production fell from 78,000 tons in 1961 to around 15,000 tons a year in the late 1960s and early 1970s.

El-Abed is Algeria's largest zinc and lead mine, producing two-thirds of the nation's output of zinc (31,000 tons in 1971) and lead (8,000 tons in 1971). Since 1967 ore has been treated by a Moroccan washing plant. It was planned that by 1973 production of zinc concentrates should reach 127,000 tons a year and output of lead concentrates 30,000 tons.

Manufacturing. Before independence the manufacturing sector was mainly confined to food processing, textiles, cigarettes, and clothing. Since 1967, however, the main emphasis has been put on heavy industry. The state steel corporation completed its Annaba steelworks in 1972, giving an annual output capacity of 1,200,000 tons for the mid-1970s. The corporation was also constructing a zinc electrolysis plant near the el-Abed mine at Ghazaouet in the early 1970s; expected output was estimated at 40,000 tons annually. Over one-tenth of the steel produced for domestic consumption was to be eventually allocated for producing machine tools, tractors, agricultural equipment, trucks, and automobiles (about 13,000 automobiles were assembled in 1970). In the early 1970s diesel engine, tractor, and cycle factories near Constantine, and two more automobile manufacturing plants, were being constructed. The state was also developing three major petrochemical centres at Arzew, Skikda, and Sétif. In addition to the gas liquefaction plants at Arzew, a nitrogenous fertilizer factory was opened in 1970, and an oil refinery and a liquid petroleum gas separation plant were completed nearby in 1972. In 1972 another gas liquefaction plant, an ethylene factory, and a liquid petroleum gas separation works were opened at Skikda, while a plastics factory and a 4,000,000-ton per year benzene refinery were to be completed by the mid-1970s. The Sétif complex, also to be finished by the mid-1970s, comprises methanol and plastics factories. Another major project in the heavy industry sector is the phosphate fertilizer factory, completed at Annaba in 1972.

Steel production

State corporations also dominate the light manufacturing sector. In 1966, 40 percent of output in this sector came from state-run canneries and mills. The state textile concern produced 307,000,000 feet of yarn in 1969, two-thirds of the country's production, and the state tannery produced over 215,000,000 square feet of leather, sufficient to produce almost 19,000,000 pairs of shoes. Output in the shoe and clothing industries was expected to rise by a third in the early 1970s, and the state also hoped to treble the 1969 paper production of 50,000 tons in the 1970s, absorbing some of the esparto crop. The state sugar corporation hoped to produce 250,000 tons of sugar a year by 1972 with two new mills, thereby meeting domestic demand. In the building supplies industry, the state was constructing a cement works near Constantine that would increase national production to 1,000,000 tons a year. Brick production also was to increase by a half to 700,000 tons by 1973. The state chemicals corporation met home demand for paints, soap, and detergents in 1971 with an output of 60,000 tons.

Energy. Following the discovery of oil in the mid-1950s the main prospecting and producing companies were two French groups, CFP-A (Compagnie Française des Pétroles Algérie) and ERAP (Entreprise de Recherches et d'Activités Pétrolières). They were later followed by some of the international oil companies, such as Mobil, Phillips, Atlantic Richfield, Esso, and Royal Dutch/Shell. Under a 1965 agreement, the French groups established a privileged position whereby they paid lower royalties and repatriated a smaller proportion of their revenues from selling the oil abroad in return for agreeing to take a leading part in prospecting for new oil fields. In the late 1960s there were frequent arguments over the terms of the agreement, the Algerians claiming that the French companies were not doing as much prospecting as had been agreed to. Consequently the state-owned Algerian oil concern, Sonatrach (Société Nationale pour la Recherche, la Production, le Transport, la Transformation et la Commercialisation des Hydrocarbures), which had been set up in 1963–64, extended its activities into exploitation and production on its own account with some success. Much of this development was accomplished with Soviet assistance and more recently by setting up a series of joint service companies with American specialist help.

The Algerians became increasingly angry in the period from 1966 to 1970 when annual production only rose by one-third to about 47,000,000 tons. Expenditure on prospecting by the foreign oil companies (1966-69) only rose from 25,000,000 dinars to 180,000,000, while Sonatrach's own exploration budget rose from 40,000,000 dinars to 280,000,000. In a series of moves from 1967 to 1970 the international oil companies were nationalized, the biggest being the Royal Dutch/Shell-owned CPA (Compagnie des Pétroles d'Algérie) with 10 percent of production. The only exception was Getty Oil, which settled its differences with the Algerian authorities in 1968. In 1971 Algeria nationalized the majority shares in the French companies operating in the country. Since then, France has not relied on Algeria for one-third of its oil supplies.

Crude-oil production was expected to rise from 47,000,000 tons in 1970 to 65,000,000 tons by 1973 because of increased pipeline capacity. Basically there are four existing pipeline links between the oil fields and the Mediterranean: the line from Hassi Messaoud to Arzew, with a 22,000,000-ton annual capacity; the twin pipes from Hassi Messaoud to Bejaïa, with an 18,000,000-ton annual capacity; the line from In Amenas and Djelem north to aṣ-Ṣukhayrah (La Skhira) in Tunisia, which has a 15,500,000-ton annual capacity; and the line from Hassi Messaoud to Skikda, opened in 1972, which began its operations with a 12,000,000-ton annual capacity.

Production of natural gas

Exports of natural gas are likely to prove far more important for the economy than sales of oil by 1980. Production of natural gas rose from about 8,000,000,000 cubic feet in 1961 to more than 96,000,000,000 cubic feet in 1970. Much natural gas is shipped to France and the United Kingdom, but far more important contracts than those concluded with these two countries were

reached with North American interests in 1969. Spain is another purchaser. The possibility of piping gas directly to Europe by two routes—one via Sicily, Italy, and the other across the Straits of Gibraltar and across Spain—was also being considered.

Domestic consumption of natural gas in the late 1960s was about 35,000,000,000 cubic feet a year, mainly in Algiers and Oran, which are fed by branches of the Arzew pipeline. Consumption was expected to treble after the opening of the Skikda pipeline (1972), which would make gas available to the eastern cities. When the petrochemical industries are established by 1975, consumption should exceed 141,000,000,000 cubic metres annually.

The consumption of electricity was expected to almost double its 1969 level and reach 2,800,000,000 kilowatt hours by 1973. The thermal stations at Algiers, Annaba, and Oran have switched from coal to gas and oil fuel in recent years, thus increasing their production. Another steam generator was being built at Annaba, and two more were planned for Oran and Skikda, as well as a number of smaller oil- and gas-fired plants in the underdeveloped regions. The hydraulic output of the Kabylie region was about 25 percent of the total in 1971.

Consumption of refined oils almost doubled between 1967 and 1971. Sales of motor oil reached 450,000 tons in 1971, and the country's main oil refinery at el-Harrach (formerly Maison Carrée) was hard pressed to meet demand. Opened in 1964, it has an annual throughput of 2,500,000 tons, brought by a spur of the main Hassi Messaoud-Bejaïa oil pipeline, which was opened in 1972. By 1973 another 2,500,000-ton refinery at Arzew was to be completed that would produce 50,000 tons of lubricants and 600,000 tons of bitumen a year, which would significantly cut back imports. There is a small (165,000-ton) refinery at Hassi Messaoud.

Financial services. An independent central bank, established at the beginning of 1963, has the sole right to issue bank notes in Algeria and also carries out the normal central banking duties. In 1967 there were five Algerian and nine foreign banks operating in the country, but since nationalization there have only been four state-owned commercial banks.

Foreign trade. The value of Algerian exports climbed from 3,747,000,000 dinars in 1963 to 4,980,000,000 in 1970, mainly due to increased shipments of oil that accounted for two-thirds of total revenue. Following the imposition of French restrictions, wine-export receipts have been halved to about 465,000,000 dinars. Exports of raw materials, natural gas, fruits and vegetables, and manufactured goods each account for about 5 percent of total earnings.

In 1969, industrial equipment cost 1,196,000,000 dinars, and consumer goods accounted for another 992,000,000, despite strict controls. Imports of steel were 12 percent of the total bill, and raw materials accounted for another 6 percent. In an average year, imports of cereals account for about 250,000,000 dinars, of milk and sugar for over 150,000,000, and of other food products, drink, and tobacco for double this figure.

Trade with France

Algerian trade with France has dropped from 80 percent of the total in 1961 to about 45 percent in 1970. Attempts to increase trade with Communist countries have partly failed, most of the former French trade going to other Common Market countries. In 1970 the European Economic Community provided over 65 percent of Algerian imports and took just under three-quarters of its exports. The 600,000 Algerians in France remit some 1,250,000,000 dinars a year to relatives in Algeria, and this is partly responsible for the healthy balance-of-payments position.

Management of the economy. *The private sector.* Following independence many firms were taken over by their workers and more recently the government has nationalized many of the remaining privately owned businesses. This was done both to give the government the direct means of dealing with chronic stagnation and mass unemployment and to bring pressure to bear on foreign interests. There is still, however, a sizable private sector in the textile, rubber, plastics, and engineering industries. In the late 1960s there was a marked rise in the number of private investment proposals submitted to the authorities, but this trend slowed in the early 1970s.

The public sector. Since independence the public sector has expanded considerably and by the 1970s there were well over 30 national corporations and boards that had monopoly powers in various industries. The most important is undoubtedly the oil, gas, and petrochemicals giant, Sonatrach; in 1971 it controlled over three-quarters of oil production, the entire refining and oil distribution network, and all natural gas operations. Other corporations and boards operate in steel, shipping, power, engineering, textiles, glass, timber, cork, ports, pasta, canning, chemicals, wine marketing, sugar, tobacco, construction, exports, road transport, tanning, and the railways.

Taxation. The tax system is substantially that inherited from France, but it was due for radical reforms in the early 1970s. The main taxes were a 50 percent corporation tax, with rather lower rates of taxation on unincorporated businesses; various taxes on salaries; a value-added tax; local sales taxes; and various stamp duties and levies. Oil was expected to provide more than one-third of the total budget for 1972.

Trade unions. The centralized nature of the economy reduces the importance of the country's national trade union, Union Générale des Travailleurs Algériens (UGTA). Following an abortive coup in 1967, dissident elements were purged; this has further weakened its influence.

Contemporary policies. The government concentrated investment in small factories in the early 1960s to alleviate the unemployment that resulted from the economic chaos and stagnation that followed the French exodus. But since 1965 a more centralized economy has been developed and the government hopes that its ambitious capital projects will give the impetus to solve the unemployment problem. The 1970–73 economic plan aimed to increase the gross national product from 15,800,000,000 dinars to 22,185,000,000. The state hoped to provide three-quarters of the funds, mostly from increased oil royalties but also by deficit spending. This explained the government's continuing attempts to get the producing nations to present a united front on oil policies.

Centralization of the economy

Problems and prospects. The government was faced with a number of economic problems in the 1970s—notably converting a predominantly rural economy into an industrial one, developing the country's vast potential for tourism, and increasing oil and natural gas revenues. The most daunting problems are those of the high birthrate, illiteracy, and vast unemployment. In the early 1970s, between 500,000 and 800,000 people were jobless in the cities, whereas only 200,000 were employed in industry. Seventy thousand new jobs a year are needed just to keep pace with the natural growth of available manpower.

(E.I.U.)

Transportation. *Road transport.* There are about 54,000 miles of roads of which about 11,360 are national highway. A good road network exists in the Tell, and there are established routes and motor transport for passengers and freight across the Sahara. Roads have also been built to link the Sahara oil fields with the coastal ports. The Société Nationale des Transports Routiers (SNTR), a government agency, was created in 1967 to coordinate road transport—more than 80 percent of the enterprises engaged in road transport have not more than two vehicles—and to established fixed rates for the transport of goods. As a majority of the vehicles in use on Algeria's roads were more than 10 years old, funds were made available to replace 70 percent of the total fleet of vehicles by the end of 1970. Small trucks have been brought into service for use in regions where the terrain or the state of the roads does not permit larger vehicles to pass.

The railroads. A central rail line runs from the Moroccan to the Tunisian frontier, with branch lines to the chief ports, as well as lines running southward from Oran to Crampel, Kenadsa, and Mehdia; from Algiers to Djelfa; and from the coastal ports Skikda and Annaba to

Touggourt, Khenchela, and Tébessa. The Algerian railroad system is operated by the Société Nationale des Chemins de Fer Algériens (SNCFA).

The ports. The principal ports are Algiers, Oran, and Annaba. Other ports are Arzew, Bejaïa, Djidjello, Ghazaouet, and Mostaganem; Skikda is an oil port.

Air services. Air Algérie, which is largely government sponsored, provides domestic as well as international air services. Algiers has an international airport, Dar el-Beïda, from which regular flights leave for France, Spain, Switzerland, East and West Germany, the United Kingdom, Italy, Belgium, the Soviet Union, Tunisia, and Morocco. Algiers airport is also a principal stop on regular air routes between Europe and Africa south of the Sahara, as well as on routes connecting the Americas and Eastern countries.

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of the government. The first Algerian Constitution, adopted by Parliament on August 28, 1963, and approved by referendum on September 8, 1963, instituted a presidential type of regime modified by the responsibility of the head of state to the National Assembly. There is a single party, the Front de Libération Nationale (FLN), which plays a preponderant role. Since its creation on June 19, 1965, the Revolutionary Council has been the supreme organ of the party as well as the government. In 1970 an Economic and Social Council was established, including representatives of all segments of the nation.

Elections to popular communal assemblies were held on February 5, 1967; and elections to assemblies of the *wilāyāt* were held on May 25, 1969. Elections to the National Assembly and to the higher organs of the state were scheduled to be held in the 1970s.

Local government. The political life of the state depends upon the life of the commune, which constitutes the basic political unit.

When Algeria won its independence, it was faced with the problem of building a strong, balanced, coherent, and united state; to do so, however, required the reconciliation of two seemingly opposed concepts—power and liberty. It is undeniable that the systematic centralization of decision making is hardly favourable to national expansion, since it leads, in effect, to a species of atrophy of thought and initiative on the part of citizens at all levels. On the other hand, it is equally true that uncontrolled initiatives that are carried to extremes may be equally harmful, since they may lead to a species of general anarchy. The Algerian system, which permits the two tendencies to operate harmoniously together, permits the active participation of each citizen in governmental action. The primary means of applying this "technique" of participation in practical terms is provided by a systematic decentralization carried out in a resolutely democratic spirit. The commune thus becomes the unit primarily responsible for economic, social, and cultural progress. Since 1965, therefore, the chosen path has been to make the people themselves, or their chosen representatives, the source from which national institutions derive their authority.

All the citizens of Algeria are grouped together in communes, that is to say in smaller or larger collective groupings each of which has its own social, economic, and human particularities.

The Algerian government has thus committed itself to the belief that each communal unit can and will undertake responsibility for its own affairs while acting within the general framework of national evolution.

Administrative and political decentralization. The constitution of the popular communal assemblies permits the members of the same territorial grouping to choose from among themselves those whom they consider to be the most qualified to represent the commune for all purposes. In this way the administrative structure encourages local autonomy and the organizing of communal interests in the direction most favourable to their development.

The popular communal assemblies are elected every four years by universal suffrage by the inhabitants of each locality; the electoral lists contain twice as many candidates as there are posts to be filled. The assembly directs the affairs of the commune. It meets four times a year in regular session, but may also meet whenever an important problem arises. The executive of the popular communal assembly is composed of a certain number of members, varying according to the population of each commune.

Economic and cultural decentralization. The communal code, which provides a legal framework for economic and social policy, establishes six separate sectors of economic activity: agriculture, industry and crafts, distribution and transport, tourism, housing and environment, and education and culture.

In agriculture, each assembly is responsible for the productivity of the territories in its domain. Its authority extends over estates no less than over tracts of land that are worked by individuals. In industry, the assembly seeks to stimulate industrial development, whether it is a matter of encouraging or expanding local craftsmanship or of building a new industrial enterprise. The assembly may create new industries or may participate by investing capital in those that are of some public interest. In distribution the assembly encourages the establishment of state outlets or of consumer cooperatives. It uses the public transport services made available by individuals operating in its territory or can, if need be, establish such services. Considerable initiatives may also be taken by the assembly with respect to tourism. The tourist attractions of Algeria are so much in evidence that immense perspectives are open. The assembly manages existing tourist facilities and is entrusted with responsibility for the profitability of new ones. The assembly is also responsible for the management of state property. It participates in educational building programs and creates and manages sporting facilities, as well as entertainment facilities. It surveys the leisure and individual or collective educational requirements of the local inhabitants, particularly so far as youth is concerned. In sum, the assembly is responsible, with the consent of the governed, for the entire direction of local affairs.

The governmental decision to invest extensive powers in the communes was carried out, following the enactment of the communal code, by a certain number of legal texts, by the terms of which the state conceded to the communes the right to exploit the national heritage, for which hitherto the state alone had been responsible. The concessions made applied to different sectors. By a decree of March 17, 1967, for example, the communes were accorded the right to use funds gained from the operation of local cinemas. A decree of April 27, 1969, granted the communes collective and individual rights related to tourist facilities in their respective areas. The income so derived, previously the property of the state, thus provided a marked increase in the revenue available for the local budget. By a decree of September 27, 1967, the State gave to the communes the buildings that constituted the former resettlement centres that were established during the struggle for independence, as well as other groups of buildings that were built for purposes of reconstruction after independence or to meet emergencies. The land upon which the buildings stood also became the property of the commune, which can reassign these properties to the occupants or to anyone it pleases. The earnings from such sales are applied to the budget of the commune and have frequently been used for the purchase of public equipment.

Economic concessions to the communes

Governmental control over the activities of the communes is restricted to the legal coordination of the different communal initiatives.

Communal finances. The resources of each commune vary. They are usually assured by the income from local taxes and by the sources of revenue turned over to the communes by the state. To render local activities more profitable, however, the commune is obliged to invest in equipment. In this connection the state intervenes to share out public finances with a threefold purpose; firstly, of correcting the more marked intercommunal imbalances from public funds; secondly, of providing direct financial aid that the commune may use to meet

its most pressing requirements; and thirdly, of guaranteeing the fiscal receipts anticipated in the commune's budget. This interaction between the state and the communes requires a constant interpretation of governmental and local responsibility.

The wilāya. Between the central government and the commune there is an intermediate local authority—that of the *wilāya* (province, or *département*). The *wilāya* represents a decentralization of power by the creation of the Assemblée Populaire de Wilaya (APW)—the popular assembly of the *wilāya*—which is elected by universal suffrage.

The APW itself has an executive council, as well as a representative of the central government, who is known as the *wali.* As with the commune, but evidently on a wider regional scale, the *wilāya* favours a balanced apportionment of tasks and responsibilities. A new territorial organization is now being studied which will tend to establish a rational geographic system of territorial division; here again the intent of the government is to palliate regional disparities that have hitherto caused some evident anomalies.

The APW is elected for a term of five years. It consists of from 35 to 55 members, depending upon the density of the population of the *wilāya;* 35 members are elected when there are less than 25,000 inhabitants, from 39 to 51 members when the population numbers from 25,000 to 250,000, and 55 members when the population numbers over 250,000. The APW normally meets three times a year. Upon the request of the *wali* and of two-thirds of the assembly's members, it may also meet in extraordinary session. The APW is particularly concerned with such regional questions as irrigation, reforestation, rural renewal, and health measures for livestock. It is also concerned with such other matters as environment, building programs, the establishment of new urban zones, road construction, public health and hospitals, schools, and tourism. The APW also establishes its own permanent commissions to study financial, economic, social, and cultural problems. Temporary commissions may also be formed for the investigation of particular problems. The APW, more than the commune, is suited to the study and promotion of regional economic development. Its budget, which is modelled on that of the central government, is based on revenues derived from taxes that it is empowered to impose. It is also empowered to raise loans and to make investments in the public interest.

The executive council of the *wilāya* is the chief regional authority. It is composed of different directors of the state services who are established in the *wilāya;* the *wilāya* is thus oriented to both regional and national levels. Under the authority of the *wali,* the *wilāya* exercises trusteeship and administrative control of the local collectives and of public establishments, as well as of independent enterprises and national societies. Within the national government it participates in the development and application of the general development plan and assists in coordination in matters that concern the *wilāya.*

The authority of the *wali*

The *wali* himself is nominated by decree. He is at the same time the sole and permanent representative of the central government as well as of the *wilāya.* He assumes those functions that must necessarily be called into play in the relationship between the governmental summit and the popular base. As the representative of the *wilāya* he presides over the implementation of the decisions of the APW. As a high state functionary, he is the delegate of the government and the direct representative of each of its ministers.

The political process. The Front de Libération Nationale (FLN), as noted above, is the sole political party. The role of the FLN was defined by President Houari Boumédienne, who stated that as the avant-garde of the Algerian Revolution, the party has the task of elaborating, giving direction, animating, and checking the revolution—but not of directing the state or of substituting itself for the state.

Since December 10, 1967, the FLN has been administered by a member of the party apparatus. In the early 1970s, a party reorganization was in progress.

Justice. Following independence, Algeria inherited colonial judicial institutions that were primarily established in the larger towns where the European communities were concentrated, having previously been oriented toward the maintenance of colonial authority. A jurisdictional duality also existed, since judicial organization was founded upon the separate foundations of so-called common law and Muslim jurisdiction. On June 19, 1965, the entire system was reformed by a decree that instituted a new judicial organization. This decree was soon followed, in June 1966, by the promulgation of new legal codes—the penal code, the code of penal procedure, and the code of civil procedure. In formulating the new codes, the primary preoccupation was to make justice available to those who demanded it by permitting the citizen to appeal for justice with a minimum of delay and a maximum of efficaciousness. Fifteen courts and 132 tribunals were created, thus establishing a system of justice that covered the entire country, including the southern areas.

Three levels of justice were established. At the first level there exists the tribunal, to which civil and commercial litigation is submitted, and which takes action in penal cases of the first instance. A single judge is empowered at this level. At the second level there is the court, which consists of a college of three members. It functions as a court of appeal for the tribunals and for the administrative jurisdictions of the first instance. At the third and highest level is the supreme court, which checks and unifies the ensemble of jurisprudence. It enjoys the powers of a final court of appeal and of appeals against the decisions of the lower courts.

Of the three fundamental legal documents on which the juridical system is based, the first—the penal code—reaffirms the principles protecting individual liberty; the second, the code of penal procedure, contains more detailed safeguards; while the third—the code of civil procedure—simplifies the conduct of judicial actions in order to eliminate delays, discourage abuses, and augment the efficiency of tribunals.

Three special courts for eliminating economic crimes against the state were established in 1966—those of Algiers, Oran, and Constantine; each is presided over by a member of the party and is composed of magistrates and of members representing the Ministry of Finance. Their decisions are without appeal.

A higher council of the magistrature is presided over by the head of state or by the minister of justice and is composed of a representative elected by magistrates, heads of the supreme court, party representatives, elected representatives, and of the administration; this higher council makes rulings on the recruitment, nomination, and promotion of magistrates. It also constitutes at the same time a disciplinary organ for the magistrates themselves.

The armed forces. The national popular army participates, within the framework of the party, in the party's political, as well as its economic and social, activities.

Ground forces. The ground forces consist of 53,000 men, with four or five battalions of Soviet T-34 and T-54 tanks and 15 artillery groups (including self-propelling guns, anti-aircraft groups, and missiles). Most of the equipment used by these forces is furnished by the Soviet Union.

Air force. About 2,000 men were trained in the Soviet Union and in Egypt. There are about 140 fighter planes and about 50 helicopters. After the sale of 28 Fouga Magister planes to Algeria by France in 1968, pilots began to train for flight at a school at Bou-Sfer near Oran.

The navy. The Algerian navy consists of 3,000 men; three minesweepers (supplied by Egypt), six coastal escort vessels, and eight motor torpedo boats comprise the naval craft.

The personnel of the police service and of auxiliary security forces are also armed. A total of about 75,000 men have been voluntarily recruited.

Compulsory national service of two years was instituted

in 1969. The recruits are employed in various kinds of national service, including road building, village housing construction, teaching, etc.

President Boumédienne assumed command of the army in December 1967.

Education. There were about 3,000,000 children of school age—that is to say, between the ages of 6 and 14 years old—in the country in the late 1960s, and by the year 1975–76 the number was expected to rise to almost 4,000,000. The educational system has experienced much difficulty in accommodating the increasing numbers of school-age children resulting from demographic growth. Despite considerable efforts, about 1,000,000 children could be accommodated in schools in the 1960s. It was anticipated, however, that by 1980 all children would be enrolled. About 55 percent of the school-age population received schooling by 1970.

Because of the scarcity of teachers, the government is utilizing educational radio and television. Although Arabization is progressing in terms of the number of hours devoted to its teaching, the level of instruction is mediocre, doubtless caused by the scarcity of qualified Arabic-speaking teachers. French is consequently being maintained as a medium of instruction not only at the higher and secondary levels but also in the primary schools. In the early 1970s a total reorganization of the teaching system was in progress: a committee for national educational reform was established in January 1970, its four subcommittees dealing with democratization, scientific and technical matters, Arabization, and the orientation and content of teaching in general.

The first institute of technology, specializing in agriculture, was opened in 1970 at Mostagenem. In compliance with the agrarian reforms instituted in 1972, more than 300 institutes of technology would be opened by 1974.

The primary university in North Africa before independence—the University of Algiers—has been able, despite difficulties, to maintain its pride of place, thanks to a highly qualified faculty and to the preservation of its links with the French university system. There is a second university in Oran and a third was opened in 1969 in Constantine.

Health and social affairs. Health policies in Algeria are primarily concerned with preventive medicine. Among health reforms recently instituted is a compulsory requirement that newly qualified doctors, surgeons, pharmacists, dentists, and midwives serve in state medical institutions for a time. Units providing free medical assistance are attached to various hospitals. Those earning middle incomes pay a part of their medical fees in proportion to their resources. Algeria as a whole is covered by a network of state pharmaceutical agencies. There are over 200 such agencies, of which 50 were established in 1968 alone, mostly in the southern regions, in the more rural areas, and on the periphery of the nation's larger towns.

The state has established a monopoly for the importation of medical drugs, with a view to obtaining lower prices. A factory for the processing of pharmaceuticals has been established and was expected to meet two-thirds of the general demand. The first drug-manufacturing plant began operation in 1971.

The Pasteur Institute, which is the central laboratory for the public health system, produces serums and vaccines. It also conducts research in microbiology, human and animal parasitology, and other areas. There is also a National Institute of Public Health.

Housing. The problem of housing is one of the most difficult of the social problems facing the nation. Apart from the maintenance and repair of many existing buildings, the government has assumed responsibility for the completion of housing programs that were abandoned following the massive exodus of French settlers to France after the achievement of independence.

Construction programs

In urban areas, the construction of over 12,000 new buildings was begun between 1969 and 1970, while 4,000 other buildings, abandoned by the French, were scheduled for completion in 1970. The construction of housing units with moderate rents has been subsidized by the government.

In rural areas, almost 10,000 buildings were completed in 1968 and 1969; 1,000 of these were in the Aurès region, 1,500 in the Tizi-Ouzou region, and 1,500 in the el-Milia region. Many of these housing projects were experimental in nature, and the lessons learned were applied to the Four-Year plan of 1970-73 as well as that of 1974-77, which envisioned bolder construction programs.

Cultural life and institutions. Press, radio, television, and the cinema are the principal media that inform the public and disseminate cultural influences. In order to reaffirm the link between modern Algerian life and its traditional roots, which predate the colonial interregnum, control of the public media has become a national responsibility. There is a national press agency, the Algerian Press Service (APS), and the national press includes such publications as *el-Moudjahid*, *al-Joumhouria-La République*, *an-Nasr*, *al-Chaab*. Television service is provided by Radiodiffusion Télévision Algérienne (RTA).

Prospects for the future. Algeria has changed greatly in the short period since the attainment of independence in 1962. After the ravages of the War of Independence, the Algerian people were faced with the necessity of reconstruction; the first fruits of the years of labour that have supervened are now being harvested.

Development policy, including plans for social and economic progress, was outlined by the Four-Year Plan of 1970–73, and the new plan for 1974–77. Both an agrarian and an industrial revolution are in progress; changes are occurring in education, the legal system, medicine, administration, and other fields. With a stable basis for future expansion, and with an effective leadership, Algeria may be expected to make further progress in overcoming the challenges it faces.

BIBLIOGRAPHY

General works: A good starting point is C.F. GALLAGHER, *The United States and North Africa: Morocco, Algeria, and Tunisia* (1963), which discusses in four sections the geography, history, politics, and foreign relations of these states not only with the U.S. but with other major countries, especially France. An introduction to the study of Algeria—geography, history and politics, economic conditions—is provided by NEVILL BARBOUR (ed.), *A Survey of North West Africa (The Mahgrib)*, 2nd ed. (1962). HILDEBERT ISNARD, *L'Algérie* (1954; Eng. trans., 1955), is a well-illustrated, authoritative travel guide covering the country region by region. A well-written, if sentimental, novel about life in Algeria is MARY MOTLEY, *Home to Numidia* (1964), which may be read for the atmosphere it evokes.

Geography: The classic human geography is JEAN DESPOIS, *L'Afrique blanche française*, vol. 1, *L'Afrique du Nord*, 3rd ed. (1964), which may be supplemented by his and RENE RAYNAL'S *Géographie de l'Afrique du Nord-Ouest* (1967), devoting a large section to Algeria. In English, J.M. HOUSTON, *The Western Mediterranean World* (1964); also to be consulted is JEAN DRESCH and PIERRE BIROT, "Afrique du Nord," in *La Méditerranée Occidentale, géographie physique et humain*, 2nd ed. (1964).

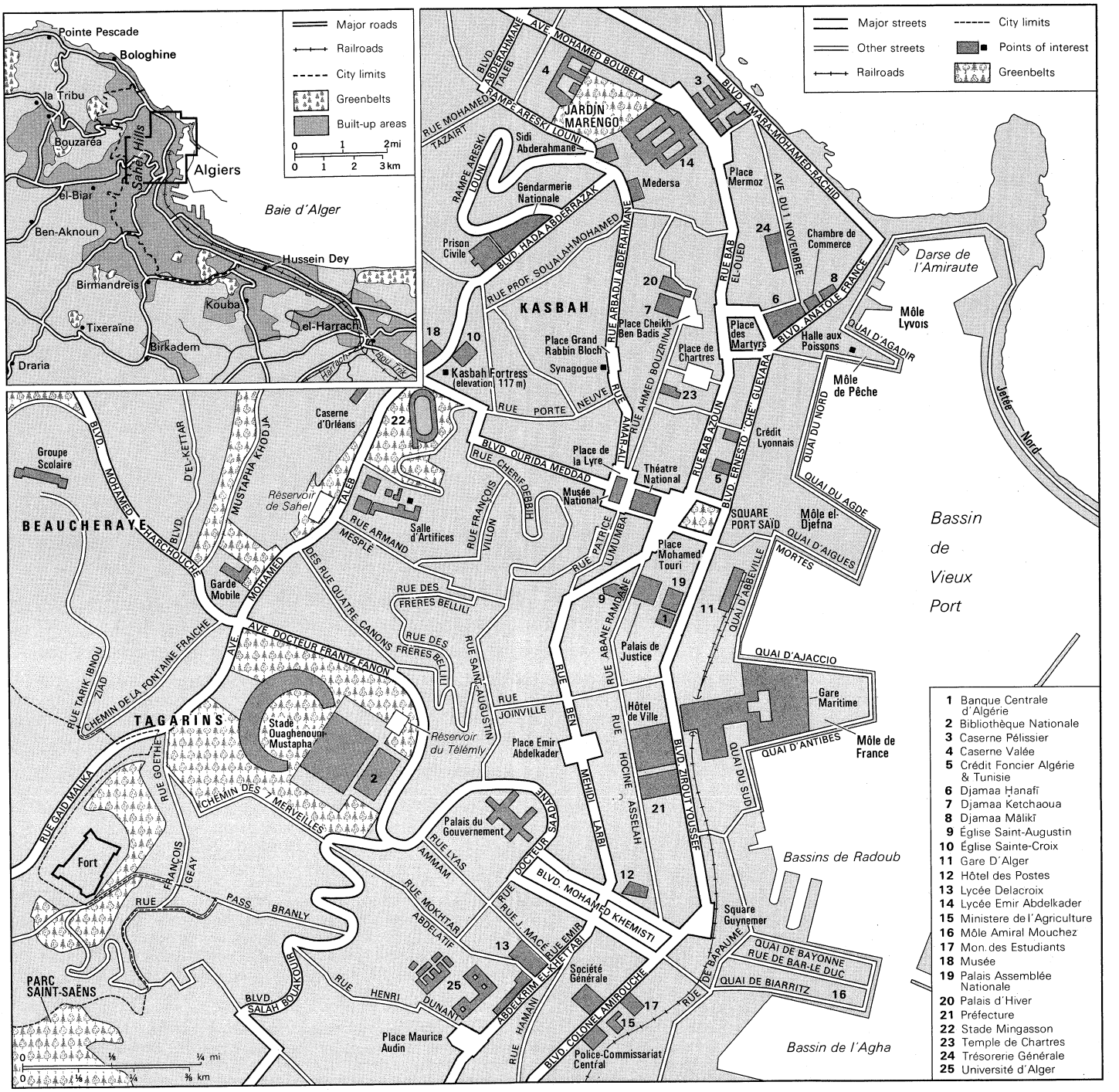
Sociology: A thorough sociological study is JACQUES BERQUE, *Le Maghreb entre deux guerres* (1962). See also PIERRE BORDIEU, *Sociologie de l'Algérie* (1961), in which the sociological analysis is conducted in terms of the main groups of peoples. A work of social anthropology relevant to Algeria is L.C. BRIGG, *Tribes of the Sahara* (1960), which includes major tribes of the Saharan regions of Algeria.

Economic conditions: J.C. PAWERA, *Algeria's Infra-Structure* (1964), is an economic study focussed chiefly on the transport, communications, and energy resources of the Algerian economy. FRANCOIS PERROUX (ed.), *L'Algérie de demain* (1962), is a collection of studies on possible paths of economic and social development. See also FRANCOIS D'ARCY, ANNIE KRIEGER, and ALAIN MARILL, *Essais sur l'économie de l'Algérie nouvelle* (1965), chiefly for socio-economic topics such as agrarian reform, administration of the state farms, and industrial economy.

(A.K.Ch.)

Algiers

Algiers (French Alger; Arabic al-Jazā'ir, "the Islands") is the capital and chief seaport of Algeria. The city and its



Central Algiers and (inset) its metropolitan area.

suburbs extend for 18 miles along the Bay of Algiers, climbing the slopes of the Sahel Hills to form a vast, white amphitheatre. The city derives its name from several small islands that formerly existed in the bay, all but one of which have been connected to the shore or obliterated by harbour works and jetties. After 1962, when Algeria became independent, the capital city underwent a considerable shift in population. About half the inhabitants of European stock left. At the same time a heavy migration of non-Europeans from the interior occurred. As a result, the total number of inhabitants of greater Algiers increased from 885,365 in 1960 to 943,551 in 1966. In 1970 the inhabitants of French descent were estimated at only 31,000, other non-native stocks at 100,000, and the total population, including the seven suburbs of Bouzaréa, el-Biar, Hussein Dey, Bologhine (formerly Saint Eugène), el-Harrach (formerly Maison-Carrée), Kouba, and Birmandreis, was over 1,100,000.

Population

History. Algiers was founded in ancient times by the Phoenicians as one of their numerous North African colonies. It was known to the Carthaginians and the Romans as Icosium. Destroyed by a series of invasions, it was revived under a Berber dynasty in the 10th century as a centre of commerce in the Mediterranean. In the 16th century it was held briefly by the Spanish. Under the Turks in the 17th century, it became a flourishing city of nearly 100,000 inhabitants, from which the much-feared Barbary pirates set out in their boats to prey upon commerce in the Mediterranean. The fortress of the Kasbah (Qaṣbah), built on rocky heights overlooking the bay, was the residence of the last two Turkish *deys*. It was in the Kasbah on April 30, 1827, that one of the *deys*, Ḥusayn, struck the French consul with a fly whisk, thereby providing the incident that led to the French conquest of Algeria. A claim by two Algerian citizens for payment for wheat delivered to France had been outstanding since

the end of the 18th century. Ḥusayn and the French consul had again been discussing that claim when an altercation arose and the incident followed. To avenge the *dey*'s insult to France, the government of Charles X instituted a naval blockade, which proved unavailing. France then sent a naval and military expedition, taking Algiers in 1830 and making it a military and administrative headquarters for its colonial empire.

The contemporary city. Around the old Muslim town with its narrow, winding streets, a new city has grown up squeezed between the hills and the bay. The main thoroughfares run north and south, paralleling the waterfront. Many of the old buildings have given way to modern apartment houses and office buildings in the southern European style, creating a white city that gleams upon the blue waters of the bay. A beautiful Ḥanafī mosque remains from the 17th century and a Mālikī mosque from the 11th. West of the main square, a short street leads to the mosque of Ketchaoua (which dates back to Turkish times, and which was consecrated as a cathedral during the French interregnum) and to the ruins of the Jénina (Arabic Janīnah) or palace of the *dey*. Nearby is the national library, part of which is housed in a fine, old Moorish palace of 1798.

The city's location between the Mediterranean and the desert gives it a highly seasonal climate. Winters are cool and wet, summers hot and dry. Average temperatures in January and February are about 55° F (13° C) and in August 79° F (26° C). Rainfall varies from about five inches in December to almost none in the summer months.

During World War II Algiers became the headquarters of Allied forces in North Africa and for a time the provisional capital of France. In the 1950s, when the Algerian rising against France began, the capital city was a focal point in the struggle.

The city after independence

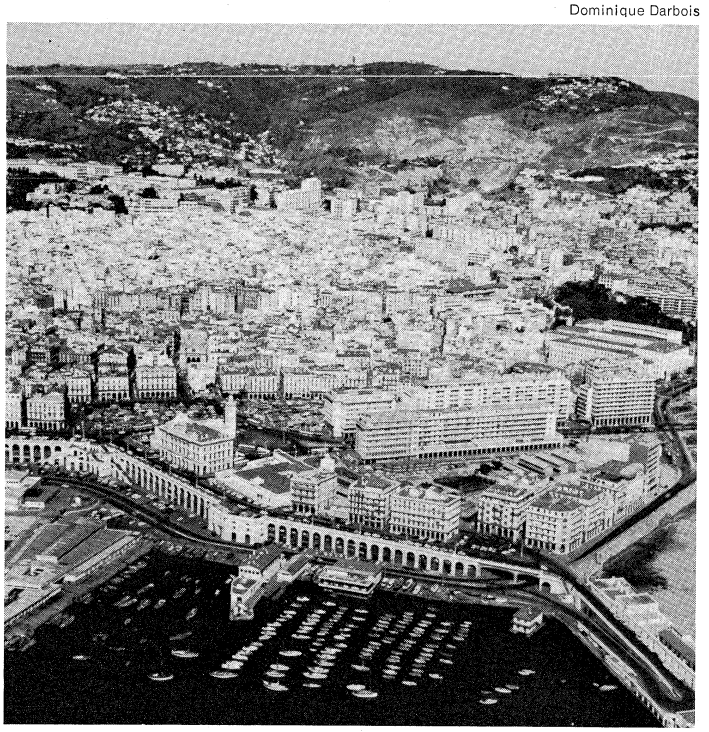
Independence brought far-reaching changes to the city as the new government set out to create a modern Socialist society out of a backward colonial one. The educational system was reorganized. In 1970 Algiers had 286 primary schools and a total of 146 higher schools, including normal schools and technical schools. Only a little more than half of the population was being educated, however, although the government intended to increase the proportion in the future. The University of Algiers had about 10,000 students in 1968. Of this number 15 percent were foreigners, coming primarily from African countries.

Two daily newspapers were published in Algiers in the early 1970s, one in French and the other in Arabic. There were several weeklies and an army paper that is published in both Arabic and French. There was one television station, and the telephone system had about 50,000 subscribers.

Algiers had once been an important French port, ranking second to Marseille in tonnage cleared. After the departure of the French it remained the busiest Algerian port. In 1969 slightly over 2,000,000 tons of merchandise were exported, about the same as in 1950; but over 3,500,000 tons were imported, as against 1,500,000 tons in 1950. The main exports are wine, early vegetables and oranges, iron ore, and phosphates.

The airport, Dar el-Beïda, is about 12 miles east of the city.

Dominique Darbois



Bay of Algiers and the city of Algiers on the slopes of the Sahel Hills.

BIBLIOGRAPHY. PIERRE BOYER, *La Vie quotidienne à Alger* (1964), on daily life in Algiers in the 18th and 19th centuries; PAUL CROMBET, "Alger au temps des Turcs," *La Revue de Paris*, 65:80–87 (1958); PIERRE FLAMENT, *La Casbah d'Alger* (1957); ROBERT GOOSSENS, *Alger* (1969), in French.

(A.K.Ch.)

Algirdas

Algirdas, grand duke of Lithuania from 1345 to his death in 1377, was mainly responsible for making his country one of the largest European states of his day.

Early years

Neither the date nor the place of his birth is known. One of the sons of the country's ruler, Gediminas, Algirdas began his long political career when he married, at his father's request, the daughter of the Prince of Vitebsk and succeeded to the Prince's lands. Together with his father he fought against the Teutonic Knights and attempted to subjugate the Russian territories of Novgorod and Pskov, which had sought Lithuanian protection both from the knights and from the Golden Horde (the Tatar overlords of Russia since the middle of the 13th century). From 1341 to 1345 he was prince of Krevo and Vitebsk and a vassal of his younger brother, Grand Duke Jaunutis, whom he removed in collusion with another brother, Kęstutis, with whose consent he became grand duke.

Defense of Lithuanian Ponemune and Podvine against the knights and their allies, invasion of the Russian and Ukrainian lands subject to the Horde, and the desire to achieve Lithuanian hegemony in the province of Volhynia were the goals of Algirdas' foreign policy. In the pursuit of these aims he relied on dynastic support and especially on his co-ruler Kęstutis. The brothers shared both their losses and their many acquisitions of fortified posts in Russia. They were supported by the princes and boyars who sat in the grand-ducal council and along with their feudal-dependent peasants, took the field under them.

Struggle with the Teutonic Knights

But though he was the leader of the Lithuanian–Slavic armies against the Teutonic Knights, Algirdas was a stranger to the lower social orders. When, during the anti-German "Uprising of the Night of Yury" in Livonia (1345), one of its leaders, a peasant, told him that he had been chosen king by the rebels and that if Algirdas followed his advice the Germans would be driven out, Algirdas had him decapitated. To the feudal prince, a peasant as king seemed a more terrible threat than the German usurpers.

Observing the ancient courtesies, Algirdas would entertain the lesser nobles at their periodic renewal of their compact with him as their overlord, but he would not tolerate an insult and in 1346 declared war on Novgorod after being scolded by the local governor at a meeting of the municipal assembly. A courageous warrior and man of strong sentiments, Algirdas, when watching his nobles besiege Kaunas but unable to come to their aid, was not above bursting into "great tears and lamentations." Although pagan to the core, he allowed his Orthodox subjects of Vilnius to build a church where formerly the gallows had stood. For political reasons he appointed his many Orthodox vice-regents in the Slavic territories of Lithuania, consistently married Orthodox princesses him-

self, and prevailed on the patriarch of Constantinople to found a Lithuanian Orthodox metropolitan see in the city of Kiev.

Algirdas saw far beyond the boundaries of his country. When the Polish king Casimir III the Great, Pope Clement VI, and Emperor Charles IV proposed to him that he accept Catholicism, he replied (1358) that he was ready to do so if they returned to him the lands between the Pregolya and Daugava rivers, liquidated the Teutonic Knights, and left him the

> empty lands between the Tatars and the Russians for their protection from the Tatars, leaving the Knights no rights whatever over the Russians but instead granting all Rus [Russia] to the Lithuanians.

But Algirdas' goals were not destined to be realized.

Early in his reign the Teutonic Knights and their allies conducted annual raids from their bases in Prussia and Livonia, ruining Lithuanian lands and subjugating White Russia as far south as Grodno. Aided by their supporters to the east and south, Algirdas and Kęstutis repelled these attacks. Yet, despite the expenditure of so much energy, Algirdas left the resolution of this historic struggle with the Teutonic Knights to his heirs.

Conflicts with Poland, the Golden Horde, and Russia

The rivalry with Poland over Volhynia was intensified when, in 1349, Casimir tricked Algirdas' brother Lubart out of one of the principal Volhynian cities. By the treaty of 1352 Lithuania took possession of Volhynia, but in 1366, preoccupied with his struggle with the Teutonic Knights, Algirdas had to yield most of Volhynia once again to his ally Casimir. After Casimir's death in 1370, however, he managed to recover part of the province by a treaty (1377) with King Louis of Hungary–Poland.

Algirdas' relations with Russia were characterized by his unsuccessful attempts to claim Pskov and Novgorod. Having met with Muscovite resistance, he concluded peace with Grand Duke Simeon of Moscow (1349). But with the decline of the Golden Horde after the death of the khan Jani Beg (Dzhanibek) in 1357, he extended his influence eastward approximately as far as Mstislavl' and Btyansk. In 1362–63 he campaigned in the territories of the Tatars, defeating three of their governors in the battle at the River Siniye Vody. He secured the principality of Kiev, which he gave to his son Vladimir, and freed little Podolia from the power of the Golden Horde.

In 1349 Algirdas married Yuliana, daughter of the Prince of Tver. Together with Tver and Smolensk he undertook three campaigns against Moscow (1368, 1370, 1372). They were, however, unsuccessful because of an increase in Moscow's prestige among the other Slavic lands.

Algirdas died in the middle of a war with the Knights. He was apparently cremated, along with 18 of his warhorses and other effects. He left his lands to his 12 sons. According to a contemporary chronicle, he

> drank not beer nor mead, nor wine nor fermented kvass. He was temperate and thus found wisdom. And by his cunning he conquered many lands and countries, subjugated many towns and principalities, and achieved great power.

(V.T.P.)

ʿAlī

ʿAlī ibn Abī Ṭālib, son-in-law of Muḥammad, the prophet of Islām, father of Muḥammad's two grandsons Ḥasan and Ḥusayn by Muḥammad's daughter Fāṭimah, reigned as the fourth caliph (successor of Muḥammad) from 656 to 661. Disagreement over ʿAlī's right to the caliphate resulted in the only major split in Islām—into Sunnites and Shīʿites. ʿAlī is revered by the Shīʿites—from Shīʿah, meaning "the party" (of ʿAlī)—as the only true successor of the Prophet.

Since the conflicts in which he was involved were perpetuated in polemical sectarian historiography, biographical material regarding ʿAlī is often biassed. There is agreement, however, that he was a profoundly religious man, devoted to the cause of Islām and the preservation of the rule of justice in accordance with the Qur'ān and the *sunnah* (the sayings and deeds of Muḥammad).

ʿAlī (meaning "high") was born in Mecca about the year 600, the son of Abū Ṭālib, chief of a local clan. When his father became impoverished, ʿAlī was adopted by Muḥammad, then still a businessman in Mecca, who himself had been cared for by ʿAlī's father as a child. When Muḥammad felt God's call to become his prophet, ʿAlī, though only ten years old, became one of the first converts to Islām and remained a lifelong devoted follower of Muḥammad. According to legend ʿAlī risked his life by sleeping in the Prophet's bed to impersonate him the night that Muḥammad fled in 622 from Mecca to Medina from enemies who were plotting to assassinate him. In addition, ʿAlī is said to have carried out Muḥammad's request to restore all the properties that had been entrusted to him as a merchant to their owners in Mecca. Only then did ʿAlī himself leave for Medina. There he married Muḥammad's daughter Fāṭimah, who bore him two sons, Ḥasan and Ḥusayn.

Adoption by Muḥammad

ʿAlī is said to have displayed rare courage in battle during the military expeditions Muḥammad conducted to consolidate Islām and always obtained a lion's share of the booty. ʿAlī was also one of Muḥammad's scribes and was chosen to lead several important missions. When the hostile inhabitants of Mecca finally accepted Islām without battle, it was ʿAlī who smashed the pagan idols in the Kaʿbah (holy shrine).

Muḥammad died on June 8, 632. Some say he had unequivocally nominated ʿAlī as his successor while he was returning from his "farewell pilgrimage" to Mecca. Others reject this claim, maintaining that Muḥammad died without naming a successor. ʿAlī, while attending the last rites of the Prophet, was confronted by the fact that Abū Bakr, Muḥammad's closest friend and father of ʿĀ'ishah, one of the Prophet's wives, had been chosen caliph. ʿAlī did not submit to Abū Bakr's authority for some time, but neither did he actively assert his own rights, possibly because he did not want to throw the Muslim community into bloody tribal strife. He retired into a quiet life in which religious works became his chief occupation. The first chronologically arranged version of the Qur'ān is attributed to him, and his excellent knowledge of the Qur'ān and the *sunnah* aided the caliphs in various legal problems.

Following the murder of ʿUthmān, the third caliph, ʿAlī was invited by the Muslims of Medina to accept the caliphate; reluctant, he agreed only after long hesitation. His brief reign was beset by difficulties due mostly to the corrupt state of affairs he inherited. Acutely aware of the neglect of the Qur'ān and the *sunnah* that his predecessors had allowed to develop, he based his rule on the Islāmic ideals of social justice and equality. His policy was a blow to the interests of the Quraysh aristocracy of Mecca who had grown rich in the wake of the Muslim conquests. In order to embarrass ʿAlī they demanded that he bring the murderers of ʿUthmān to trial, and when he rejected their request a rebellion against him was instigated in which two prominent Meccans along with ʿĀ'ishah, Muḥammad's widow and daughter of Abū Bakr, the first caliph, took a leading part. This rebellion, known as "the battle of the camel" (the camel ridden by ʿĀ'ishah), was quelled. A second rebellion was on the point of being crushed when its leader, Muʿāwiyah, a kinsman of ʿUthmān and the governor of Syria, averted defeat by proposing arbitration. ʿAlī saw through the stratagem but was forced by his army to accept adjudication, which greatly weakened his position. Soon, moreover, he had to fight the very people who had earlier forced him to accept arbitration but now denounced it. Known as Khārijites ("seceders"), they were defeated by ʿAlī in the battle of Nahrawān. Meanwhile, Muʿāwiyah followed an aggressive policy, and by the end of 660 ʿAlī had lost control of Egypt and of the Hejaz. Early one morning while praying in a mosque at Kūfah in Iraq, ʿAlī was struck with a poisoned sword by a Khārijite, intent on avenging the men slain at Nahrawān. Two days later ʿAlī died and was buried near Kūfah.

Acceptance of the caliphate

ʿAlī's political discourses, sermons, letters, and sayings, collected by ash-Sharīf ar-Raḍī (died 1015) in a book entitled *Nahj al-balāghah* ("The Road of Eloquence") and later commented upon by Ibn Abī al-Ḥadīd (died 1258), are well known in Arabic literature.

BIBLIOGRAPHY. An article on 'Alī appears in both the old and new editions of the *Encyclopaedia of Islam.*

(I.K.P.)

Alienation

Despite its popularity in the analysis of contemporary life, the idea of alienation remains an ambiguous concept with elusive meanings, the following variants being most common: (1) *powerlessness*, the feeling that one's destiny is not under one's own control but is determined by external agents, fate, luck, or institutional arrangements; (2) *meaninglessness*, referring either to the lack of comprehensibility or consistent meaning in any domain of action (such as world affairs or interpersonal relations) or to a generalized sense of purposelessness in life; (3) *normlessness*, the lack of commitment to shared social prescriptions for behaviour (hence widespread deviance, distrust, unrestrained individual competition, and the like); (4) *cultural estrangement*, the sense of removal from established values in society (as, for example, in the intellectual or student rebellions against conventional institutions); (5) *social isolation*, the sense of loneliness or exclusion in social relations (as, for example, among minority group members); and (6) *self-estrangement*, perhaps the most difficult to define and in a sense the master theme, the understanding that in one way or another the individual is out of touch with himself.

Recognition of the concept in Western thought has been similarly elusive. Although entries on "alienation" did not appear in major reference books of the social sciences until as late as 1935, the concept had existed implicitly or explicitly in classical sociological works of the 19th and early 20th centuries (by Marx, Durkheim, Tönnies, Weber, and Simmel). This rereading, as it were, of recent intellectual history is in no small part a result of the discovery (in 1932) of the early philosophic writings of Karl Marx (*Economic and Philosophic Manuscripts of 1844*), in which alienated labour is described. Thus, it is to Marx that are due both an early treatment of alienation and the recent reinvigoration of efforts to apply the idea to contemporary institutions.

History. The roots of the idea of alienation can be found in the work of Plotinus and in the theology of St. Augustine and Martin Luther (expressed, for example, in the struggle to dissociate—or alienate—oneself from one's own imperfections by identifying with a transcendental perfect Being). The most powerful pre-Marxian source, however, was Hegel, from whom Marx took the word alienation while at the same time secularizing the conception of it. For Hegel, alienation was an ontological fact, rooted in the nature of man's existence in the world. There was an inherent dissociation between man as subject and man as object (that is, between man as a creative subject seeking to be and to realize himself and man as an object influenced and manipulated by others), so that man's own creations (his art, language, science, and so forth) stand outside him as alien objects—as objectifications of what is essential and prior, namely, mind and individual consciousness.

This view, rooted in German idealism and in a metaphysical perspective, became the basis for a radical critique of society and its institutions by those philosophers who received Hegel's ideas and gave them a socialistic interpretation. Responding to the social crises of the 1840s, the German philosopher Ludwig Feuerbach (1804–72) and Marx in particular transformed alienation into a secular and materialistic idea. The evil that was alienation was no longer abstract and inescapable, for alienation was a product of specific forms of social organization. For Feuerbach, the source of alienation lay in the institution of religion: the myths of divine power were merely ways in which man projected his own humanity outside himself, locating his own capacities and sensibilities elsewhere. Overcoming alienation required a humanistic religion of man not of God.

It was Marx who then completed the secularization process. For him, the alienation of man from his true being was grasped by neither Hegel's existential duality nor Feuerbach's religious projection. It was man's nature, Marx held, to realize himself in work, but the possibility of doing so was denied to him by the economic system. Thus, the key problem was alienated labour under capitalism: work was compelled rather than spontaneous and creative; workers had little control over the work process; the product of labour was expropriated by others to be used against the worker; and the worker himself became a commodity in the labour market. Alienation consisted of the fact that man did not fulfill his "species being" in work; the essence of man remained unrealized.

Non-Marxian traditions

The Marxian tradition, however, represents only one stream of thought concerning alienation. A second stream, considerably less sanguine about the prospects for de-alienation, is embodied in what has come to be called the theory of "mass society." Observing the dislocations brought about by industrialization in the 19th and early 20th centuries, Émile Durkheim and Ferdinand Tönnies—and eventually Max Weber and Georg Simmel as well—each, in his own way, lamented the passing of traditional society and the consequent loss of the sense of community. Modern man was isolated as he had never been before—anonymous and impersonal in an urbanizing mass, uprooted from old values, yet without faith in the new rational and bureaucratic order. Perhaps the clearest expression of this theme is contained in Durkheim's notion of "anomie" (from Greek *anomia*, "lawlessness"), a social condition characterized by rampant individualism and the disintegration of binding social norms. Both Weber and Simmel carried the Durkheimian theme further. Weber emphasized the fundamental drift toward rationalization and formalization in social organization; personal relations became fewer, and impersonal bureaucracy became larger. Simmel emphasized the tension in social life between the subjective and personal, on the one hand, and the increasingly objective and anonymous, on the other.

Whereas both Marx and the mass-society theorists secularized alienation by focussing on the mundane world of work and social organization, a third stream of influence retained the existential emphasis of Hegel. Thus, Søren Kierkegaard, stressing the importance of subjective experience over objective knowledge, held that attaining an adequate sense of self in a world dominated by purposelessness and despair was the central problem of alienation. The same theme, without the religious solution offered by Kierkegaard, continues into the recent work of Martin Heidegger, Jean-Paul Sartre, and other existentialist philosophers.

Contemporary definitions. The classification of modern definitions given in the first paragraph of this article —powerlessness, meaninglessness, normlessness, cultural estrangement, social isolation, and self-estrangement— can serve only as a rough guide because contemporary writers often develop radically different conceptions within any one of the categories. Thus, with respect to self-estrangement, one can be "out of touch" with oneself in several quite different ways:

1. Borrowing heavily from Marx, the psychoanalyst Erich Fromm and the philosopher Herbert Marcuse indict modern civilization for failing to realize the creative potential of human nature. In their view, man in the capitalistic marketplace has become estranged from himself; he conducts himself according to ceremonial gestures that are detached or "alienated" from any real inner feelings or capabilities that he may have. "Sales personality" and "false consciousness" become merely instrumental, not creative. Such, according to this neo-Marxist interpretation, is the distortion of character imposed under capitalism. Perhaps the most profound form of alienation, in this view, is the estrangement of those who have adjusted themselves to alienated life styles without being aware of the distortion.

2. For Sigmund Freud and his followers, the self-estrangement is seen to lie in the split between conscious and unconscious forces in the personality. The individual is out of touch in the sense that repressed and unacknowledged desires motivate his behaviour. (The extreme form

of this dissociation is manifested in the common psychiatric disorders to which the French, before Freud, had attached the name *aliéne*.)

3. For Sartre and other existentialists, self-estrangement is a natural state of affairs in a world bereaved of meaning and purpose. Irremediable by either Marxian social innovation or Freudian clinical understanding, the absurdity of existence is held to be fundamental. To be "authentic" is to have a lucid consciousness that life is indeed meaningless and yet commit oneself through active choice (such as choosing to resist Nazi rule or to remain identified as a Jew under the humiliation of anti-Semitism).

4. Still others seek to limit the scope of self-estrangement to more manageable proportions while still retaining the insights of Marx and Freud. Thus, being out of touch may be read as a discrepancy between the real self and the ideal self (as, for example, the self-hatred among minority members who have accepted the majority's negative definition of them). It may also be read to mean that the individual is out of touch with his inner self because he is not truly involved in the task at hand: the task is not intrinsically rewarding but is a mere means to future or extrinsic rewards (salary, promotion, or the like). This view borrows heavily from Marx's depiction of the worker engaged in unfulfilling labour, though now the estrangement lies not in the split between human nature and the production system but in the rupture between required activity and subjective experience. This conception applies to all activities (such as housewifely duties, social conversation, and classroom studies) to the degree that they are not experienced as intrinsic rewards.

In these terms, sociologist David Riesman's "other-directed" type of personality represents a portrait of self-estrangement. In Riesman's formulation, the personality of man historically has proceeded from "tradition-directed" to "inner-directed" to "outer-directed." The tradition-directed type, characteristic of primitive and folk cultures, looks to tradition for models of behaviour. The inner-directed type, exemplified by 19th-century bourgeois man, was guided by personal aspirations and moral ideals that family authority had imbued in him from childhood. Modern outer-directed man, on the other hand, though just as success oriented as inner-directed man, gets his cues for behaviour not from an inner voice but from his peers—from their approval or disapproval. He thus conforms to a source "alien" to or outside himself.

Contemporary writers differ not only in their definitions of alienation but also in the assumptions that underlie these definitions. Two such contrasting assumptions are the normative and the subjective. First, those who hold most closely to the tradition of Marx (Marcuse and Fromm in the U.S., for example, or Georges Friedmann and Henri Lefebvre in France) treat alienation as a normative concept: as an instrument for criticizing the established state of affairs in the light of some standard based on human nature, "natural law," or moral principle. Marxian theorists, secondly, insist upon alienation as an objective condition quite independent of individual consciousness—hence, to be a "happy robot" at work is to be alienated irrespective of one's acceptance of the work experience. Alternatively, some writers—most commonly the American empiricists—emphasize that alienation is a social-psychological fact—it is the *experience* of powerlessness, the *sense* of estrangement. Such an assumption is often found in analyses and descriptions of deviant behaviour (and in the work of such theoreticians as Robert K. Merton and Talcott Parsons).

Causes of alienation. There are at least five causal theories concerning alienation: the economic, technological, sociological, philosophic-existential, and psychological. The economic theory is identified with Marx, for whom the epitome of evil was the property system and especially private ownership of the means of production. This economic root established the basis for the expropriation of the worker, both as a person and as a producer, and led inevitably to the alienation of man from man. It is an ironic fact of the concept's history, however, that the early philosophical Marx should be the one to whom we owe this account of alienation; in his later work (for instance, *Capital*) the idea of alienation disappears, as such, to be replaced by the economist's detailed analysis of surplus value, exploitation, class relations, and the like. To what extent Marx, in fact, abandoned the idea of alienation, or simply transformed his interest in it, remains a debatable issue.

The role of technology

Those who emphasize technological sources of alienation (for example, philosophers Lewis Mumford in the U.S. or Jacques Ellul in France) are concerned not simply with the direct impact of automation but basically with the ascendance of the technological spirit. The argument is that man becomes alienated by being forced to adjust his life style to the machine.

The dominating sociological thesis today is that alienation is the middle term in a formulation that reads: the decline of the local community based on tradition and the corresponding emergence of a secular society based on large-scale and mass action leads to a sense of alienation (powerlessness, isolation, and so on), which in turn finds expression in a wide variety of disorders—including political apathy, intergroup hostility, and volatile social movements seeking direct influence on the political process. The inference that alienation has increased in the modern era is difficult to document and is probably based on a too romantic idea of the premodern community.

The philosophic-existential theories—either in the religious version associated with such men as Martin Buber and Paul Tillich or in the secular form of Sartre and Albert Camus—emphasize that alienation is inherent in the finite and isolated character of man's existence as a "stranger" in the world.

On the psychological side, two basic elements in Freudian theory combine to form an explanation of alienation: first, there is the principle of Oedipal conflict, denoting a boy's unconscious hostility toward his father whom he feels to be a rival for the affections of his mother (or the analogous conflict involving a girl and her parents); second, there is the frustration inherent in civilized society (expressed most clearly in Freud's pessimistic *Civilization and its Discontents*). Both of these strands are incorporated, for example, in psychologist Kenneth Keniston's analysis of the "uncommitted" or culturally estranged college student, for whom there is an Oedipal drama in which the father is rejected for his inadequacy (his weak character and false values) and to whom the surrounding culture seems equally frustrating, fragmented, and inadequate.

Manifestations of alienation. It is one of the weaknesses, not strengths, of the concept of alienation that it is invoked to explain so much—everything, indeed, from the rise of fascism to the fall of fluoridation proposals. In the case of fluoridation (and similar issues) the argument is that such electoral defeats are produced by people who, feeling generally powerless, take the opportunity of expressing their underlying resentment against political authority by voting no. In the case of totalitarianism, fascist or otherwise, it is held that the readiness to seek a secular messiah and submit to his charismatic leadership derives from the fact that the individual feels so lost and unfulfilled in mass society.

Because the list of the supposed consequences of alienation is so extensive (psychosis, suicide, delinquency, psychosomatic disorder, wildcat strikes, ethnic prejudice, civil riots, and so forth) and so often contradictory (political passivity, on the one hand, and protest, on the other), it is important that matters be organized by specifying the arenas, the conditions, and the mechanisms involved in producing such manifestations of alienation. The task is not an easy one. Empirical studies of alienation have only recently been available, and cross-cultural investigations of the phenomenon are especially rare. Still, with respect to *arenas*, it can be noted that the symptoms of alienation occur in the physiological, the psychological-behavioural, and the sociological realm—respectively, for example, in psychosomatic illness, neurotic depres-

sion or suicide, and student protest movements. With respect to *conditions*, the main principle to be recognized is that alienation is not a free-floating element or force but operates, rather, within the limits of social structure—which is to say that the emergence of protest movements (or the failure of fluoridation proposals) has at least as much to do with the level and kind of social control applied by authorities (or the political structure of the community) as it does with the degree of alienation. Finally, one of four major *mechanisms* may generally be invoked to make predictions about the manifestations of alienation. Alienation either (1) turns inward on the self (hence, the development of physiological stress symptoms, suicide, or the like); (2) generates the need for substitute satisfactions (hence, the search for "community" in social movements or the emphasis on leisure pursuits among alienated workers); (3) generalizes out of the domain of its origin into other spheres of life (for example, the individual's sense of low control at work may be expressed also as powerlessness in politics, family life, and so on); or (4) operates as a functional agent in the individual's behavioural economy (for instance, it causes those who feel powerless to avoid politically relevant information because in their world view knowledge is not power, since control is impossible).

Cross-cultural comparisons

There is no reason to suppose that these mechanisms operate differently in different societies; hence the same general principles should apply to predictions made cross culturally. There is, however, reason to suppose that the varying structure of different societies would (1) generate different degrees and types of alienation and (2) affect the manner in which alienation of one kind or another is behaviourally expressed. Thus, if membership in a work organization, which presumably mediates between the corporation and the individual, diminishes the worker's sense of powerlessness (as mass theory would argue and as has been demonstrated to be the case in the United States), one would expect the same principle to hold in Sweden and in France (as, again, has been demonstrated). The principle may hold even though the absolute level of powerlessness, as well as the moderating effect produced by organization, might be quite different in view of the fact that labour organization plays a different role in Sweden and France than in the U.S. (in the former, it is more closely tied to national politics, and in the latter, it is weaker in contract negotiation). Similarly, one might expect to find some countries (such as Sweden) in which the problem of alienation is more a matter of "social isolation" than one of "powerlessness," though cross-cultural differences in types of alienation are difficult to demonstrate. International interest in the problem of alienation (including research on the functioning of workers' councils in such socialist countries as Yugoslavia and the U.S.S.R.) should guarantee the availability of sounder cross-cultural generalizations within the foreseeable future.

Overcoming alienation. It is a striking fact that the literature on the subject rarely provides a detailed portrait of unalienated man; it appears easier to recount contemporary troubles than to describe some healthy states or prescribe remedies for de-alienation. The prescriptions offered depend, naturally enough, on the diagnosis; the deeper the roots of alienation, the more radical the required transformation.

For those who follow in the Marxian tradition, the necessity for economic change is central because they hold that socialism will not only alter the work relationship and the production process but will also sustain the necessary moral climate and thoughtways (including attitudes about consumption and international relations). And if the economic revolution is not easily achievable at the national level, there are instructive models to be found in smaller utopian and communitarian societies where the aim is to encourage active and shared participation in meaningful work (as in the work communities of Europe or the kibbutzim in Israel). Barring such developments, there are still opportunities for change in the work process—such as programs of "job enlargement" or job rotation, based on the idea that the evils of specialization and monotony can be significantly attenuated if received notions of efficiency can be challenged.

If "creative work" is the primary solution for the Marxist, "commitment" is the key idea for the existentialist, and "awareness" for the Freudian. Insofar as a solution is possible (and in this respect non-Marxians tend to be considerably more pessimistic), the existentialist finds a solution in "action" and "being" rather than in "detachment" and "objectivity": man can transcend himself either by creating a meaningful existence out of his own freedom of choice or by courageously searching for an authentic experience of God. In the Freudian perspective, the solution is not merely a matter of bringing an individual's repressed strivings and unconscious needs to awareness through therapy but also a difficult matter of developing a less repressive society and culture.

Among the non-Marxian sociologists, there are few grand programs for amelioration; their common theme seems to be captured in the idea of "pluralism." Recognizing both overorganization and disorganization as ingredients of the problem, they encourage the development of community-based action groups whose counterbalance and interplay can provide the individual with a sense of serious engagement in policy determination and can provide society with instruments to combat both superorganization and anomic disintegration.

Though the alienation perspective recounted here reflects a melancholy view of modern life, there is in it a moral concern about the good life that is basically hopeful and visionary. And there is, as well, some recognition that alienation is not exclusively evil, providing as it sometimes does the basis for creative insight and the impetus for social change.

BIBLIOGRAPHY. KARL MARX, *Selected Writings in Sociology and Social Philosophy*, ed. by T.B. BOTTOMORE and M. RUBEL (1964), includes Marx's classic and crucial early work on alienation; SIGMUND FREUD, *Das Unbehagen in der Kultur* (1930; Eng. trans., *Civilization and Its Discontents*, 1958), the clash between the conflictful individual and repressive society is elaborated in this late and classic work; HERBERT MARCUSE, *Reason and Revolution: Hegel and the Rise of Social Theory* (1941, 2nd ed. reissued 1960), an authoritative treatment of the Marxian view in its relation to Hegel, Feuerbach, Kierkegaard, and others; ERICH FROMM, *The Sane Society* (1955, reissued 1965), a recent classic joining the Freudian and Marxian perspectives on alienation; R.A. NISBET, *The Sociological Tradition* (1966), reviews the basic ideas in classical sociological thought, including a chapter on alienation in Tocqueville, Marx, Durkheim, Weber, and Simmel; F.H. HEINEMANN, *Existentialism and the Modern Predicament* (1953, 2nd ed. reissued 1958), a thorough and readable review of the existentialists from Kierkegaard to Sartre; E. AND M. JOSEPHSON (eds.), *Man Alone: Alienation in Modern Society* (1962), a collection of empirical and theoretical papers on the many manifestations of alienation.

(M.S.)

Alismales

The Alismales, or water-plantain order, is a group of flowering plants belonging to the monocots (one of the two large classes of flowering plants) and containing three families of aquatic plants that grow in freshwater, either entirely submerged or partly exposed to the air. Although widely distributed, this group of plants is small in both number of species (about 100 in 18 genera) and population. They are of little economic value but hold considerable botanic interest through their relationships with the environment and with other aquatic plants. In their role as temperate-climate and tropical weeds, they are important factors in conservation, navigation, irrigation, and fisheries.

General features. *Importance to man.* Their nutritional status in the present-day world is insignificant, but the underground storage organs are swollen with food reserves, and this carbohydrate was once utilized by North American Indians and others. In the Orient, rhizomes (buried-stem structures) of *Sagittaria* have been eaten both cooked and fresh, and it is reported that Eskimos still make bread from dried and powdered rhi-

zomes of *Butomus*. In Java, juvenile plants of *Limnocharis flava* are highly esteemed as a vegetable, and this species, together with *Sagittaria trifolia*, is also collected and used as manure and for cattle and pig fodder in tropical Africa, India, and Southeast Asia. Although these plants have no medicinal importance today, in medieval times some species were considered important and were featured in early herbals.

Many plants of the Alismales are decorative in form and have been used as ornamental models; *Sagittaria trifolia*, for example, appears as a motif in Japanese religious metalwork of the 16th century. Ornamental species are presently cultivated in formal pools, and *Sagittaria* and *Butomus* occur frequently in landscaped water gardens. Tropical species are much used in tropical-fish aquariums to both aesthetic and biological advantage.

Size range and diversity of structure. There is some variation in size among species, which range from five to 150 centimetres (two to 60 inches) in height, but none attains the extreme dimensions common in many plant orders. Members of all three families of the order are rather uniform anatomically. Surface hairs, for example, are absent in the order except for the large, simple, and branched ones on the undersurfaces of *Echinodorus macrophyllus* leaves. Stomata (small air pores) are present in the leaves of all species of the order except *Wiesneria schweinfurthii*. They usually occur in greater numbers on the upper surfaces of the leaves—markedly so in *Hydrocleys*, in which the upper side of the leaf is exposed to air and the lower side is submerged.

In all families there are species with massive leafstalks, which may be triangular, horseshoe-shaped, circular, or elliptical in cross section. Internal tissues are largely filled with air spaces, which occupy 80 percent of the total plant volume. Vascular tissue is universally present, although the xylem, the water-conducting tissue, is primitive and may be nearly absent in submerged species. The phloem, or food-conducting component, is more marked. The number of vascular bundles in any plant organ can be correlated with its size and degree of elaboration. In the Alismales order they are not randomly scattered, in contrast to the condition prevailing in most monocots, except in the flower stems of *Butomus* and *Limnophyton*.

Arrangement of vascular tissues

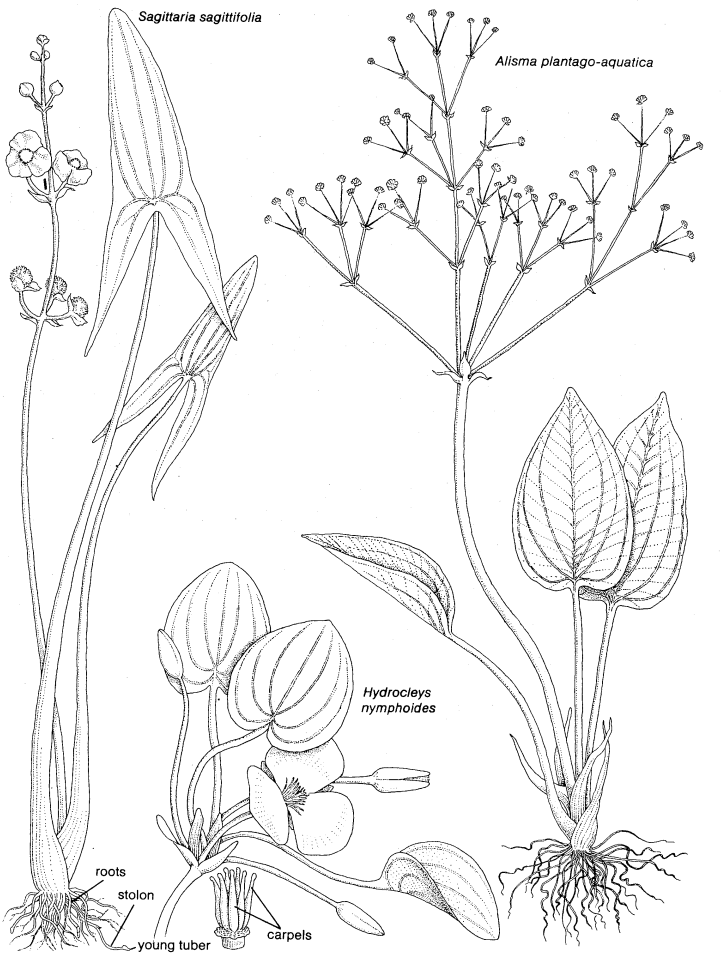
The most frequent arrangement is in arcs or circles. In every species, the first-formed xylem (protoxylem) disintegrates, forming conspicuous canals. Conducting elements in all parts of the plant except the root consist of tracheids, long, thin-walled cells that are primitive in type. Protoxylem tracheids may have ringlike or spiral cell-wall thickening, but later formed elements, called metaxylem, may have more complex wall sculpturing. The true vessel, a more advanced type of conducting cell, is confined to the root, where it is constantly present.

In the families Alismaceae and Limnocharitaceae, metaxylem elements of the leaf are arranged in a crescent-shaped arc surrounding the protoxylem canal, with phloem similarly disposed around the metaxylem. The vascular strand is V-, U-, or C-shaped in section, open to the inside facing the main axis of the plant. In the genus *Butomus*, the metaxylem is inverted, enclosing the oval phloem group, with the V-arms facing outward.

Mechanical tissue is reduced in all members of the order, and there is a lack of cell-wall thickening in conducting tissue. Secretory canals exuding a type of latex are frequent in aerial parts of members of the Alismaceae and Limnocharitaceae families but are not present in *Butomus*. Crystals and tannin occur in the cells of *Butomus* and some members of the Alimaceae but are not found in the Limnocharitaceae.

Some species show more than one form of leaf structure. Stomata, for example, are absent from submerged strap-shaped leaves of *Sagittaria sagittifolia* but are present in floating leaves, less frequently in the aerial arrow-shaped leaf. Surface cells also differ in shape.

Root structure shows little variation among members of the order except in minor details such as number of vascular strands, and in general architecture it resembles that of other monocots and water plants. The outer layers contain many air spaces.



Typical plants and structures of the water-plantain order.
Drawing by M. Moran

Comparing different organs of the plant, *Hydrocleys* is structurally more homogeneous than other genera, probably because all its parts are submerged. Leafstalk, flower stem, and runners are similar in shape and diameter and are almost identical in organization of their tissues, differing only in number and distribution of vascular bundles. In other genera the histology of different plant organs is more varied.

Distribution and abundance. The order is broadly distributed throughout temperate and tropical regions and includes many widespread and familiar water plants. All three families are small in size, and one (Butomaceae) has only a single species. Populations are small and are frequently included in mixed communities. Common genera with larger numbers of species are widely spread; *e.g.*, *Alisma*, *Sagittaria*, and *Echinodorus*. *Sagittaria* and *Butomus*, less frequently found, are more local in distribution than *Alisma*. *Butomus*, native to England, where it is reasonably common, has been introduced in North America. *Alisma plantago-aquatica*, the water plantain, which gives its name to the family and order, is the most abundant member of the group.

Geographical range

As with other water plants, those of the Alismales order have a great geographical range and at the same time a high frequency of species confined to a particular region only, mostly in the tropics.

Botanical importance. Although commercially unimportant, the external and internal form and type of leaf of the Alismales order have been studied intensively from the earliest days of botanical research. This ancient group, with its key position in the taxonomic system, is significant in speculations about the relationships between the two major groups of flowering plants (monocots and dicots) and their past origin. The Alismales order forms a link in the evolutionary chain between them. Since the order is restricted to water habitats, its

status is heightened in concepts that stress aquatic ancestry of the monocots and their hypothetical origin from aquatic dicots or predecessors.

Natural history. Reproduction is both sexual and vegetative. The latter method is more variable in mechanism and possibly of greater significance in multiplication.

Sexual reproduction. In the reproductive phase, vascular water plants betray their land-based ancestry. The form and structure of their vegetative organs are modified in relation to their aquatic habitat, but the reproductive organs are remarkably similar to those of related land plants in both general organization and microscopic structure. The flowers are adapted to aerial life and are pollinated by insects or wind.

There is evidence that ability to reproduce is related to age, a minimum juvenile period elapsing before flowering takes place. In genera such as *Sagittaria*, with more than one type of leaf, flowers are typically associated with adult aerial foliage. In contrast, *Alisma gramineum* can flower when possessing only submerged foliage.

Inflorescences (flower clusters) arise from the junctions of stem and leaves, and vegetative growth of the axis is resumed by subsequent development of other lateral buds. Aerial flowers are mainly bisexual and are borne on erect or floating axes in whorls. The petals, though small, are significant and attractive, and the species of most genera are pollinated mainly by flies, with beetles and snails possibly playing a minor role.

The fruit develops aerially. In the Alismaceae family the separate female structures develop into a head of dry, one-seeded fruits. In the Butomaceae family the fruit is primitive, consisting of a cluster of pods that split open when mature.

Seeds of many species have a prolonged dormancy. In the single-seeded indehiscent achenes (dry, one-seeded fruits that do not split open along definite seams), the embryo is protected by a hard seed coat that is thick and tight fitting. Some fruits retain their viability even after being frozen in ice and mud for some weeks. The ovary walls of *Alisma* fruits contain large intercellular spaces providing buoyancy for water dispersal. Seeds of *Sagittaria* have lateral wings containing air-space tissue. These seeds float in water, and the period of flotation may last several days to two months. In *Baldellia* the fruits sink immediately, but the young seedlings rise to the surface and float for several weeks before rooting.

Seed dispersal

Asexual, or vegetative, reproduction. Aquatic plants resort to vegetative reproduction more frequently than land plants, although methods of propagation are fairly universal. The uniform environment and abundance of carbon dioxide and water encourage leaf development and sustain vegetative growth. Vegetative reproduction is facilitated by fragmentation of the plant body followed by regeneration from any small part bearing a bud.

Multiplication by means of modified stems such as rhizomes, stolons, and runners is widespread in temperate species. These specialized organs also serve the purpose of sustaining plants during adverse conditions. In hot climates vegetative growth may be more continuous. Vegetative parts that can grow into new plants may be transported by water or animals, further contributing to the distribution of a species.

There are many instances of pseudovivipary, wherein asexual reproductive units replace some or all of the normal sexual flowers in the inflorescence. This method is common in the family Alismaceae, in which the inflorescence may set seed above water but bear young plantlets when submerged.

In many genera there is an abbreviated axis from which arises the crown of leaves. This rootstock becomes swollen with starch grains and serves as a storage organ during the resting period when foliage dies down. Many plants with underground stems reproduce vegetatively and spread horizontally by means of runners at the soil-surface level or by similar underground stems. These become erect at intervals to form new plants with adventitious roots. Later, intermediate parts may disintegrate and the secondary plants become independent.

Species of *Sagittaria* remain alive during adverse seasons by the production of special terminal tubers behind the apical bud. These structures are two to five centimetres long (one to two inches) and may be brightly coloured blue and yellow. They remain dormant during the winter but sprout into a new plant the following spring.

Vegetative reproduction is of the utmost importance in the maintenance and spread of some species. In all three families of this order it is correlated with aquatic existence, and this perhaps accounts for its success and frequency in comparison with sexual reproduction that is not adapted to aquatic conditions.

Ecology. Members of these families, together with other species, participate in water-habitat plant succession but do not play a particularly dominant role at any stage. Plants with largely immersed shoots, such as *Sagittaria* and *Alisma*, mostly occur in silted waters just beyond or replacing reed swamp and in fairly shallow water. Both belong to a mixed aquatic community, and *Alisma* is a ready colonizer of fresh ponds, though not necessarily always present in the succession. *Sagittaria* occurs frequently in rivers with moderate current and may dominate submerged communities in its strap-leaved form. It may act as pioneer for the reed-swamp community and is frequent in the upper Thames and River Cam in England. *Alisma* also participates in reed swamp and, together with *Sagittaria* and *Butomus*, can form a shoreline reed swamp in muddy rivers that are subject to flooding.

Typical habitats

Temperature is a factor of great uniformity in an aquatic environment and may partly explain the vast geographical range and latitudinal extension of some species. The requirement for water itself is perhaps the most restrictive factor in the occurrence of these plants.

Form and function. *Floral characters.* Flower structure of the Alismales order includes several features held to be relatively unspecialized. Male and female structures are found in the same flower. The flower parts arise at the base of the ovary. The ovaries are separate, and there is an absence of fusion in other floral parts. The Alismaceae family shows a trend from crowded or spirally arranged female structures to a single whorl. The families of the order are considered to be advanced in having whorled flower parts in two circles of three petals and three sepals. *Sagittaria* shows an interesting tendency toward unisexual flowers that are male in the upper and female in the lower parts of the flower cluster. Flowers are aerial, borne erect on floating axes, and are similar to those of land plants in structure. Members of the family Alismaceae have nectar-producing structures between the flower parts.

Vegetative characters. The plants are nonwoody and persist through more than one growing season (*i.e.*, they are perennial). They are characterized by a structural uniformity that adapts them to an aquatic existence. Variations in leaf form and shape are associated with the degree of immersion in water, and some species have leaves of two or more forms. In many others it is possible to discern a gradation in leaf shape along the plant.

Butomus (Butomaceae) has a broad, horizontally creeping, underground stem with thick, fibrous roots from which arise tall leaves in two vertical series. Leaves are linear, erect, and triangular in section but are not differentiated into a separate leaf and blade. The leaf base is broadly sheathing (*i.e.*, it wraps around the stem), and the tip may be slightly flattened. The whole leaf is morphologically equivalent to the petiole (leafstalk) of other genera and is thus a phyllode (a flattened leafstalk that functions as a leaf blade).

Limnocharis (Limnocharitaceae) consists of a short rootstock with fine basal roots. The leafy shoot is erect and partly emergent. Leaves are massive and grow out of the water with the blade and upper stalk aerial. The sturdy stalk is three-angled, and the blade is inclined to the upright, varying in shape from broadly oval to elongated lance-shaped.

Hydrocleys (Limnocharitaceae) is a floating form with runners producing small plantlets at intervals. The leaf is submerged. Long, rounded leafstalks float sinuously, and the flattened leaf blades rest horizontally near or on

the water surface with their upper faces exposed to the atmosphere.

Leaf variation with water depth

In the Alismaceae family the main plant axis is similar throughout, with crowded, sheathing leaf bases arranged in two ranks. Leaf type and shape vary according to locality and depth of water, and there may be two or more forms of leaf, together with intermediates, on a single plant. Leaves of *Sagittaria* show the greatest degree of elaboration in the family. The leaf blade develops paired and pointed lateral lobes, giving rise to the typical arrowhead shape. Submerged leaves may be linear, transparent, and floating. All leaves have the parallel venation typical of the monocots.

Evolution. It has been suggested that the dicotyledon group that gave rise to monocotyledons must have had many features in common with the plants of the Alismales order. The Nymphaeales order is such a group and itself was probably derived from some ancient, vesselless stock of the Magnoliales order. Nymphaeales plants are aquatics without vessels and show tendencies towards cotyledon (seed-leaf) fusion. It is certainly feasible that an ancient stock of plants resembling the modern Nymphaeales order gave rise to the Alismales order.

Monocots are considered to be of aquatic ancestry, later evolving land forms that in turn secondarily gave rise to groups that have returned to the water. The Alismales order shows a progressive adaptation to an aquatic habitat, and the closely allied order Hydrocharitales has marine as well as freshwater species.

One of the fundamental differences between the two great groups of flowering plants, the monocots and dicots, is the difference in leaf-vein arrangement between them. The Alismales order is important in being positioned, in an evolutionary sense, near the bottom of the monocot sequence, and some of its members show leaf structure that can be viewed as showing the evolution of the parallel-veined monocot leaf from the dicot leafstalk. This interpretation is well supported by *Sagittaria*, which shows all the transitional stages between a normal leaf with blade and stalk and a linear leaf with parallel veins, all the stages being present within a single population. The occurrence of the various forms appears to depend on the water depth at which each is located. There is evidence that when present and developed, the parallel-veined leaf blade is derived from the secondarily expanded tip of the stalk, which becomes flattened and expanded through the evolutionary series. Thus, in this interpretation, the monocot and dicot leaf blades are not homologous.

Phylogenetic trends within the order are difficult to discern because they are masked by characters that are more readily correlated with ecological factors. *Butomus* has shorter and wider vessel elements in its roots than members of the Limnocharitaceae family, suggesting that the latter family is more primitive. With respect to the xylem water-conducting elements, dimensions, and many other characters, the Alismaceae family appears to be close to and more advanced than the Limnocharitaceae family. The Butomaceae family seems to have diverged from the other two families or some common ancestor of all three, becoming more specialized. There are indications of possible affinities with the Scheuchzeriaceae family of the order Potamogetonales (also known as Najadales).

The Alismales order, in common with other water plants with soft, vegetative organs, have been rarely preserved in fossil deposits. In the absence of woody parts, usually only resistant hard fruits, seeds, or pollen grains remain, the identification of which is uncertain.

Classification. *Annotated classification.*

ORDER ALISMALES

Aquatic herbs. Flowers superior (ovary positioned above other flower parts) with biseriate perianth usually differentiated into 3 sepals and 3 petals. Gynoecium (female component) of several or many free carpels (structures comprising the ovary). Placentation (attachment of ovules) laminar, ovules numerous and scattered. Mature seeds without endosperm, endosperm development helobial (later disappearing). Embryo with prominent basal cell and nucleus of high ploidy (multiples of the basic chromosome number) horseshoe-shaped or straight. Stamens numerous, pollen 3-celled, trinucleate, and multi-aperturate (except *Butomus*). Vegetative parts with laticiferous (latex-bearing) canals (except *Butomus*), paracytic (parallel-celled) stomata, specialized aerenchyma (cellular tissue with air spaces) with diaphragms, protoxylem canals, metaxylem tracheids, and restriction of vessels to roots.

Family Butomaceae

Leaves erect linear phyllodes, not differentiated into petiole (stalk) and lamina (blade). Inflorescence umbellate (with flower stalks radiating from a common point, as in an umbrella). Flowers hermaphrodite, numerous, emergent from water. Perianth 6 in 2 whorls, petaloid, persistent. Stamens 9, red. Anthers introrse (opening inward toward the flower axis). Pollen grain uni-aperturate. Carpels 6, red, free, with numerous ovules. Placentation laminar, ovules scattered over carpel walls. Fruit a follicle (a fruit consisting of a single carpel that splits open along the ventral suture). Embryo sac monosporic, *Polygonum* type. Embryo straight. Endosperm helobial. A single perennial genus and species, *Butomus umbellatus*, distributed throughout the North Temperate Zone.

Family Limnocharitaceae

Plants emergent or submerged. Leaves erect or floating, differentiated into petiole and lamina. Laticifers present. Sepals 3, green. Petals 3, delicate, deciduous (falling early). Carpels 3, 6, 9, or 15–20. Placentation laminar, ovules scattered. Pollen grains with 4 pores. Embryo sac bisporic *Allium* type. Embryo curved or folded, horseshoe-shaped. Four genera, 12 species, distributed in the tropics.

Family Alismaceae

Plants emergent or submerged. Leaves heterophyllous (with different forms), linear when submerged or laminate when emergent. Small axillary scales present. Laticifers present. Inflorescence branched, racemose or cymose. Flowers hermaphrodite or unisexual. Sepals 3. Petals 3. Stamens 6 to many or 3. Pollen multi-aperturate. Anthers extrorse (opening away from flower axis). Carpels 3 to many. Fruit a group of achenes. Ovules 1 or rarely 2, basal or on a marginal placenta. Embryo sac bisporic, *Allium* type. Embryo horseshoe-shaped. Endosperm helobial or nuclear. Thirteen or 14 genera and 90 species, mainly confined to the Northern Hemisphere but widely found in tropical and temperate regions.

Critical appraisal. These groups convey an overall impression of unity in form, structure, and habit. Within the families, the genera and species are distinguished by minor characters not always immediately apparent to the untrained eye. Some authorities isolate the Butomales and Alismales as the first two orders of the monocot sequence. In the Butomales order of those systems are two families, the Butomaceae and Hydrocharitaceae. Other taxonomists afford order status to the latter family as the Hydrocharitales, near to and derived from the Alismales order. The Hydrocharitaceae family can be separated from the three families of the Alismales order by its inferior ovary (located below the other flower parts), although there is support for the idea that the family may have descended from ancestors similar to the Butomaceae family.

The inclusion of the Limnocharitaceae family together with *Butomus* in one family, the Butomaceae, as is sometimes done, is not supported by the evidence. In the past, genera belonging to all three families have been combined under the Alismaceae family. It is clear now, however, that there is a great weight of anatomical, morphological, palynological (analysis of pollen types), embryological, and biochemical characters all splitting *Butomus* from the genera that can be included in the Limnocharitaceae family. This last group exhibits closer affinities with the family Alismaceae, *Limnocharis* with its large elaborate leaves particularly resembling *Alisma* and *Sagittaria*, whereas the submerged and reduced leaves of *Hydrocleys* are in some respects more reminiscent of members of the Hydrocharitaceae family.

BIBLIOGRAPHY. A. ARBER, *Water Plants: A Study of Aquatic Angiosperms* (1920), a classic study of water-habitat plants; C.D. SCULTHORPE, *The Biology of Aquatic Vascular Plants* (1967), an up-to-date and comprehensive monograph on water plants in general; M.Y. STANT, "Anatomy of the Alismataceae," *J. Linn. Soc.*, 59:1–42 (1964); and "Anatomy of the Butomaceae," *J. Linn. Soc.*, 60:31–60 (1967), the only comprehensive accounts on vegetative morphology and structure of members of the Alismales order.

(M.Y.S.)

Alkali Metals and Their Compounds

The alkali metals are the elements in Group Ia of the periodic table of the elements. In order of increasing atomic number they are lithium (Li), sodium (Na), potassium (K), rubidium (Rb), cesium (Cs), and francium (Fr). Sodium and potassium are the sixth and seventh most abundant of the elements, comprising respectively 2.6 and 2.4 percent of the earth's crust. The other alkali metals are considerably more rare, with rubidium, lithium, and cesium respectively forming 0.03, 0.007, and 0.0007 percent of the earth's crust. Francium, a natural radioactive isotope, is very rare and was not discovered until 1939. Because of their high reactivity, the alkali metals are never found as free metals in their natural state. They generally are found combined with other elements in the form of simple or complex compounds. The simpler compounds of the alkali metals are soluble in water and therefore are easily extracted and subjected to chemical operations for purposes of separation and purification. Minerals belonging to this class, such as halite (sodium chloride, NaCl), sylvite (potassium chloride, KCl), and carnallite (a potassium-magnesium chloride, $KCl \cdot MgCl_2 \cdot 6H_2O$), although somewhat rare, are the most important commercial sources of the alkali metals; the more complex, water-soluble minerals are far more abundant in the Earth's crust.

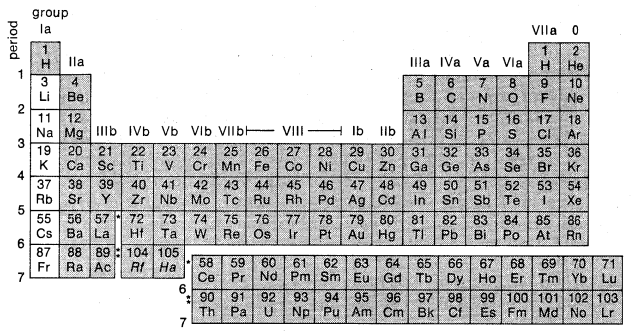


Figure 1: Alkali metals in the periodic table of elements.

The alkali metals have all of the physical properties generally associated with metals, including silver-like lustre, high ductility, and excellent conductivity of electricity and heat. Lithium is the lightest metallic element. The alkali metals are low melting, ranging from a high of 179° C for lithium to a low of 28.5° C for cesium. Alloys of alkali metals exist that melt as low as −78° C.

The alkali metals are extremely reactive and combine readily with most of the substances found in the atmosphere. (Only lithium, however, reacts with nitrogen.) The alkali metals all react vigorously, and often violently, with water, releasing hydrogen and forming strong caustic solutions. Most common nonmetallic substances such as the halogens, halogen acids, sulfur, and phosphorus react with the alkali metals. The alkali metals themselves react with many organic compounds, particularly those containing an active hydrogen atom (that is, one that can be replaced readily).

Sodium is by far the most important alkali metal in terms of industrial use. It is employed in the reduction of organic compounds, in the preparation of sodium peroxide and sodium cyanide, and in the manufacture of tetraethyl lead. As a free metal, it is used as a heat transfer fluid in nuclear reactors. Its compounds, of the widest industrial importance and manufactured in hundreds of thousands of tons annually, include common salt (NaCl), baking soda ($NaHCO_3$), soda ash (Na_2CO_3), and caustic soda (NaOH). Potassium has considerably less use than sodium as a free metal. Potassium salts, however, are consumed in considerable tonnages in the manufacture of fertilizers. Lithium metal is used in certain light-metal alloys, and it is employed as a reactant in organic syntheses. Lithium also has potential use as a battery anode in electrically propelled automobiles. Rubidium and cesium and their compounds have but limited use.

This article covers the history and the physical and chemical properties of the alkali metals and then treats in more detail the individual members of the group. Information on the technological aspects of sodium and potassium is available in the article SODIUM AND POTASSIUM PRODUCTS AND PRODUCTION. Further details of the relation of alkali metals to other chemical elements appear in the article PERIODIC LAW.

HISTORY

Alkali metal salts were known to the ancients. The Old Testament refers to a salt called *neter* (sodium carbonate), which was extracted from the ash of vegetable matter. Saltpetre (potassium nitrate) was used in gunpowder, which was introduced into Europe in about the 12th century. Saltpetre was also one of the chief ingredients of the incendiary mixture called Greek fire.

Saltpetre

In October 1807 the English chemist Sir Humphry Davy isolated potassium and then sodium, the discoveries coming only a few days apart. The term sodium is derived from the Italian *soda*, a term applied in the Middle Ages to all alkalies; potassium comes from the French *potasse*, a name used for the residue left in the evaporation of aqueous solutions derived from wood ashes.

Lithium was discovered by the Swedish chemist J.A. Arfvedson in 1817 while analyzing the mineral petalite. The name lithium is derived from *lithos*, the Greek word for stony. The element was not isolated in pure form until Davy produced a minute quantity by the electrolysis of lithium chloride.

While the German chemists Robert Bunsen and Gustav Kirchhoff were investigating the mineral waters in the Palatinate in 1860, they obtained a filtrate that was characterized by two lines in the blue region of its spectrum (the light emitted when the sample was inserted into a flame). They suggested the presence of a new alkali element and called it cesium, derived from the Latin *caesius*, used to designate the blue of the sky. The same researchers, on extracting the alkalies from the mineral lepidolite, separated another solution, which yielded two spectral lines of violet colour. They proposed the name rubidium for the element in this solution from the Latin *rubidus*, which was used for the darkest red colour. Francium was not discovered until 1939 by Marguerite Perey of the Curie Institute in Paris.

In the 19th century the only use for the alkali metals was the employment of sodium as a reagent in the manufacture of aluminum. When the electrolytic process for aluminum purification was established, it appeared that large-scale use of sodium would cease. Subsequent improvements in the electrolytic production of sodium, however, reduced the cost of this element to such an extent that it could be employed economically in processes such as the preparation of sodium peroxide. The chief improvement in the electrolytic process was the introduction of continuous electrolysis of sodium hydroxide, a technique called the Castner process. The Castner process was used until the evolution of the Downs-cell process in 1926. This process, in which a molten sodium chloride–calcium chloride mixture is electrolyzed, has largely supplanted the older process in the production of sodium.

Electrolytic production

GENERAL PROPERTIES

Physical properties. Metals are ordinarily characterized by their high thermal and electrical conductivity and by their lustre, ductility, and malleability. The alkali metals as a class exhibit these properties to a high degree. Chemically, each of the alkali metals has a valence of one (*i.e.*, one atom of an alkali metal combines with one atom of another univalent element, such as fluorine); this property is a consequence of their having only a single electron in the outermost shell of their atoms. In passing from lithium to francium, this single electron tends to be less strongly held. Generally, the energy necessary to remove the outermost electron from the atoms of an element, the ionization energy, decreases in the periodic table toward the left and downward in each vertical file, with the result that the most easily ionizable element in

Some Properties of the Alkali Metals

	lithium	sodium	potassium	rubidium	cesium	francium
Atomic number	3	11	19	37	55	87
Atomic weight (or stablest isotope)	6.94	22.99	39.1	85.47	132.91	(223)
Colour of element	silver	silver	silver	silver	silver	—
Melting point (°C)	179	97.8	63.65	38.89	28.5	—
Boiling point (°C)	1317	892	753.9	688	671	—
Density at 20° C (g/cm³)	0.534	0.97	0.862	1.53	1.89	—
Volume increase on melting (percent)	1.51	2.63	2.81	2.54	2.66	—
Valence	1	1	1	1	1	1
Electronic configuration	$1s^2 2s^1$	$1s^2 2s^2 2p^6 3s^1$	$(Ar)4s^1$	$(Kr)5s^1$	$(Xe)6s^1$	$(Rn)7s^1$
Isotopic abundance (terrestrial, percent)	^{7}Li (92.6), ^{6}Li (7.4)	^{23}Na (100)	^{39}K (93.1), ^{41}K (6.88), ^{40}K (0.0119)	^{85}Rb (72.15), ^{87}Rb (27.85)	^{133}Cs (100)	—
Colour imparted to flame	red	yellow	violet	dark red	blue	—
Main spectral emission lines (wavelength, Å)	6708, 6104	5890, 5896	7699, 7665	4216, 4202	4593, 4555	—
Heat of fusion (cal/mole)	690	622	598	540	520	—
Specific heat (liquid, cal/g °C)	1.05	0.33	0.188	0.0880	0.0572	—
Electrical resistivity at 0° C (microhm-cm)	8.55	4.2	6.15	2.5	20.0	—
Magnetic susceptibility (cgs unit)	14.2×10^{-6}	16.0×10^{-6}	20.8×10^{-6}	17.0×10^{-6}	29.0×10^{-6}	—
Crystal structure	BCC*	BCC*	BCC*	BCC*	BCC*	—
Radius						
Atomic (Å)	1.52	1.85	2.31	2.44	2.62	—
Ionic (M^+, Å)	0.68	0.97	1.33	1.47	1.67	1.80
Metallic (single bond, Å)	1.225	1.572	2.025	2.16	2.35	—
Ionization energy (first, kcal/mole)	124.3	118.4	100	96.3	89.7	—
Oxidation potential at 25° C ($M \rightarrow M^+ + e^-$, volts)	3.04	2.71	2.92	2.92	2.92	—
Electronegativity	1.0	0.9	0.8	0.8	0.7	0.7

*Body-centred cubic.

the entire table is francium, followed closely by cesium. The alkali metals, which make up the extreme left-hand file, have ionization energies ranging from 124.3 kilocalories per mole (Li) to 89.7 kcal/mole (Cs) (omitting the rare radioactive element francium). The alkaline-earth metals, the next group to the right, have higher ionization energies, ranging from 214.9 (beryllium) to 120.1 kcal/mole (barium).

Electronegativities

The electronegativity scale of the elements compares the ability of the atoms of the various elements to attract electrons to themselves. In the periodic table the electronegativities range from 0.7 for cesium, the least electronegative of the elements, to 4.0 for fluorine, the most electronegative. Metals are ordinarily considered to be those elements having values less than 2.0 on the electronegativity scale. As a group the alkali metals are the least electronegative of the elements, ranging from 0.7 to 1.0 on the scale, while the alkaline earths, the next group on the table, have electronegativities ranging from about 0.9 to 1.5.

The relatively simple electronic structure of the atoms of the alkali metals (a single electron in the outermost shell of each) results in a considerable degree of regularity and predictability in their physical and chemical properties. Bonds of the alkali metals with all nonmetals are essentially ionic (resulting from mutually attractive forces existing between positive and negative electrical charges). The physical properties vary in a regular pattern and, to a considerable extent, can be attributed to the effects of the mobile outermost electron.

The Table summarizes the important physical and thermodynamic properties of the alkali metals. Generally, the mobile electrons in metals are responsible for their high thermal and electrical conductivity and their high heat capacity; this is especially true of the alkali metals. These metals are all characterized by a body-centred cubic crystallographic arrangement (a standard pattern of atoms in their crystals), with eight nearest neighbours to each atom.

The closest distance between atoms, a characteristic property of crystals, increases with increasing atomic weight of the alkali metal atoms. As a group, the alkali metals have a looser crystallographic arrangement than any of the other metallic crystals, and cesium, because of its greater atomic weight, has an interatomic distance that is greater than that of any other metal.

Vapour-pressure data for the alkali metals and for two alloys formed between elements of the group are given in Figure 2, which shows that the vapour pressures increase in regular fashion with increasing atomic weight.

Melting points and boiling points

The melting points of the alkali metals as a group are lower than those of any other nongaseous group of the periodic table, ranging between 179° C for lithium and 28.5° C for cesium. Only mercury among the metallic elements has a lower melting point than cesium. The low melting points of the alkali metals are a direct result of the large interatomic distances in their crystals and the

From J. Mansteller, F. Tepper, and S. Rodgers, *Alkali Metal Handling and Systems Operating Techniques* (1967); Gordon & Breach

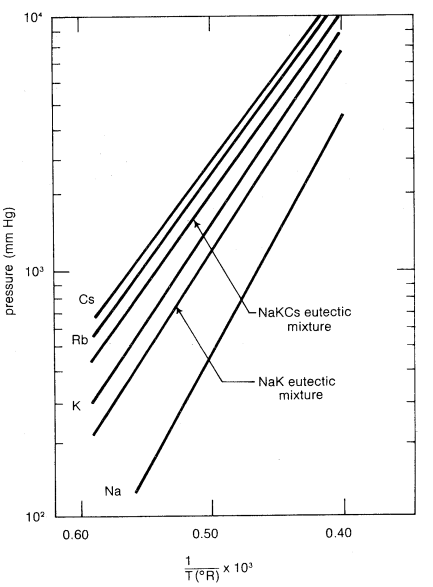


Figure 2: Vapour pressure in millimetres of mercury (mm Hg) plotted against the inverse of absolute temperature on the Rankine scale (T [°R]), for the alkali metals sodium (Na), potassium (K), rubidium (Rb), and cesium (Cs). As the atomic weight increases toward cesium the vapour pressure is higher at any temperature chosen. Alloys of metals that have single melting points, called eutectic mixtures, follow the same rule.

weak bond energies associated with such loose arrays. These same factors are responsible for the low densities, low heats of fusion, and small volume differences during fusion of the metals. Lithium, sodium, and potassium are less dense than water.

Cesium is the most volatile of the alkali metals with a boiling point of 671° C. The boiling points of the alkali metals increase in regular fashion as the atomic numbers decrease, the highest boiling point, 1,317° C, being that of lithium, the alkali metal of lowest atomic number.

Chemical properties. Since the alkali metals are the most electropositive (the least electronegative) of elements, they react with a great variety of nonmetals.

Reactions with oxygen. All of the alkali metals react with oxygen—lithium and sodium to form monoxides and the heavier alkali metals to form superoxides. The rate of reaction with oxygen, or with air, depends upon whether the metals are in the solid or liquid state, as well as upon the degree of mixing of the metals with the oxygen, or air. In the liquid state alkali metals can be ignited in air with ease, generating copious quantities of heat and a dense choking smoke of the oxide.

In its chemical reactivity lithium more closely resembles Group IIa of the periodic table than it does the other metals of its own group. It is less reactive than the other alkali metals with water, oxygen, and halogens and more reactive with nitrogen, carbon, and hydrogen.

The free energy of formation (a measure of stability) of the alkali metal oxides at 25° C varies widely from a high of −133 kilocalories per mole for lithium oxide to −63 kilocalories per mole for cesium oxide. The close approach of the small lithium ion to the oxygen atom results in the unusually high free energy of formation of the oxide. Normal oxide formed on combustion of lithium is the simple oxide (Li_2O); it also is possible to make the peroxide (Li_2O_2) by passing oxygen through a liquid-ammonia solution of lithium metal. Sodium peroxide (Na_2O_2) can be similarly made, although the commercial method is by oxidation of sodium monoxide (Na_2O) with oxygen. Sodium superoxide (NaO_2) can be prepared with high oxygen pressures, whereas the superoxides of rubidium, potassium, and cesium can be prepared directly by combustion in air. By contrast, no superoxides have been isolated in pure form in the case of the alkaline-earth metals, although the heavier members of that group can be oxidized to the peroxide state. The ozonides of potassium, rubidium, and cesium, which are less stable than the lower oxides, can be prepared by the reaction of the superoxides with ozone.

Oxides, peroxides, and superoxides

Reactions with water. The alkali metals all react violently with water. The rate of the reaction depends on the degree of metal surface presented to the liquid. With small metal droplets or thin films of alkali metal, the reaction can be explosive. The rate of reactivity of water with the alkali metals increases with increasing atomic weight of the metal. With the heavier alkali metals the hydroxides are highly soluble; thus they are removed readily from the reacting surface, and the reaction can proceed with unabated vigour. The reaction involves equimolar mixtures (that is, equal numbers of atoms or molecules) of the alkali metal and water to form a mole (an amount equal to that of the reactants) of alkali metal hydroxide and half a mole of hydrogen gas. These reactions are highly exothermic (give off heat), and the hydrogen that is generated can react with oxygen to further increase the heat that is generated.

Reactions with nonmetals. Of the alkali metals, only lithium reacts with nitrogen, and it forms a nitride (Li_3N). In this respect it is more similar to the alkaline-earth metals than to the Group Ia metals. Lithium also forms a relatively stable hydride, whereas the other alkali metals form hydrides that can be more readily dissociated. Lithium forms a carbide (Li_2C_2) similar to that of calcium. The other alkali metals do not form stable carbides, although they do react with the graphite form of carbon to give intercalation compounds (substances in which the metal atoms are inserted between layers of carbon atoms in the graphite structure).

The alkali metals can be burned in atmospheres of the various halogens to form the corresponding halides. The reactions are highly exothermic, producing up to 235 kcal/mole for lithium fluoride. The alkali metals react with nonmetals in Groups Va and VIa of the periodic table. Sulfides can be formed by the direct reaction of the alkali metals with elemental sulfur, furnishing a variety of sulfides. Phosphorus combines with the alkali metals to form phosphides with the general formula M_3P.

Formation of alloys. The characteristics of alloy behaviour in alkali metals can be evaluated in terms of the similarity of the elements participating in the alloy. Elements with similar atomic volumes form solid solutions (that is, mix completely in all proportions); some dissimilarity in atomic volumes results in eutectic-type systems (solutions formed over limited concentration ranges); and further dissimilarity results in totally immiscible systems.

The elements potassium, rubidium, and cesium, which have rather similar atomic volumes and ionization energies, form complete solid solutions and mixed crystals. Sodium, which is a significantly smaller atom than potassium and has a higher ionization energy, tends to form eutectic systems with potassium, rubidium, and cesium. Even greater dissimilarity exists in the atomic volumes of sodium and lithium, resulting in insolubilities of the liquid phases. The consular temperature (the temperature at which the two liquids become completely miscible) increases on going from the lithium–sodium alloy system to the lithium–cesium system. Lithium and cesium can coexist as two separate liquid phases at temperatures up to at least 1,100° C.

Eutectics

There is only one example of solid miscibility in alkali–alkaline-earth metal binaries—the lithium–magnesium system, in which the two elements are very similar. Sodium forms compounds only with barium in the alkaline-earth-metal series. The heavier alkali metals all tend to form immiscible liquid phases with the alkaline earths.

Several of the elements of the Group IIb of the periodic table (zinc, cadmium, and mercury) react with the alkali metals to form compounds. Mercury forms at least six compounds, commonly termed amalgams, with each of the five alkali metals; and, with the exception of the amalgam with lithium, the highest melting point compound in each series has the formula MHg_2. Lithium and sodium also form compounds with cadmium and zinc.

LITHIUM

Occurrence and production. Lithium occurs as salts in mineral springs, and its concentration in sea water is 0.1 parts per million. Crude phosphate, containing 20 percent lithia, or lithium oxide (Li_2O), is recovered from brines at Searles Lake, California. Lithium is found in pegmatite ores, such as spodumene ($LiAlSi_2O_6$) and lepidolite (of varying structure), or in amblygonite ($LiAlFPO_4$) ores, with Li_2O contents ranging between 4 and 8.5 percent.

Lithium is manufactured by the electrolysis of a fused mixture of lithium and potassium chlorides. The lithium chloride consumed in the electrolysis may be obtained from lithium-containing silicate minerals by a number of different processes. Spodumene, for example, is calcined (heated), treated with sulfuric acid, and then neutralized with calcium carbonate to form lithium sulfate solution. Soda ash (sodium carbonate) is added to produce lithium carbonate, the basic material from which other lithium compounds are derived. Lithium chloride may be volatilized from aluminosilicate materials by heating with carbon in a stream of chlorine. Lithium chloride also may be prepared in high yield by passing hydrogen chloride over lepidolite ore at 935° C.

The addition of potassium chloride to lithium chloride allows electrolysis to occur at 400°–420° C. Since the voltage at which decomposition of lithium chloride takes place is lower than that of potassium chloride, lithium is deposited in purities greater than 97 percent. Graphite anodes are used in the electrolytic production of lithium, while the cathodes are made of steel. The pure lithium formed at the cathode coalesces at the surface of the elec-

trolyte to form a molten pool, which is protected from reaction with air by a thin film of the electrolyte. The lithium is ladled from the cell and cast by pouring it into a mold at a temperature only slightly above the melting point, leaving the solidified electrolyte behind. The solidified lithium is then remelted, and materials insoluble in the melt either float to the surface or sink to the bottom of the melt pot. The remelting step reduces the potassium content to less than 100 parts per million (ppm).

Many lithium alloys are produced directly by the electrolysis of molten salts, containing lithium chloride in the presence of a second chloride, or by the use of cathode materials that interact with the deposited lithium, introducing other elements into the melt.

Lithium production is second only to sodium production among the alkali metals. During World War II substantial amounts of lithium were produced for the manufacture of lithium hydride, which served as a convenient source of hydrogen for military purposes. The maximum capacity for lithium production in the U.S. in 1955 was estimated to be 200 tons, approximately 85 percent of the total world production capacity. The cost of elemental lithium is between $8.00 and $9.50 U.S. per pound, depending on the quantity purchased.

Significant uses. The isotope lithium-6, the less common of the two chief isotopes, yields tritium gas when bombarded with neutrons. Tritium so produced is employed in the manufacture of hydrogen bombs, among other uses. The principal industrial applications for lithium are in metallurgy, where the active element is used as a scavenger (remover of impurities) in the refining of such metals as iron, nickel, copper, and zinc, and their alloys. A large variety of nonmetallic elements are scavenged by lithium, including oxygen, hydrogen, nitrogen, carbon, sulfur, and the halogens. Lithium is utilized to a considerable extent in organic synthesis. Organolithium compounds act somewhat like organomagnesium compounds and undergo reactions similar to the Grignard reaction, a standard synthetic procedure in organic chemistry. Lithium compounds also serve as catalysts in synthetic rubber manufacture.

Lithium has potential value as a heat-transfer fluid for high power-density nuclear reactors. The lithium-7 isotope, the more common stable isotope, has a low nuclear cross section (that is, absorbs neutrons very poorly) and is an ideal primary coolant for nuclear reactors in which coolant temperatures above about 800° C are required.

Lightweight lithium–magnesium alloys have structural applications in the aerospace and other industries. Metallic lithium is used in the preparation of compounds such as lithium hydride.

Chemical properties and principal compounds. In many of its properties lithium exhibits the same characteristics as do the more common alkali metals sodium and potassium; thus, the metal is highly reactive and forms strong hydroxide solutions. In many respects lithium also shows similarities to the elements of the alkaline-earth group, especially magnesium, which has similar atomic and ionic radii. This similarity is seen in oxidation properties, the monoxide being normally formed in each case.

Lithium forms a hydride that is more thermodynamically stable than sodium hydride. Lithium also forms carbides and nitrides by direct combination of the elements; sodium does not. Both lithium and magnesium carbonates are more stable thermally than are the common alkali metal carbonates; the lithium and magnesium carbonates are rather insoluble in water, as are the fluorides and phosphates. Lithium halides readily absorb moisture from the air and, in this respect, resemble magnesium and calcium salts more than sodium salts. The halides and other salts of lithium are more soluble than other alkali-metal halides in organic solvents like methanol.

Lithium hydrides

Lithium hydride, produced by the direct reaction of the elements at approximately 700° C, is a white crystalline solid that melts at 680° C and is crystallographically related to the lithium halides. The hydrogen in the molecule is the anion (bears the negative charge) and is released at the anode (the positive pole) when the hydride is electrolyzed. Lithium hydride is used chiefly as a convenient source for hydrogen since it reacts vigorously with water to form lithium hydroxide and hydrogen. Lithium aluminum hydride, a useful reagent in organic chemistry, can be prepared by the action of lithium hydride on aluminum chloride.

Lithium oxide

Lithium oxide is used widely as a constituent of glass; it also has application in the manufacture of enamels and glazes. The peroxide (Li_2O_2) may be directly prepared by oxidation of lithium–ammonia solutions followed by autoclaving (heating under pressure) in the presence of oxygen. The peroxide is used in atmosphere regeneration, as in space capsules, where it reacts with water to release oxygen, the lithium hydroxide that forms absorbing excess carbon dioxide.

Lithium hydroxide (LiOH) is a common starting material for preparing many lithium compounds. A process widely used in its production is the reaction of lithium carbonate with lime (calcium oxide or its hydrate) according to the equation:

$$Li_2CO_3 + Ca(OH)_2 \rightarrow 2\,LiOH + CaCO_3.$$

The hydroxide also may be prepared by digesting lithium ores, such as spodumene, with lime followed by leaching with water. Lithium hydroxide is used directly in atmosphere regeneration because of its capacity for absorbing carbon dioxide. It has other uses, including the mercerization (pretreatment) of cotton, and it is employed as an additive in alkaline storage battery electrolytes.

Lithium halides

The principal lithium halides of industrial importance are the chloride and bromide. Single crystals of lithium fluoride are used in optical components. Lithium fluoride is the only lithium halide that is not hygroscopic (water absorbing), and consequently it is used as a flux (metal cleaning agent) in a number of joining processes. The production of lithium chloride by hydrogen chloride treatment of lithium carbonate is carried out on a large scale. A number of patented processes exist for the recovery of the chloride directly from minerals. Anhydrous lithium chloride has a crystallographic pattern like that of sodium chloride; it is highly water soluble and quite hygroscopic. Its principal use is as a source of lithium (see above). The eutectic LiCl–KCl system is used as an electrolyte in thermal batteries. Lithium chloride also is used to lower the melting points of the fused-electrolytic baths used in the electrolytic production of molybdenum, titanium, and other exotic metals. Lithium bromide and iodide are used as drying agents in air conditioning, in the dehydration of mineral oil, and in gas drying.

Nuclear properties. Lithium, which exhibits no natural radioactivity, has two isotopes of mass number 6 and 7 respectively. The ^7Li : ^6Li ratio is between 12 and 13. The isotopes ^8Li (half-life 0.855 sec) and ^9Li (half-life 0.17 sec) have been produced by nuclear bombardment. The ^7Li isotope has a low nuclear cross section (0.003 barn) permitting consideration of the substance as a coolant for nuclear reactors in such mobile systems as satellites and spacecraft.

Biological properties. The widespread occurrence of lithium in plants results in a wide, although low-level, distribution of lithium in animals. Lithium salts have complex effects when absorbed into the body. They are not highly toxic, although they can affect body processes dependent upon salt balance. Lithium poisoning can occur, particularly with patients on low sodium diets. Lithium salts that are ingested are absorbed and are rapidly excreted in the urine.

SODIUM

Occurrence and production. Sodium is the most abundant of the alkali metals. Sodium chloride (table salt) is the most common compound of sodium, but many others also are known. Sodium is an important constituent of a number of silicate materials, such as feldspars and micas. There are huge deposits of rock salt in various parts of the world, and sodium nitrate deposits exist in Chile and Peru. The sodium content of the sea is approximately

1.05 percent, corresponding to a concentration of approximately 3 percent of sodium halides. Sodium has been identified in both the atomic and ionic forms in the spectra of stars and the interstellar medium. Analysis of meteorites indicates that the silicate material present has an average content of approximately 4.6 atoms of sodium for every hundred atoms of silicon.

Sodium is the alkali metal of most commercial importance by far. In 1968 sodium production in the United States totalled about 157,000 tons. Most processes for the production of sodium involve the electrolysis of sodium chloride. The element is used chiefly to produce tetraethyl lead, which increases the antiknock rating of gasoline. The cost of sodium is approximately 17 cents U.S. per pound in large quantities.

The principal sodium compounds of commerce are the chloride, carbonate, and sulfate.

The annual U.S. consumption of sodium chloride is greater than 31,000,000 tons, much of it employed in the production of other heavy (industrial) chemicals. The consumptions of the carbonate (soda ash) and the sulfate are respectively 6,000,000 and 1,000,000 tons.

Significant uses. Two of the earliest uses of sodium were in the manufacture of sodium cyanide and sodium peroxide, both uses that continue in importance. A present major use for sodium is in the manufacture of tetraethyl lead by means of the reaction of lead–sodium alloy with ethyl chloride. The amount of sodium consumed for this purpose alone in 1953 was 137,000,000 pounds. Substantial amounts of sodium (50,000,000 pounds in 1953) are also used in the manufacture of sodium alkyl sulfates for the growing synthetic-detergent industry, in which the sodium alkyl sulfates are employed as the principal ingredient.

Sodium also is used as a starting material in the manufacture of sodium hydride. In addition, sodium is employed in the production of dyes and dye intermediates, in the synthesis of perfumes, and in a wide variety of organic reductions. It is used in the purification of hydrocarbons and in the polymerization of unsaturated hydrocarbons. In many organic applications, sodium is used in the form of dispersions in hydrocarbon liquid media.

Heat-transfer uses of sodium

Sodium is an excellent heat-transfer fluid, and because of this property, it has potentially large-scale uses. It is anticipated that, by the year 2000, fast-breeder nuclear reactors, cooled by sodium, will generate much of the world's electrical power. Sodium forms an alloy with potassium (NaK) which melts at −10° C. At one time NaK was considered for use as the heat-transfer fluid in submarine reactor systems. Later, sodium was investigated as the secondary coolant system in a high power-density reactor system for space use. The alloy also is used for cooling crucibles in consumable arc-melting processes for preparing titanium. Another large-scale potential use for sodium is in secondary batteries for electric automobiles. Considerable effort was being made in the early 1970s to evolve a molten sodium–sulfur battery that would provide primary propulsion in automobiles.

Because of the excellent electrical conductivity of sodium, the substance has been considered as a substitute for copper in electrical conductors. It is used extensively in metallurgy as a deoxidant and as a reducing agent for the preparation of calcium, zirconium, titanium, and other transition metals. Commercial production of titanium involves reduction of titanium tetrachloride with sodium.

Chemical properties and principal compounds. Generally, elemental sodium is more reactive than lithium, and in aqueous solution it reacts to form a stronger base than LiOH. Its chemistry is well explored.

Reaction with oxygen. Sodium is ordinarily quite reactive with air, and the reactivity is a function of the relative humidity, or water-vapour content, of the air. The corrosion of solid sodium by oxygen also is accelerated by the presence of small amounts of impurities in the sodium. In ordinary air, sodium metal reacts to form a sodium hydroxide film, which can rapidly absorb carbon dioxide from the air, forming sodium bicarbonate. Sodium does not react with nitrogen. It is significantly more reactive in air as a liquid than as a solid, and the liquid can ignite at about 125° C. In a comparatively dry atmosphere, sodium burns quietly, giving off a dense, white, caustic smoke, which can cause choking and coughing. The temperature of burning sodium increases rapidly to more than 800° C, and under these conditions the fire is extremely difficult to extinguish.

Sodium monoxide (Na_2O) is ordinarily formed upon oxidation of sodium in dry air. The superoxide (NaO_2) can be prepared by heating metallic sodium to 300° C in an autoclave (heated pressure vessel) containing oxygen at high pressure (2,700 pounds per square inch). Another route to the superoxide is oxidation of sodium peroxide treated to have a large surface area.

Sodium that is heavily contaminated with the monoxide may be readily purified by filtration since the solubility of the oxide in molten sodium is low. This low solubility is utilized to a considerable extent in continuous purification processes of the sodium in large liquid-metal reactor systems. A second technique for removing the oxide, called cold trapping, involves running the molten sodium through a cooled, packed bed of material, upon which the oxide can precipitate. Filtration and cold trapping also are effective in removal of gross quantities of carbonate, hydroxide, and hydride.

Reaction with water and hydrogen. The reaction with water of liquid sodium having a high surface area can be explosive. The sodium–water reaction is highly exothermic (that is, heat is given off):

$$Na + H_2O \rightarrow NaOH + \tfrac{1}{2} H_2 +$$
33.67 kilocalories per mole at 25° C.

Tests have indicated, however, that sodium and water cannot be mixed fast enough to produce the shock waves characteristic of high explosives. The explosive hazards of the reaction are associated primarily with the hydrogen gas that is formed.

Preparation of sodium hydride

Pure sodium begins to absorb hydrogen appreciably at about 100° C, the rate of absorption increasing with the temperature. Pure sodium hydride can be formed at temperatures above 350° C by exposing sodium to hydrogen gas at a high flow rate. The dissociation of sodium hydride is considerably greater than that of lithium hydride but slightly less than that of potassium hydride.

Reactions with nonmetals. Generally, alkali metals react with halogen gases, the degree of reactivity decreasing with increasing atomic weight of the halogen. Sodium is no exception to this statement. Under certain conditions of reaction, sodium and halogen vapours react to produce light (chemiluminescence). Halogen acids, such as hydrochloric acid, react vigorously with sodium, yielding the sodium halides. The reactions are highly exothermic, with heats of reaction (energy given off) of −71.8 and −76.2 kilocalories respectively for the reactions with hydrofluoric and hydrochloric acids. Sodium is attacked by other strong mineral acids to form the corresponding salts. It reacts with the fumes of nitric acid at 15° C to form sodium nitrate and with acetic and sulfuric acids to form sodium acetate and sodium sulfate. With molten sulfur it reacts violently to produce polysulfides; under more controlled conditions it reacts with organic solutions of sulfur. Liquid selenium and tellurium both react vigorously with solid sodium to form selenides and tellurides.

Sodium shows relatively little reactivity with carbon, although the existence of lamellar (layer-like) materials prepared from graphite with the formula NaC_{64} has been reported. At 625° C carbon monoxide reacts with sodium to form sodium carbide and sodium carbonate.

With the exception of the oxides of the Group IVb metals (titanium, zirconium, and hafnium), the oxides of the transition metals are all reduced to the respective metals with elemental sodium. Sodium also reacts with a large number of metallic halides, displacing the metal from the salt and forming a sodium halide in the process. This reaction is used in the preparation of several of the transition metals themselves, including titanium and tantalum.

Reactions in liquid ammonia

Sodium dissolves in liquid ammonia to give intense blue solutions, and at ordinary temperatures a slow reaction occurs in such solutions, resulting in the formation of sodamide, $NaNH_2$, and hydrogen. Conversion of alkali metals to their amides takes several days in the case of lithium and sodium, but for the heavier alkali metals conversion occurs in a half-hour or less—and with cesium it takes only 15 minutes. The reaction of sodium and liquid ammonia can be catalyzed by the addition of any of a number of metal and metal oxide catalysts.

Liquid ammonia often is used as a solvent for sodium, allowing a number of reactions to occur at ordinary temperatures that would otherwise need heat. Sodium superoxide (NaO_2), for example, can be formed by passing oxygen through ammonia solutions of sodium at −77° C. Ammonia also serves as a solvent for reactions of sodium with arsenic, tellurium, antimony, bismuth, and a number of other low-melting metals.

Organic reactions. The organic reactions of sodium have been studied to a greater extent than those of any of the other alkali metals. Sodium reacts with anhydrous alcohols to form the respective alcoholates. The reaction is most vigorous with methanol and decreases with increasing molecular weight of the alcohol. Sodium methoxide is produced on an industrial scale by reaction of sodium with excess methanol. Organic acids react with sodium to form sodium salts.

The high free energy of formation of sodium halides permits the dehalogenation of a large number of organic halides, the formation of the sodium halide being energetically favoured. The so-called Wurtz reaction—based on this principle—is used in organic synthesis to a considerable extent:

$$2RCl + 2Na \rightarrow R\text{–}R + 2NaCl.$$

By this reaction, octane can be made from bromobutane and sodium. Organosodium compounds include a number in which the sodium atom is bonded directly to a carbon atom; an example is methylsodium, $Na\text{–}CH_3$. Such compounds can be prepared by the action of sodium on mercury dialkyls or diaryls, as in the following equation:

$$Hg(CH_3)_2 + 2Na \rightarrow 2NaCH_3 + Hg.$$

Reaction with halogenated hydrocarbons

Sodium reacts violently with a number of halogenated hydrocarbons. A violent explosion occurs when a mixture of carbon tetrachloride and sodium is subjected to shock, and even when the sodium is diluted to a considerable extent—as in sodium amalgam—a brisk reaction with carbon tetrachloride occurs.

Alloys. Sodium is completely miscible with the alkali metals below it in the periodic table (potassium, rubidium, and cesium). A eutectic (that is, an alloy that melts lower than its components) melting at −10° C is formed in the sodium–potassium system and is known commercially as NaK. Its composition is approximately 78 percent potassium, and it is used as a heat-transfer fluid and as an organic reactant. The eutectics formed in the sodium–rubidium and sodium–cesium binary systems melt respectively at −4.5° and −30° C. Sodium is the minor component with potassium and cesium of the ternary alloy NaKCs, melting at −78° C. This fluid is the lowest melting liquid alloy yet isolated.

Sodium also forms alloys with the alkaline-earth metals. Beryllium is soluble in sodium only to the extent of a few atomic percent at approximately 800° C. Liquid sodium and magnesium are only partially miscible. The degree of solubility in sodium of the alkaline-earth metals increases with increasing atomic weight, with the result that the solubility of calcium is 10 percent by weight at 700° C. In the sodium–strontium system, there is a considerable degree of miscibility. Sodium forms a number of compounds with barium, and several eutectics exist in the system.

The precious metals, silver, gold, platinum, palladium, and iridium, and the white metals, such as lead, tin, bismuth, and antimony, alloy to an appreciable extent with liquid sodium. Cadmium and mercury also react with sodium, and a number of compounds exist in both binary systems. Seven sodium–mercury compounds, or amalgams, exist, with Hg_2Na having the highest melting point (354° C). Sodium amalgams are used chiefly for carrying out reactions in situations in which pure elemental sodium would be violently reactive and difficult to control. The solubility of transition metals in alkali metals is generally very low, often in the one to ten parts per million range even at temperatures in excess of 500° C.

The sodium compounds of major commercial importance are the chloride, hydroxide, carbonate, sulfate, and, to a lesser extent, the peroxide, sulfide, and halides other than the chloride.

Nuclear properties. In its nuclear properties, sodium is a simple element, with but one natural isotope, ^{23}Na, exhibiting no natural radioactivity. Several isotopes have been prepared by nuclear reactions, with the ^{22}Na isotope having the longest half-life—2.6 years. The ^{24}Na isotope (half-life 15 hours) is produced by irradiation in a nuclear reactor. Because of this reaction a sodium-cooled reactor must have a second heat-transfer loop so that radioactive sodium does not come in contact with the environment.

Biological properties. Sodium salts, particularly sodium chloride, are found almost everywhere in biological material. Sodium is an essential element for life, as is potassium, and the two elements maintain a definite balance within the cell structure. Most of the biological effects of sodium salts are the result of the cation (sodium ion), with the anion (negative ion) apparently not playing a dominant role.

The presence of salinity in soils is often detrimental to plant growth. Sodium ions replace calcium and other ions in clay complexes, transforming the clay to a sticky mass; water percolation is then drastically reduced, and the basicity of the soil rises markedly. Sodium nitrate is widely used as a fertilizer because it serves as a nitrogen source for crops, although the more expensive potassium nitrate is much more effective.

Sodium as a dietary requirement

The tolerance of fish to changes in salinity is often quite remarkable. Many marine bacteria and diatoms are able to tolerate salt concentrations as great as 25 percent. The minimum sodium requirement for mammals appears to be 0.05 percent of the diet, corresponding in a normal man to a requirement of one to two grams of salt per day, resulting in an average sodium content of body tissues of 0.24 percent. There is a wide variation of sodium content in the different tissues, with whole blood containing approximately 0.62 percent sodium chloride, whereas skin has a sodium content of less than 0.1 percent. There is a relationship between salt content and water balance of the body, a low intake of salt causing loss of water. Considerable quantities of sodium are lost through the skin by sweating, and considerable quantities can be excreted in the urine.

POTASSIUM

Occurrence and production. Igneous rocks, shale, and sediment are minerals containing much of the potassium that is found in nature. The potassium content of the Dead Sea is estimated at approximately 1.7 percent potassium chloride, and many other salty bodies of water are rich in potassium. The waste liquors from certain salt works may contain up to 40 grams per litre of potassium chloride and are used as a source of potassium. The salt deposits in Stassfurt, Germany, which are rich in potassium sulfate and chloride, are one of the most important sources of potassium in the world. Potassium also is found in nature as the nitrate.

Extraction and purification of potassium salts

The techniques used for extracting potassium salts depend to a considerable extent on the nature of the ore body and the character of the foreign salts mixed with the potassium salt. The Stassfurt salt deposit contains large deposits of carnallite, $MgCl_2 \cdot KCl \cdot 6H_2O$. These deposits became so important with respect to potash (potassium hydroxide) production that Germany had a virtual monopoly on production of that substance up to the time of World War I. The other chief sources of potash

are in France, Austria, Spain, India, and Chile. Seawater, brines, many lake deposits, and the ashes of vegetation also are used as sources of potash.

Potassium chloride can be separated from carnallite ore by preferential precipitation of potassium chloride during evaporation of the mother liquor. In the Searles Basin, California, deposit, brine is evaporated in vacuum pans, and the sodium chloride and sulfate are deposited. The resulting liquor is boiled to a point where potassium chloride is separated, producing a crude potash deposit containing about 66 percent potassium chloride.

Production of potassium

Potassium is produced by sodium reduction of molten potassium chloride at 870° C. Molten KCl is continuously fed into a packed distillation column, while sodium vapour is passed up through the column. By condensation of the volatile potassium at the top of the distillation tower, the reaction

$$Na + KCl \rightarrow K + NaCl$$

is forced to the right. Efforts to devise a scheme for the electrolytic production of potassium have been unsuccessful because there are few salt additives that can reduce the melting point of potassium chloride to temperatures where electrolysis is efficient.

Commercial grades of potassium (more than 98 percent pure) sell for $2 U.S. a pound, approximately ten times that of sodium on a weight basis. The price difference between the two metals is further magnified when the relative costs are considered on a molar basis—a mole of sodium weighing considerably less than a mole of potassium. (Twenty-three pounds of sodium are chemically equivalent to 39 pounds of potassium.) Chiefly because of the price difference, sodium is produced in quantities greater than 150,000 tons per year in the United States, whereas the total U.S. potassium production is only 100 tons. Potassium is used mainly for the manufacture of potassium superoxide (KO_2). Only about 10 to 15 tons of elemental potassium are marketed in the U.S., and a similar tonnage of the alloy with sodium.

Significant uses. Most of the potassium produced is converted to the superoxide directly by combustion in dry air. The superoxide is used as a source of oxygen in respiratory equipment, since it generates oxygen and also absorbs carbon dioxide.

Sodium–potassium alloy (NaK) is used only to a limited extent as a heat-transfer coolant, although it has significant potential for this purpose. It is presently being evaluated for use in crucible cooling in titanium arc melting. The alloy also is used as an organic catalyst.

Chemical properties and principal compounds. The chemical properties of potassium are similar to those of sodium, although the former is considerably more reactive. Potassium differs from sodium in a number of respects. Whereas sodium is essentially unreactive with graphite, potassium, rubidium, and cesium react to form a series of interlamellar compounds, the richest in metal having the formula MC_8. Compounds are formed with carbon–alkali metal atomic ratios of 8, 16, 24, 36, 48, and 60 : 1. The graphite lattice is expanded during penetration of the alkali metal between the layers. Potassium reacts with carbon monoxide at temperatures as low as 60° C to form an explosive carbonyl ($K_6C_6O_6$), a derivative of hexahydroxybenzene. Sodium is essentially nonreactive with carbon monoxide.

Liquid potassium and NaK both are more reactive than liquid sodium with air and oxygen. Potassium superoxide (KO_2) is the preferred oxide, whereas sodium forms the monoxide (Na_2O). Oxidation of NaK produces Na_2O preferentially. The superoxide of potassium is a yellow solid consisting of K^+ and O_2^- ions. It also can be formed by oxidation of potassium amalgam with dry air or oxygen.

Potassium reacts violently with water to produce one-half mole of hydrogen per mole of potassium and water and generating approximately 47 kcal/gram-atom of heat. It can be stored under nitrogen gas with no reaction. It reacts with hydrogen at approximately 350° C to form the hydride.

Potassium is highly reactive with halogens and detonates when it contacts liquid bromine. Violent explosions also have been observed when mixtures of potassium and halogen acids are subject to shock. Explosions also have occurred when potassium is being mixed with a number of metal halide salts or with organic-halogen compounds.

At elevated temperatures, potassium reduces carbon dioxide to carbon monoxide and carbon. Solid carbon dioxide and potassium react explosively when subjected to shock. Oxidation of potassium amalgam with carbon dioxide results in the formation of potassium oxalate. Potassium is not reactive with benzene, although heavier alkali metals such as cesium add across the double bonds between carbon atoms in the benzene molecules, giving organometallic products.

Potassium alloys and compounds

In addition to the alloys of potassium with lithium and sodium (see above), alloys with other alkali metals are known. Complete miscibility exists in the potassium–rubidium and potassium–cesium binary systems. The latter system forms an alloy melting at approximately −38° C. Modification of the system by the addition of sodium results in a ternary eutectic melting at approximately −78° C. The composition of this alloy is 3 percent sodium, 24 percent potassium, and 73 percent cesium. Potassium is essentially immiscible with all of the alkaline-earth metals, as well as with zinc, aluminum, and cadmium.

The chief use of potassium compounds is in agriculture, 94 percent of them being used for agricultural purposes, with the other 6 percent consumed in such diverse chemical manufacturing products as detergents, glass, ceramics, and textiles.

Nuclear properties. The two stable isotopes of potassium are ^{39}K and ^{41}K, which together comprise 99.99 percent of the element as it occurs in nature, about 13.5 times as much of the former isotope normally being present. The natural radioactivity of potassium is due to beta radiation from the ^{40}K isotope (10^9 years half-life), which is present in the element in nature to the extent of about 0.01 percent. The disintegration of ^{40}K is used in geological age calculations.

Biological properties. Potassium is an essential ingredient in all forms of life. Each organism has a closely maintained potassium level and a relatively fixed potassium–sodium ratio. Potassium is the primary inorganic cation within the living cell, and sodium is the most abundant cation in extracellular fluids. Potassium is an essential ingredient of fertilizers. The potassium content of plants varies considerably, but it is ordinarily in the range of 0.5 to 2 percent. In man, the ratio of potassium between the cell and plasma is approximately 27 : 1. The potassium content of muscle tissue is approximately 0.3 percent, whereas that of blood serum is about 0.01 to 0.02 percent. The dietary requirement for normal growth is approximately 3.3 grams of potassium per day, but the ingestion of more than 20 grams of potassium results in distinct physiological effects. Excess potassium is excreted in the urine, and a significant quantity may be lost during sweating.

RUBIDIUM AND CESIUM

Because rubidium and cesium are very similar in their physical and chemical properties, they are conveniently discussed together.

Occurrence and production. Rubidium and cesium often occur together in nature. Rubidium, however, is more widely scattered and seldom forms a natural mineral; it is found only as an impurity in other minerals, ranging in content up to 5 percent in such minerals as lepidolite and pollucite. Brine samples have been analyzed that contain up to six parts per million of rubidium.

The earth's crust is estimated to contain only 7 ppm of cesium, whereas it contains 310 ppm of rubidium. Pollucite ($Cs_4Al_4Si_9O_{26} \cdot H_2O$), however, is a cesium-rich mineral resembling quartz. It contains 40.1 percent cesium on a pure basis, and impure samples are ordinarily separated by hand-sorting methods to greater than 25 percent cesium. Large pollucite deposits have been found in

Rhodesia and in the lithium-bearing pegmatites at Bernic Lake, Manitoba. Cesium also occurs in a rare mineral called rhodizite, and it is found in low concentrations in lepidolite and in salt brines and saline deposits.

Source of rubidium

The principal commercial source of rubidium is accumulated stocks of a mixed carbonate produced as a by-product in the extraction of lithium salts from lepidolite. Primarily a potassium carbonate, the by-product also contains approximately 23 percent rubidium and 3 percent cesium carbonates. The primary difficulty associated with the production of either pure rubidium or pure cesium is that these two elements are always found together in nature and also are mixed with other alkali metals; because these elements are very similar chemically, their separation presents numerous problems. Before development of procedures based on thermochemical reduction and fractional distillation, the elements were purified in the salt form through laborious crystallization techniques. Once pure salts have been prepared by precipitation methods, it is a relatively simple task to convert them to the free metal. This end is ordinarily accomplished by calcium reduction in a high-temperature vacuum system in which the highly volatile alkali metal is distilled from the reaction mixture.

Direct reduction of the mixed carbonates from lepidolite purification, followed by fractional distillation, is perhaps the most important of the commercial methods for producing rubidium. The mixed carbonate is treated with excess sodium at approximately 650° C, and much of the rubidium and cesium passes into the metal phase. The resulting crude alloy is vacuum distilled to form a second alloy considerably richer in rubidium and cesium. This product is then refined by fractional distillation in a tower to produce elemental rubidium more than 99.5 percent pure.

The direct thermochemical reduction of pollucite ore with sodium metal is the primary commercial source of cesium metal. Use of this process may lead to significant decreases in cesium prices if the market for cesium increases. In the process ground pollucite ore is treated with elemental sodium at approximately 650° C to form a sodium–cesium alloy containing some rubidium impurity. Fractional distillation of this alloy in a distillation tower at approximately 700° C produces 99.9 percent pure cesium metal.

Significant uses. Rubidium metal has few commercial uses, and it is of very minor economic significance. The high prices and uncertain and limited supply of the metal discourage the development of any uses. Prices quoted for rubidium metal range from $300 to $425 U.S. per pound. Production of rubidium metal in the United States, for example, is probably no greater than a few hundred pounds per year. Only a few commercial uses have been developed for the metal itself, although the salts are of slightly greater economic significance because of their use in photocells and radio vacuum tubes. The total U.S. production of cesium metal in 1969 was probably less than one ton. In small quantities the cost of cesium is approximately $250 U.S. per pound, although prices of $15 per pound have been quoted for large quantities. Cesium has been evaluated for power systems for space applications and as a potential rocket fuel for interplanetary travel. Because the cesium atom can be ionized thermally and the positively charged ions accelerated to great speeds, cesium systems can provide specific impulses for rocket propulsion that are extraordinarily high. Cesium also has application in thermionic converters that generate electricity directly within nuclear reactors. Another potentially large application of cesium metal is in the production of low-melting NaKCs eutectic alloy. Cesium salts are used in moderate quantities in the manufacture of photomultiplier tubes, infrared lamps, scintillation counters, and spectrophotometers.

Space applications of cesium

Chemical properties and principal compounds. Rubidium and cesium are difficult to handle because they react spontaneously in air. If the surface area of a sample of the metal is large, it can burn to form superoxides. Rubidium superoxide (RbO_2) is a yellow powder; cesium superoxide has a more reddish cast. Peroxides of both metals (Cs_2O_2 and Rb_2O_2) can be formed by oxidation of the metal with the required amount of oxygen. Rubidium forms two other oxides (Rb_2O and Rb_2O_3), but the reactions of cesium with oxygen are much more complex. A number of suboxides exist, many of them having melting points below that of cesium metal itself. These suboxides include Cs_7O (melting point 3° C), Cs_7O_2, and Cs_3O; the monoxide is Cs_2O.

Rubidium and cesium both form hydrides when hydrogen is passed over the metal at 400° and 600° C, respectively. These hydrides have dissociation pressures slightly greater than those of sodium and potassium hydrides. Neither cesium nor rubidium forms a nitride, although they react with gaseous ammonia to form amides. Both elements burn with considerable chemiluminescence in halogen vapours, forming the respective halide salts. Rubidium and cesium also are highly reactive with water, generating hydrogen in the fashion typical of alkali metals. Rubidium and cesium also react with carbon to form graphitic compounds.

Rubidium and cesium form organic compounds with carbon monoxide that are derivatives of the aldehyde glyoxal. These metals dissolve in liquid ammonia to form the blue solution characteristic of alkali metals, and this solution reacts with oxygen at about −16° C to form the peroxides and amides of the metals.

An organic reaction peculiar to cesium is the reaction with aromatic hydrocarbons, including benzene and toluene, in which cesium atoms add across the aromatic bond, forming saturated ring compounds. The product of reaction with benzene is $C_6H_6Cs_6$, a substance that is spontaneously inflammable in air and reacts with water to give hydrogen and cyclohexane. In the presence of cesium, ethylene is hydrogenated even at room temperature.

Rubidium and cesium are miscible in all proportions and have complete solid solubility; a melting-point minimum of + 9° C is reached. Both rubidium and cesium form a number of mercury amalgams. Because of the increased specific volumes of rubidium and cesium, as compared to the lighter alkali metals, there is less tendency for them to form alloy systems with other metals.

Rubidium and cesium hydroxides may be prepared readily by electrolysis of an aqueous solution of the corresponding chloride, using a mercury cathode; the amalgam formed at the cathode is decomposed by water to give the free hydroxide. The use of rubidium and cesium hydroxides in alkaline-storage and mercury-cell batteries has been suggested. The hydroxides would be substituted for sodium and potassium hydroxides in order to obtain more efficient operation at lower temperatures.

The halides and polyhalides are the most important compounds of rubidium and cesium, but even these are expensive specialty chemicals. Rubidium and cesium halides exhibit some distinct differences from the halides of the lighter alkali metals. Unlike the lighter alkali metals, rubidium and cesium form stable polyhalides. The stability of the polyhalides increases with increasing atomic weight of both the alkali metal and the halogen, with the result that cesium tri-iodide (CsI_3) is the most stable of the metal polyhalides. Rubidium and cesium chlorides will react with fluorine to form perfluorides, such as cesium trifluoride (CsF_3). Rubidium and cesium halides also are distinctive in their ability to form stable acid fluorides of the type CsF • 2HF. The melting points of these compounds decrease with increasing atomic weight of the alkali metal, whereas the maximum number of molecules of hydrogen fluoride that can combine with the metal fluoride increases with the atomic weight of the metal. The compound CsF • 6HF has a melting point of −42.3° C. Another distinction of the heavier alkali metal halides is their increased miscibility with the parent metal. Liquid cesium chloride is completely miscible with liquid cesium whereas the solid salt is soluble in the liquid metal to the extent of about 15 mole percent.

Nuclear properties. Two rubidium isotopes are found in nature. The stable rubidium-85 makes up 72 percent of the mixture. Radioactive rubidium-87 makes up 28

percent; it emits beta rays with a half-life of about 6×10^{11} years. Cesium has only one naturally occurring isotope, with a mass number of 133. A number of cesium isotopes with short half-lives have been produced artificially by nuclear reactions.

FRANCIUM

The existence of francium was predicted by Dmitry Ivanovich Mendeleyev in his periodic classification of the elements. In 1939 the French chemist Marguerite Perey discovered the radioactive element, which has beta activity and a half-life of 22 minutes (^{223}Fr). Nineteen other isotopes of francium also have been reported. It has been calculated that the total quantity of francium in the earth's crust is only 24.5 grams. Information on biological and chemical aspects of the element is virtually nonexistent.

ANALYTICAL CHEMISTRY OF THE ALKALI METALS

Chemical separation of lithium can be effected by the use of the organic nitrogen-containing compound picrolonic acid, which precipitates the heavier alkali metals from chloride solutions, leaving lithium behind. Lithium can be detected with a number of azo dyes that give highly sensitive colour reactions with lithium ions in alkaline solution. A modification of the uranyl acetate test (the precipitation of an insoluble sodium salt with uranyl acetate) is used as a standard test for the presence of sodium. The use of a cobaltinitrite solution permits separation of potassium from sodium by precipitation of the insoluble potassium salt. There are essentially no satisfactory analytical methods for rubidium and cesium based on the use of reagents in solution; these elements are commonly determined spectroscopically—that is, by measurements of their flame spectra or light-emission patterns.

Instrumental analysis

Ordinary methods of chemical analysis generally require extensive preliminary separation. With the alkali metals this process frequently can be avoided by the use of flame tests, which are quickly and easily applied and able to detect small amounts of the alkali metals in the presence of other elements. When salt solutions of the alkali metals are atomized in the coal gas burner of a simple flame photometer, which measures emitted light and must be provided with a suitable interference filter (to remove unwanted emissions), the metals can be determined generally in the parts-per-million range. Determination of one alkali metal in the presence of another, however, can result in interference, which can be reduced using specially prepared standard solutions containing known amounts of the interfering metals for comparative purposes.

Emission spectrographic analysis (measuring light emitted at particular wavelengths, a more precise method than by a photometer) is a useful method for analyzing trace quantities of one alkali metal in another and is also very helpful for analyzing metallic impurities. Colorimetric methods, which compare colours produced in solution, are often used for the analysis of metallic impurities.

The analysis of the alkali-metal samples for the presence of nonmetallic elements, such as oxygen, carbon, hydrogen, and nitrogen, requires specialized techniques. The oxygen content of sodium and potassium samples can be determined by extraction of the free alkali metal with mercury, leaving behind mercury-insoluble oxides and carbonates, which can subsequently be analyzed using solution methods. The oxygen content of rubidium and cesium can be accurately determined by precise measurement of the freezing point of these two elements.

The carbon content of alkali metals can be analyzed by oxidation of the alkali metal in pure oxygen, followed by infrared measurement of the carbon dioxide generated during combustion. For the analysis of nitride in lithium, the nitride commonly is converted to ammonia, and the ammonia is measured by colorimetric analysis.

The sampling of alkali metals to obtain representative examples for the analysis of nonmetallic elements is a difficult art. Such sampling ordinarily is carried out by melting the whole batch of alkali metal and extracting a sample by dipping a bucket into the melt, withdrawing it, and analyzing the contents of the bucket.

SPECIAL PROBLEMS IN HANDLING ALKALI METALS

The high degree of reactivity of the alkali metals with most other substances requires special precautions in handling. Caution is particularly called for if there is the chance of contact of the free metal with water, air, or certain organic compounds, such as the halogenated hydrocarbons. Many of the more reactive compounds of the alkali metals, such as the superoxides, also require special handling.

Reactions with water

The intensity of reactivity with water increases greatly when the metals are in the liquid state. The reaction of water with solid alkali metals frequently generates enough heat to melt the mass, thereby accelerating the reaction. The hydrogen that is produced may cause secondary fires or explosions. In industrial situations alkali residues left in metal containers are extracted with water only under very carefully controlled conditions and behind barricades. Alcohols are used to clean alkali metals from systems; methanol is the most reactive, and higher alcohols are often used if the reaction is too vigorous.

Alkali metal fires can be very spectacular, since the flames are coloured and are accompanied ordinarily by dense white smoke. The alkali metals increase in their susceptibility to air ignition with increase in atomic weight; they are more reactive at high temperatures and when highly subdivided. Once ignited, however, the lighter alkali metals such as lithium and sodium burn more vigorously than the others and sometimes reach temperatures higher than 800° C within the molten mass.

Alkali-metal fires are very difficult to extinguish. The surest method is the exclusion of air. Extinguishants ordinarily used in fire fighting such as water and carbon dioxide are prohibited in the case of alkali metals. Inorganic salts are widely used to extinguish alkali-metal fires. Their action is to limit the flow of air to the surface of the metal, thereby smothering the fire. These salts are commercially available in resin-coated form, which keeps them dry and free flowing. They are not effective with lithium fires, however, because they sink into the melt. Graphite is used in fighting lithium fires.

Proper protective equipment and clothing, especially eyeglasses and gloves, are necessary when working with alkali metals. Rubber gloves are adequate with small laboratory systems, but insulated leather gloves are more desirable with liquid metals at high temperatures. In the latter cases, a hard hat, face shield, and noncombustible clothing are desirable.

If alkali metal comes in contact with the skin, it is advisable to wash the area contacted with water and with a 3-percent acetic acid solution to neutralize the strong base formed. If large particles of free alkali metal adhere to the skin, mineral oil should be applied around the local area and the free bits of metal should be removed with tweezers.

Storage and handling

Alkali metals are shipped and stored in metal containers, under either mineral oil or an inert cover gas. Mineral oil can be removed from solid sticks of the metal by washing with petroleum ether. Only lithium and sodium should be cut with a knife because of the possibility of explosions with the more reactive alkali metals. Molten metals can be contained up to temperatures of about 870° C by iron, nickel, and cobalt alloys, although lithium rapidly dissolves nickel. Above 870° C the alkali metals are kept in containers of niobium or tantalum alloy. Alkali metal peroxides and superoxides also require special care, because they are active oxidants and often react explosively with reducing agents.

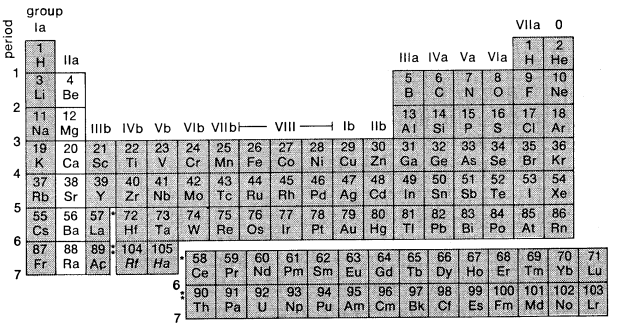
BIBLIOGRAPHY. J.W. MELLOR, *Inorganic and Theoretical Chemistry*, vol. 2, suppl. 2, *The Alkali Metals* (1961), a comprehensive work devoted to the alkali metals and their compounds, including history, production methods, analysis, alloy systems, and physical, chemical, biological, and spectrochemical properties; J.W. MAUSTELLER, F. TEPPER, and S.J. RODGERS,

Alkali Metal Handling and System Operating Techniques (1967), a monograph on alkali metal technology devoted chiefly to laboratory handling aspects, including a summary of physical properties, chemical properties of the pure metals, analytical techniques, handling, and safety procedures; M. SITTIG, *Sodium* (1956), a monograph covering the manufacture, inorganic and organic reactions, physical properties, handling, and uses of sodium; *The Alkali Metals*, an international symposium held at Nottingham, July 1966 (1967), a symposium of 44 papers, covering most aspects of alkali metals, including preparation, chemical and physical properties, and compounds.

(F.T.)

Alkaline-Earth Metals and Their Compounds

The term alkaline-earth metals is the generic name for the chemical elements beryllium (Be), magnesium (Mg), calcium (Ca), strontium (Sr), barium (Ba), and radium (Ra), which are the members of group IIa of the periodic table (see Figure) of the elements. Prior to the 19th century, substances that were nonmetallic, insoluble in water, and unchanged by fire were known as earths. Those earths, like lime, that resembled the alkalies (soda ash and potash) were designated alkaline earths. Alkaline earths were thus distinguished from the alkalies and from other earths, such as alumina and the rare earths. By the first decade of the 19th century it became clear that the earths, formerly considered to be elements, were in fact oxides, compounds of a metal and oxygen. The metals whose oxides make up the alkaline earths then came to be known as the alkaline-earth metals and have been classified in group II of the periodic table ever since the Russian chemist Dmitry Ivanovich Mendeleyev proposed his first table in 1869.



Alkaline-earth metals in the periodic table of the elements.

Chemical and physical properties

The alkaline-earth metals are extremely electropositive; that is, like the alkali metals of group Ia (*e.g.*, sodium, potassium), their atoms easily lose electrons to become positive ions (cations). Most of their typical compounds are therefore ionic: salts in which the metal occurs as uniformly divalent cations, M^{2+}, where M represents any group IIa atom. The salts are colourless unless they include a coloured anion (negative ion). Typical alkaline-earth compounds, calcium chloride ($CaCl_2$) and calcium oxide (CaO), may be contrasted with the compounds of the uniformly monovalent alkali metals, sodium chloride (NaCl) and sodium monoxide (Na_2O). The oxides of the alkaline-earth metals are basic (*i.e.*, alkaline, in contrast to acidic). A fairly steady increase in electropositive character is observed in passing from beryllium, the lightest member of the group, to radium, the heaviest; as a result of this trend, beryllium oxide is only weakly basic and even shows acidic properties, whereas barium and radium oxide are strongly basic. The metals themselves are highly reactive reducing agents; that is, they readily give up electrons to other substances that are, in the process, reduced.

All the metals and their compounds find commercial application to some degree, especially magnesium alloys and a variety of calcium compounds. Magnesium and calcium, particularly the latter, are abundant in nature and play significant roles in geological and biological processes. Radium is a rare element; all its isotopes are radioactive.

HISTORY

The earliest known alkaline earth was lime (Latin *calx*), which is now known to be calcium oxide; it was used in ancient times in the composition of mortar. Magnesia, (the name derives probably from the ancient district of Magnesia in Asia Minor), the oxide of magnesium, was shown to be an alkaline earth different from lime by the Scottish chemist Joseph Black in 1755; he observed that magnesia gave rise to a soluble sulfate, whereas that derived from lime was known to be insoluble. In 1774 Carl Wilhelm Scheele, the Swedish chemist who discovered oxygen, found that the mineral called heavy spar or *barys* (Greek, heavy) contained a new earth, which became known as baryta (barium oxide). A further earth, strontia (strontium oxide), was identified by Adair Crawford, a London physician and chemist, in 1790 on examining a mineral (strontium carbonate) found in a lead mine at Strontian in Argyllshire, Scotland. Beryllia (beryllium oxide) was extracted from the mineral beryl and recognized as an earth by the French analytical chemist Louis-Nicolas Vauquelin in 1797. Though at first confused with alumina (aluminum oxide) because both dissolve in alkali, beryllia was shown to be distinct; unlike alumina, it reprecipitated when the alkaline solution was boiled for some time. Beryllia was originally called *glucina* (Greek *glykys*, sweet) because of its sweet taste. (This etymological root is retained in France, where the element beryllium is also known as glucinium.)

Discovery of alkaline-earth metals

Magnesium, calcium, strontium, and barium, elements derived from alkaline earths, were isolated as impure metals by the English chemist Sir Humphry Davy in 1808 by means of the electrolytic method he had previously used for the isolation of the alkali metals potassium and sodium. The alkaline-earth metals were later produced by reduction of their salts with free alkali metals, and it was in this way (the action of potassium on beryllium chloride) that beryllium was first isolated by the German chemist Friedrich Wöhler in 1828. Radium was discovered in 1898 by means of its radioactivity by Pierre and Marie Curie, who separated it with great difficulty from barium.

PHYSICAL AND CHEMICAL BEHAVIOUR

The alkaline-earth elements are highly metallic and are good conductors of electricity. They have a gray-white lustre when freshly cut but tarnish readily in air, particularly the heavier members of the group. Beryllium is sufficiently hard to scratch glass, but barium is only slightly harder than lead. The melting points and boiling points of the group are higher than those of the corresponding alkali metals; they vary in an irregular fashion, magnesium having the lowest (mp 650° C and bp 1,105° C) and beryllium the highest (mp 1,283° and bp about 2,500°). The elements crystallize in one or more of the three regular close-packed metallic crystal forms. Chemically, they are all strong reducing agents. The free metals are soluble in liquid ammonia—the dark-blue solutions of calcium, strontium, and barium arousing considerable interest because they are thought to contain metal ions and the most unusual species, solvated electrons, or electrons resulting from the interaction of the metal and the solvent. Highly concentrated solutions of these elements have a metallic, copper-like appearance, and further evaporation yields residues containing ammonia, which correspond to the general formula $M(NH_3)_6$. The solutions are strong reducing agents and are useful in a number of chemical processes.

Electronic configuration

The atoms of the alkaline-earth elements all have similar electronic structures, consisting of a pair of electrons (designated *s* electrons) in an outermost orbital, within which is a stable electronic configuration corresponding to that of a noble gas. The noble gas elements—helium (He), neon (Ne), argon (Ar), krypton (Kr), xenon (Xe), and radon (Rn)—have generally complete electron shells. Strontium has the configuration $1s^2 2s^2 2p^6 3s^2$

Some Properties of the Alkaline-Earth Metals

	beryllium	magnesium	calcium	strontium	barium	radium
Atomic number	4	12	20	38	56	88
Atomic weight (stablest isotope)	9.0122	24.312	40.08	87.62	137.34	(226)
Colour of element	gray	silvery white	silvery white	silvery white	silvery white	silvery white
Melting point (°C)	1,283	650	842–48	769	725	700
Boiling point (°C)	~2,500	1,105	1,487	1,384	1,140	<1,737
Density at 20° C (g/cm³)	1.85	1.74 (5° C)	1.54	2.54	3.51	(about 5)
Oxidation number	2	2	2	2	2	2
Electronic configuration	$1s^2 2s^2$	(Ne)$3s^2$	(Ar)$4s^2$	(Kr)$5s^2$	(Xe)$6s^2$	(Rn)$7s^2$
Isotopic abundance (terrestrial, percent)	^{9}Be(100)	^{24}Mg(78.70), ^{25}Mg(10.13), ^{26}Mg(11.17)	^{40}Ca(96.97), ^{42}Ca(0.64), ^{43}Ca(0.145), ^{44}Ca(2.06), ^{46}Ca(0.0033), ^{48}Ca(0.18)	^{84}Sr(0.56), ^{86}Sr(9.86), ^{87}Sr(7.02), ^{88}Sr(82.56)	^{130}Ba(0.101), ^{132}Ba(0.097), ^{134}Ba(2.42), ^{135}Ba(6.59), ^{136}Ba(7.81), ^{137}Ba(11.32), ^{138}Ba(71.66)	—
Radioactive isotopes (mass numbers)	7, 10–11	20–23, 27–28	37–39, 41, 45, 47, 49, 50	80–83, 85, 89–93 95	123, 125–29, 131, 133, 139–43	213–17, 219–30
Electrical resistivity at 0° C (microhm-cm)	4.0	4.45	3.91	23	60	—
Crystal structure*	HCP	HCP	FCC, HCP, BCC	FCC, HCP, BCC	BCC	—
Radius, ionic (M^{2+}, Å)	0.31	0.65	0.99	1.13	1.35	1.40
Diameter, atomic (Å) (coordination number of 12)	2.25	3.20	3.93	4.30	4.48	—
Ionization energy (electron volts)						
First	9.32	7.64	6.11	5.69	5.21	5.28
Second	18.21	15.03	11.87	11.03	10.00	10.14
Third	153.85	80.12	51.21	—	—	—
Electrode potential (according to international convention), $M^{2+} + 2e^- \rightarrow M$, at 25° C, volts	−1.85	−2.37	−2.87	−2.89	−2.90	−2.92
Electronegativity (Pauling)	−1.5	−1.2	−1.0	−1.0	−0.9	−0.9

*HCP = hexagonal close packed, FCC = face-centred cubic (cubic close packed), BCC = body-centred cubic.

$3p^6 3d^{10} 4s^2 4p^6 5s^2$, which may be written as (krypton core) $5s^2$, or simply (Kr)$5s^2$. Similarly, Be may be designated as (He)$2s^2$, Mg as (Ne)$3s^2$, Ca as (Ar)$4s^2$, Ba as (Xe)$6s^2$, and Ra as (Rn)$7s^2$. The prominent lines in the atomic spectra of the elements, obtained when the elements are heated under certain conditions, arise from states of the atom in which one of the two *s* electrons has been promoted to a higher energy orbital. The *s* electrons are relatively easily ionized (removed from the atom), and this ionization is the characteristic feature of alkaline-earth chemistry. The ionization potential (the energy required to strip an electron from the atom) falls continuously in the series from beryllium (9.32 electron volts [eV]) to barium (5.21 eV); radium, the heaviest in the group, has a slightly higher ionization potential (5.28 eV). The small irregularities observed in the otherwise smooth change as one proceeds down the group as it appears in the periodic table are explained by the uneven filling of electron shells in the successive rows of the table. The *s* electrons may also be promoted to *p* orbitals of the same principal quantum number (within the same shell) by energies similar to those required to form chemical bonds; the atoms are, therefore, able to form stable covalently bonded structures, unlike helium, which has the otherwise analogous electronic configuration of $1s^2$.

Zinc, cadmium, and mercury, the group IIb elements, are often compared with the alkaline-earth elements calcium, strontium, and barium. Cadmium, for example, has the electronic configuration (Kr)$4d^{10}5s^2$, with the ten 4*d* electrons taking virtually no part in chemical bonding. The $5s^2$ electrons, however, are much less readily ionized in cadmium than they are in strontium, for the 4*d* electrons act as an ineffective shield for the corresponding increased charge on the cadmium nucleus. The chemistry of the IIb metals, therefore, is markedly less ionic than the chemistry of the alkaline-earth metals.

Ionic character. The chemistry of the alkaline-earth metals, like that of the alkali metals, is for the most part reasonably interpreted in terms of an ionic model for the compounds formed. This model is less satisfactory for the chemistry of beryllium and magnesium than for the heavier alkaline-earth metals. In fact, most beryllium compounds are molecular (covalent) rather than ionic. Although there is some evidence for the transient existence of singly charged alkaline-earth ions, in most cases the chemistry of these elements is dominated by the formation and properties of the doubly charged M^{2+} ions, in which the outermost *s* electrons have been stripped from the metal atom. The resulting ion is stabilized by electrostatic interaction with a solvent, like water, which has a high dielectric constant and a great ability to absorb electrical charge, or by combination with ions of opposite charge in an ionic lattice such as is found in salts. The extra energy required to remove the second *s* electron (the second ionization potential being approximately twice the first) is more than compensated for by the extra binding energy present in the doubly charged ion. The removal of a third electron from an alkaline-earth atom, however, would require an expenditure of energy greater than could be recouped from any chemical environment. As a result, therefore, the alkaline-earth metals show a constant valence of +2 in their compounds.

Ionic sizes

The sizes of the ions of the alkaline-earth elements increase steadily from Be^{2+}, which has a radius of 0.31Å or 31×10^{-10} cm, to Ra^{2+} with a radius of 1.40Å. The ionic radius of the europium ion Eu^{2+} (1.12Å) is very close to that of strontium Sr^{2+} (1.13Å); this means that Eu^{2+} ions can sometimes be used as a "probe" for the alkaline-earth metals, substituting for strontium ions in situations where advantage can be taken of the spectroscopic and magnetic properties that make Eu^{2+} readily identifiable. The ionic radius of the cadmium ion Cd^{2+} (0.97Å) is very similar to that of calcium Ca^{2+} (0.99Å). A quantitative comparison of cadmium and calcium chemistry, therefore, clearly shows up the less ionic character of cadmium chemistry without complications due to differences in ionic size. A related comparison may be made between mercury and strontium, because of the similar ionic radii of Hg^{2+} (1.10Å) and Sr^{2+} (1.13Å).

The salts of the alkaline-earth metals have crystal structures that may be understood in terms of the packing of oppositely charged spheres of appropriate radii representing the cations and anions composing them. The mode of packing in each case is dictated by the relative sizes of the spheres. Thus, magnesium fluoride, MgF_2, with a cation–anion radius ratio of 0.48, has a crystal structure in which each magnesium cation is surrounded by six fluoride anions (the so-called rutile structure), whereas CaF_2, SrF_2, and BaF_2, each with larger radius ratios, 0.73, 0.83, and 0.99, respectively, are able to adopt a crystal structure with eight fluoride ions surrounding the metal ions (a fluorite structure). When the lattice energies (*i.e.*, energies released on forming a crystal lattice from its component ions) of the alkaline-earth salts are calculated in terms of their electrostatic properties, the results are in surprisingly good agreement with those obtained experimentally.

Chemical reactions and ion size

The ionic model of alkaline-earth compounds also predicts semiquantitatively the course of many chemical reactions. Thus in the double decomposition reaction involving an oxide anion (smaller) and a sulfide anion (larger), $MO + M'S \rightarrow M'O + MS$, the reaction goes from left to right when the radius of M'^{2+} is less than that of M^{2+}, for the largest total lattice energy is obtained when the smaller cation is combined with the smaller anion. Similarly in the decomposition of a carbonate, $MCO_3 \rightarrow MO + CO_2$, the reaction goes to the right more readily with M^{2+} ions of smaller radius. The driving force of this reaction may be considered to be the increase of lattice energy arising from the changeover from a large carbonate anion (CO_3^{2-}) to the smaller oxide anion (O^{2-}); this gain in energy will increase as the radius of the M^{2+} ion decreases.

The chemistry of radium is less well investigated than that of the other alkaline-earth metals. As expected, however, it is in general an extrapolation of the chemistry of calcium, strontium, and, especially, barium.

The group IIa ions are readily hydrated, with the strength of bonding to the water molecules increasing with decreasing ionic radius. Although the number of water molecules directly attached to the metal ion may be greater with the larger ions for purely steric (geometrical) reasons, the total number of water molecules associated with the metal ion nevertheless increases inversely with the size of the ion itself (as shown by migration experiments conducted in aqueous solution). Large anions, such as sulfate, tend to form weak ion-pair complexes more readily with the larger metal ions of the family, but weak-acid anions, such as acetate, tend to form stronger complexes with the smaller metal ions, particularly those of magnesium and beryllium. That many of these complexes are molecular rather than ionic is shown by their ready extraction from aqueous solution (which preferentially dissolves ionic substances) into organic solvents (which dissolve molecular ones).

Chemical compounds. *Hydrides.* All of the alkaline-earth metals form hydrides, of molecular formula MH_2, but those of magnesium and, more especially, beryllium are unstable. MgH_2 has an ionic structure, whereas BeH_2 is regarded as an electron-deficient polymer somewhat akin to the higher boron hydrides. The remaining alkaline-earth metal hydrides are colourless and saltlike, with hydride anions occupying two kinds of position in an approximately hexagonal close packing of metal atoms in the crystals. The alkaline-earth metal hydrides are strong reducing agents that react with water to liberate hydrogen; thus calcium hydride (CaH_2), like lithium hydride (LiH), may be used as a convenient source of hydrogen without the need for pressure cylinders. On heating, the hydrides lose a portion of their hydrogen to form other hydrides of irregular composition.

Halides. The halides of beryllium are unlike most metallic halides in being very poor electrical conductors in the fused state. This suggests a tendency to molecular (covalent) rather than ionic character, which is also reflected in their structures; thus beryllium chloride, $BeCl_2$, has a chainlike structure. The beryllium halides are very readily hydrated, and their solutions on evaporation yield basic salts. On passing down the group, as it appears in the periodic table, the character of the halides becomes progressively more ionic: $MgCl_2$ has a layer structure whereas $CaCl_2$ has a crystalline lattice. Bromides and iodides show similar changes. Magnesium and calcium salts are good drying agents, but hydrated magnesium halides, when heated, tend to lose hydrogen halide to form magnesium oxide or a basic halide. With the exception of the beryllium salt the fluorides have very low solubilities in water.

Oxides. The lattice energies of the alkaline-earth metal oxides decrease with increasing size of the metal ion. This is reflected in their hardness (as shown on Mohs scale where diamond is 10): BeO, 9.0; MgO, 6.5; CaO, 4.5; SrO, 3.5; BaO, 3.3. A similar but much less regular effect is found in the melting and boiling points.

Beryllium oxide is weakly basic and also shows some acidic properties, since it is soluble in alkali. In this way beryllium shows a relationship to aluminum, the second element of group IIIa of the periodic table. Beryllium oxide is not known to combine with water to give the hydroxide, $Be(OH)_2$, although at high temperatures its volatility is dependent upon the water vapour pressure, which suggests that the hydroxide is formed under these circumstances. The oxide is soluble in solutions of many beryllium salts; oxo or hydroxo complexes are thought to be formed with Be–O–Be or Be–OH–Be bridges, respectively. Similar complexes are believed to exist in concentrated solutions of $Be(OH)_2$.

Magnesium oxide, MgO, is a weak base with no acidic properties; it combines slowly with water to form the insoluble hydroxide, $Mg(OH)_2$. The remaining oxides in the series are all strongly basic and combine readily with water giving the respective hydroxides. The stability to heat of the hydroxides (*i.e.*, their resistance to conversion to the oxide by loss of water) increases steadily from $Ca(OH)_2$ through $Sr(OH)_2$ to $Ba(OH)_2$.

Peroxides. Beryllium does not form a peroxide, and it is doubtful if the so-called magnesium peroxide of commerce is a true peroxide of magnesium. Calcium peroxide, CaO_2 (used for bleaching and disinfecting), may be prepared by the action of hydrogen peroxide on milk of lime (calcium hydroxide), but not by direct action of the elements; some calcium peroxide, however, is said to be formed by the action of oxygen on solutions of calcium in liquid ammonia. Whereas only small amounts of strontium peroxide, SrO_2, may be made by heating strontium oxide in oxygen under pressure, barium peroxide, BaO_2, is readily formed by heating the oxide in air or oxygen. The formation and decomposition of barium peroxide was the basis of a method formerly used for the extraction of oxygen from the air; BaO_2 also is employed as an intermediate in the production of hydrogen peroxide.

Method for oxygen extraction

Sulfates. The sulfates, MSO_4, become more stable and more insoluble in water as one proceeds down the group. Beryllium sulfate, $BeSO_4$, loses sulfur trioxide, SO_3, at about 580° C; $BeSO_4$ forms many hydrates and is readily soluble in water. Magnesium sulfate, $MgSO_4$, is stable up to white heat; it also forms many hydrates (the heptahydrate is known as Epsom salts) and is soluble in water. Calcium, strontium, and barium sulfates lose sulfur trioxide at progressively higher temperatures (1,400° C for $BaSO_4$). Calcium sulfate is sparingly soluble in water, whereas the barium salt is extremely insoluble and radium sulfate, even more so. The precipitation of barium sulfate is used in the standard gravimetric procedures for determining the percentage of barium or sulfate in a sample. Because of its high insolubility, $BaSO_4$ forms very fine particles, which are difficult to filter. Care must be taken, therefore, to ensure slow precipitation; one method is to make use of the slow conversion of sulfamic acid (NH_2SO_3H) to sulfate in water, which produces the precipitate only gradually. Barium sulfate dissolves in concentrated sulfuric acid as $Ba(HSO_4)_2$.

Carbonates. Normal beryllium carbonate does not exist, but a basic carbonate may be precipitated from a solution of a beryllium salt by addition of sodium carbo-

nate. Here again beryllium demonstrates its weakly basic and weakly metallic nature. The remaining carbonates, like the sulfates, are relatively insoluble, with the solubility decreasing as one proceeds down the group. The resistance of the carbonates to heat, like that of the sulfates, increases down the group, although the decomposition temperatures of the carbonates are lower ($MgCO_3$, 400° C; $CaCO_3$, 900°; $SrCO_3$, 1,280°; $BaCO_3$, 1,360°).

Water hardness

Magnesium and calcium carbonates are soluble in carbonic acid solutions to give the bicarbonates, $M(HCO_3)_2$. When they are present in water, these bicarbonates contribute to its hardness; the magnesium and calcium ions react with soaps (fatty-acid anions) to form precipitates, and the decomposition of the bicarbonate anions on heating, with attendant loss of carbon dioxide, leads to the formation of boiler scale (insoluble carbonates).

Other inorganic compounds. The basic acetate of beryllium, $Be_4O(C_2H_3O_2)_6$, is employed as a means of purifying the metal. The acetate may be extracted from aqueous solution into organic solvents. In the molecules of the acetate, the central oxygen atom is surrounded by the four beryllium atoms, and each beryllium by four oxygen atoms from the acetate groups.

The alkaline-earth elements are widely distributed in natural silicate minerals. The filtration of water through zeolites, hydrated sodium aluminum silicates, is used to soften water, because sodium ions from the zeolite replace the calcium and magnesium ions in the water, and these ions are taken up by the zeolite. The zeolite may be regenerated with strong solutions of sodium chloride. Cement and most varieties of glass contain calcium silicate.

The alkaline-earth metals combine directly with the nonmetals not yet referred to (silicon, selenium, tellurium, nitrogen, phosphorus, arsenic, antimony, carbon, silicon, germanium, and boron) to give binary compounds (Mg_3N_2, MgC_2, Mg_2C_3, etc.). The compounds formed with the more electronegative nonmetals are commonly classified as ionic, but they are better regarded as only semiionic because they also show a number of the characteristics of alloys. In the structures of these compounds the metal atoms have small numbers of nonmetal atoms as near neighbours, with the nearest metal atoms substantially further away; direct bonding between the nonmetal atoms within these molecules also is quite common, especially in the silicides and borides. Thus calcium silicide, $CaSi_2$, contains layers of silicon atoms, and calcium boride, CaB_6, contains octahedral B_6 groups joined together at their corners in a three-dimensional network. The most characteristic chemical property of these compounds is that they react with water to give the metal hydroxide and a hydride (or hydrides) of the nonmetal. Because of its ready reaction with water to produce acetylene, calcium carbide (CaC_2) is commonly used as a source of that material, and mixtures of silanes, germanes, or boranes (binary hydrides of silicon, germanium, and boron) may be made by the hydrolysis of magnesium silicide, germanide, or boride, respectively.

Organometallic compounds. Beryllium alkyls (containing paraffinic hydrocarbon radicals, such as ethyl -C_2H_5), and beryllium aryls (containing aromatic hydrocarbon radicals, such as phenyl,-C_6H_5), may be prepared by the reaction of beryllium halides with Grignard reagents or with lithium alkyls or aryls. In these organometallic compounds the metal is bonded directly to a carbon atom in the organic portion of the molecule. These substances are usually liquids or low-melting solids of high chemical reactivity. Their structures are those of polymerized molecular species. Of the organometallic compounds, those of magnesium are the most widely used; as Grignard reagents (named for their discoverer, the French chemist Victor Grignard) they are the key materials in a number of important reactions in organic and organometallic chemistry. The Grignard reagents are prepared by the direct interaction of metallic magnesium and an alkyl or aryl halide in a solvent, commonly ether. Their empirical formula is RMgX, where R is an alkyl or aryl group and X is a halogen. In solution, the RMgX species are in equilibrium with MgR_2 and MgX_2; the exact distribution between these and their significance in Grignard reactions has been the subject of much controversy. The alkyls and aryls of calcium, strontium, and barium are markedly ionic, more so than is generally true of organometallic compounds. They are highly reactive and of little practical importance.

Grignard reagents

Molecular complexes. The complexes, or loose molecular associations, of the alkaline-earth elements, like the compounds, are held together by essentially electrostatic (ionic) forces; thus they are formed with polar, usually oxygen-containing, ligands (donor molecules) and small (*e.g.*, fluoride) or highly charged (*e.g.*, phosphate) anions. Unlike the ions of the transition metals, however, the alkaline-earth metal ions fail to form stable complexes with polarizable ligands in aqueous solution. As a general rule, weak-acid anions (acetate, hydroxide) tend to form stronger complexes with the lighter members of the alkaline-earth series than they do with the heavier members. Some multidentate weak-acid anions (*e.g.*, tartrate, ethylenediaminetetraacetate), however, form weaker complexes around the smaller Mg^{2+} cation because of steric hindrance. Large (strong-acid) anions (*e.g.*, sulfate, nitrate) form ion-pair complexes more readily with the heavier members of the series.

The complexing of alkaline-earth cations, especially Ca^{2+} and Mg^{2+}, is of considerable importance in washing processes because soaps (long-chain alkyl carboxylates) and many detergents (long-chain alkyl sulfonates) form complexes with Ca^{2+} and Mg^{2+} ions. Water softeners are even stronger complexing agents for Ca^{2+} and Mg^{2+}, often by virtue of their many coordinating groups (chelating agents); they include zeolites, phosphates, and sulfonated polystyrene ion exchangers. The formation of complexes may also be used to produce electrodes that are selective to certain ions.

Analysis. For gravimetric analysis (by weight of solid precipitates) beryllium is normally converted to its oxide, magnesium to the pyrophosphate or oxinate, calcium to the oxalate, strontium to its sulfate, and barium to either sulfate or chromate. Some of these insoluble materials readily form the basis of volumetric procedures (*e.g.*, oxinate, oxalate, chromate). Complexometric titration, dependent upon complex formation, with ethylenediaminetetraacetate is also a convenient volumetric procedure; the end point of the titration may be observed by displacement of a more weakly coordinated indicator (*e.g.*, the dye Eriochrome Black T). The pigment sodium rhodizonate is a colorimetric reagent for barium and strontium, with which it complexes through its negatively charged oxygen atoms. Volatile compounds of the alkaline-earth metals (especially the chlorides) impart characteristic colours to the flame of an ordinary laboratory or Bunsen burner: brick red for calcium, carmine for strontium, yellow-green for barium, and crimson for radium.

Flame analysis

INDIVIDUAL ALKALINE-EARTH METALS

Beryllium. Beryllium is widely distributed in the earth's crust, of which it composes about 0.001 percent; its cosmic abundance is 20 on the scale in which silicon, the standard, is one million. Although there are about 30 recognized minerals containing beryllium, only three—beryl ($3BeO \cdot Al_2O_3 \cdot 6SiO_2$), phenacite ($2BeO \cdot SiO_2$), and chrysoberyl ($BeO \cdot Al_2O_3$)—are of any significance, and only the first is of industrial importance. The precious forms of beryl, emerald and aquamarine, have a composition closely approaching that given above, but industrial ores contain less beryllium; most beryl is obtained as a by-product of other mining operations, the larger crystals being picked out by hand. The world production of beryl (11 percent BeO) was 6,900 metric tons in 1967.

Exploration for beryllium has been considerably facilitated by the development of the berylometer, a portable device that detects very small quantities of beryllium in almost any mixture. The device depends on the bombardment of the sample with gamma rays from the radioactive isotope antimony-24. The gamma rays interact

specifically with beryllium atoms to give neutrons that can be counted electronically.

The extraction of beryllium is complicated by the fact that beryllium oxide is found only as a minor constituent, tightly bound to alumina and silica. Treatment with acids, roasting with complex fluorides, and liquid–liquid extraction have all been employed to concentrate beryllium oxide. The oxide is converted to fluoride via ammonium beryllium fluoride and then heated with magnesium to form elemental beryllium. The element is purified by vacuum melting.

Metallic beryllium is grayish, and its chemical properties somewhat resemble those of aluminum. Though resistant to air oxidation under normal conditions, it is readily attacked by acids and alkalies. The applications of beryllium depend upon its unusual combination of physical properties. A very light metal, it has a high mechanical strength and a high melting point. Because of its low atomic weight it has a high transparency to X-rays (17 times greater than aluminum), which makes it ideal as a material for windows of X-ray tubes. Its ability to slow down fast neutrons has found considerable application in atomic piles. Beryllium has been developed as a structural material for specific applications particularly within the aerospace industry. Its market value is about $60 per pound.

Uses of beryllium

Beryllium is used in a number of commercial neutron sources. The alpha particles released by radioactive decay of radium atoms react with atoms of beryllium to give, among the products, neutrons with a wide range of energies—up to about 5×10^6 eV. If radium is encapsulated, however, so that none of the alpha particles reach beryllium, neutrons of energy less than 600,000 eV are produced by the more penetrating gamma radiation from the decay products of radium.

Compounds. Most beryllium compounds are fairly easily converted to the oxide. As expected for the first member of the alkaline-earth group, the hydroxide is only a weak base; its salts are readily hydrolyzed, and the oxysalts, such as nitrate or sulfate, are readily decomposed by heating. Basic beryllium acetate, $Be_4O(C_2H_3O_2)_6$, can, however, be crystallized in a very pure form and serves as a useful source for making pure beryllium compounds. Beryllium oxide is a high-temperature refractory substance with the unusual combination of high electrical resistance and dielectric strength with high thermal conductivity. It finds numerous applications (*e.g.*, as a crucible material).

Inhalation of beryllium compounds can cause an acute respiratory disease or a chronic disease (berylliosis), which is an even greater industrial hazard because it may not appear for periods of from a few months to as long as 20 years after exposure. Skin contact with soluble beryllium salts leads to dermatitis.

Magnesium. Magnesium, comprising about 2.5 percent of the earth's crust, is the eighth most abundant element and, after aluminum and iron, the third most plentiful structural metal; its cosmic abundance is estimated as 9.1×10^6 atoms (Si = 10^6 atoms). It occurs as carbonates (magnesite, $MgCO_3$, and dolomite, $CaCO_3 \cdot MgCO_3$) and in many common silicates, including asbestos, talc, and olivine. It also is found as the hydroxide (brucite), chloride (carnallite, $KCl \cdot MgCl_2 \cdot 6H_2O$), and sulfate (kieserite). About 6,600,000 tons of magnesium are contained in each cubic mile of sea water, which is therefore an almost inexhaustible source.

Magnesium in sea water

Magnesium is produced chiefly by the electrolysis of molten magnesium chloride and by the thermal reduction of magnesium oxide with ferrosilicon (see MAGNESIUM PRODUCTS AND PRODUCTION). The world production of magnesium metal was 207,000 tons in 1967. Its market price is around 36 cents per pound.

It is the lightest metal that can be commonly used for structural purposes (its density is 65 percent that of aluminum and 22 percent that of iron), a property widely exploited. Its thermal and electrical conductivity and its melting point are very similar to those of aluminum. Whereas aluminum is attacked by alkalies but is resistant to most acids, magnesium is resistant to most alkalies but is readily attacked by most acids (chromic and hydrofluoric acids are important exceptions). At normal temperatures it is stable in air and water due to the formation of a thin protective skin of oxide (but burns rapidly when heated in air), and it is attacked by steam. Magnesium is a powerful reducing agent and is used to produce other metals from their compounds (*e.g.*, titanium, zirconium, and hafnium). It reacts directly with many elements; with organic halides in ether solution it forms Grignard reagents. Because of its ready combustibility, magnesium finds application in explosive and pyrotechnic devices.

Uses of magnesium

Magnesium is mainly used in the form of alloys, which usually contain 10 percent or less of other elements, generally added to increase the strength of the metal. The most important alloys are those of aluminum and zinc. Casting, rolling, extruding, and forging techniques are all employed with the alloys, and further fabrication of the resulting sheet, plate, or extrusion is carried out by normal forming, joining, and machining operations. Magnesium is the easiest structural metal to machine and has often been used when a large number of machining operations are required.

Compounds. Magnesium oxide, MgO, is produced by calcining (heating to a high temperature) the carbonate or hydroxide; it is used as a refractory in steel and other furnaces. The sulfate finds application in a number of industrial processes including tanning and dyeing, and it is employed as a treatment for magnesium deficiency in soils. Magnesium compounds such as the sulfate (Epsom salts) and hydroxide (milk of magnesia) are used medically as laxatives and purgatives.

Magnesium is essential to all living systems. The photosynthetic function of plants depends upon the action of chlorophyll pigments, which contain magnesium at the centre of a complex, nitrogen-containing ring system (porphyrin). These magnesium compounds enable the energy of light to be used to convert carbon dioxide and water to carbohydrates and oxygen and thus directly or indirectly provide the key to nearly all living processes. Magnesium also takes part in a number of enzyme reactions, which control energy transfer in living cells.

Calcium. Calcium is the fifth most abundant element in the earth's crust, to which it contributes an estimated 3.64 percent; its cosmic abundance is estimated at 4.9×10^4 atoms (Si = 10^6 atoms). One of the most widely distributed elements, it occurs as carbonate (chalk, limestone, marble, calcite), sulfate (anhydrite, gypsum), fluoride (fluorite or fluorspar), and phosphate (apatite). It is also found in a large number of silicates and aluminosilicates, in salt deposits, and in natural waters, including the sea. Calcium carbonate deposits dissolve in water that contains carbon dioxide to form calcium bicarbonate, $Ca(HCO_3)_2$. This process often results in the formation of caves, and may reverse to deposit limestone as stalactites and stalagmites. Calcium is essential to plant and animal life. Many living organisms concentrate calcium in their shells or skeletons, and indeed in higher animals calcium is the most abundant inorganic element. Many important carbonate and phosphate deposits owe their origin to living organisms.

Natural sources of calcium

The metal is produced by thermal reduction of lime with aluminum under high vacuum and by electrolysis of fused calcium chloride (see CALCIUM PRODUCTS AND PRODUCTION). It reacts with water and, on heating, with oxygen, nitrogen, hydrogen, halogens, boron, sulfur, carbon, and phosphorus; its commercial applications depend largely on these reactions. Although it compares favourably with sodium as a reducing agent, calcium is more expensive and less reactive. In many deoxidizing, reducing, degasifying, and alloying applications, however, calcium often is preferred because of its lower volatility. Small percentages of calcium are used in many alloys for special purposes. The annual world production of calcium metal is about 100,000 pounds, and its market price about $2 per pound.

Compounds. Calcium oxide, CaO (lime, quicklime), is a white solid produced by decomposing the carbonate in

large kilns at temperatures below 1200° C. On reaction with water it slakes to form the hydroxide, $Ca(OH)_2$, which is widely used as a raw material in the chemical industry. The carbonate, $CaCO_3$, is used on a large scale in the cement, glass, and alkali industries and in agriculture to neutralize acid soil.

Slaked lime

The sulfate (as uncalcined gypsum) is employed as a soil corrector. Calcined gypsum is used in making tile, wallboard, lath, and various plasters. Plaster of Paris, the hemihydrate, $CaSO_4 \cdot \frac{1}{2}H_2O$, is produced by partial calcination at about 120° C; mixed with water, it forms a plastic mass that hydrates to a hard white plaster.

The dihydrogen sulfite, $Ca(HSO_3)_2$, is made by the action of sulfur dioxide on a slurry of $Ca(OH)_2$. Its aqueous solution under pressure dissolves the lignin in wood to leave cellulose fibres and thus finds considerable application in the paper industry.

The phosphates are the principal minerals for the production of phosphate fertilizers and for a whole range of phosphorus compounds. The rock is usually treated with sulfuric acid to form $Ca(H_2PO_4)_2$, which may be applied directly to the land. The fluoride, CaF_2, is important to the production of hydrofluoric acid, which is made from CaF_2 by the action of sulfuric acid.

The chloride, $CaCl_2$, occurs in nature but is mainly obtained as a by-product of the ammonia–soda (Solvay) process of manufacturing sodium carbonate. It is very soluble in water and, with ice, produces an excellent freezing mixture, obtaining temperatures as low as −50° C. The solution is used as a refrigerating brine. It also is applied to roads to prevent ice formation, and, in a partly dehydrated form, it is used as a drying agent. The hypochlorite, widely used as bleaching powder, is produced by the action of chlorine on slaked lime. The hydride, CaH_2, formed by the direct action of the elements, provides a useful source of hydrogen when treated with water.

Calcium carbide

The carbide, CaC_2, is produced on a mass scale by the reaction between carbon and CaO in large electric furnaces at temperatures of 1800° C to 2100° C. It is used for the production of acetylene by reaction with water and for the production of calcium cyanide, $Ca(CN)_2$, by heating with nitrogen.

Strontium. Strontium comprises about 0.025 percent of the earth's crust; its cosmic abundance is estimated as 18.9 atoms (Si = 10^6 atoms). Although it is widely distributed with calcium, there are only two principal ores of strontium alone, celestite ($SrSO_4$) and strontianite ($SrCO_3$). Metallic strontium may be obtained by the electrolysis of strontium chloride or by reduction of the oxide with aluminum, but there is little commercial demand for the metal because calcium and barium are more readily obtainable and fulfill the functions for which strontium might be used. The price of the metal is around $8 per pound.

Compounds. The chemistry of strontium is quite similar to that of calcium. Strontium hydroxide, $Sr(OH)_2$, can be used in the extraction of sugar from molasses since it forms a soluble saccharide from which the sugar may be regenerated by the action of carbon dioxide. Strontium nitrate, $Sr(NO_3)_2$, and strontium chlorate, $Sr(ClO_3)_2$, are used in various pyrotechnic devices, flares, and tracer bullets because they produce an intense red colour. Strontium monosulfide, SrS, is used in some luminous paints and as a depilitory. The biological properties of strontium are also very close to those of calcium and distinct from those of barium, whose soluble compounds, for example, are poisonous.

Hazards of strontium-90

The artificial isotopes strontium-89 (52-day half-life) and more especially strontium-90 (28-year half-life), found in radioactive fallout, are extremely hazardous. They become deposited in bones, where they replace calcium; their radiation then damages bone marrow, impairs the process of forming new blood cells, and may induce cancer.

Barium. Barium occurs to the extent of about 0.45 percent in the earth's crust; its cosmic abundance is estimated as 3.7 atoms (Si = 10^6 atoms). Two minerals are common: the more important sulfate ($BaSO_4$, barite, barytes or heavy spar), used principally in oil-drilling "mud," and the carbonate ($BaCO_3$, witherite). The metal may be produced by electrolysis of barium chloride, but the most effective method is the reduction of the oxide by heating with aluminum or silicon in a high vacuum. Only a few tons of the metal are produced each year; the market price is around $9 per pound. It is used chiefly as a "getter" in vacuum tubes, where it removes the last traces of various gases.

Compounds. The chemistry of barium is rather similar to that of calcium and strontium; one important distinction is the increased solubility of the hydroxide in water. Most barium compounds are produced from the sulfate via reduction to the sulfide, BaS. A number of uses of these compounds depend upon the ready formation of the highly insoluble sulfate. Thus barium carbonate, $BaCO_3$, which may be made by the action of carbon dioxide on barium sulfide, is used to remove sulfate from salt brines before they are fed into electrolytic cells (for the production of chlorine and alkali). Further, barium chloride is used in the preparation of white leather, in which white barium sulfate is precipitated in the leather by reaction with a solution of sodium sulfate. Barium sulfate is widely used as a filler (in paper, rubber, etc.) and finds an important application as an opaque medium in the X-ray examination of the gastrointestinal tract. Lithopone, a mixture of barium sulfate and zinc sulfide, is a brilliant white pigment. Barium nitrate, $Ba(NO_3)_2$, has pyrotechnic applications somewhat similar to those of strontium nitrate except that a green colour is produced. The peroxide, BaO_2, is formed by heating the oxide in air and has been used for both oxygen production and as a source of hydrogen peroxide. All soluble barium compounds are poisonous, as is barium carbonate, which will dissolve in the hydrochloric acid of the stomach; the antidote is a solution of sulfate.

Uses of barium sulfate

Radium. About 16 isotopes of radium, all radioactive, are known; their half-lives, except for radium-226 (1,600 years) and radium-228 (5.8 years), are less than a few weeks. The long-lived radium-226 is found in nature as a result of its continuous formation from uranium-238 decay. Radium thus occurs in all uranium ores, but it is more widely distributed because it forms water-soluble compounds; the earth's surface contains an estimated 1.8×10^{13} grams of radium.

The discovery of radium

The story of the discovery and isolation of radium by Pierre and Marie Curie (1898) is well known. The Curies observed that the major part of the radioactivity of pitchblende was caused by a previously unknown substance that could be concentrated with barium and then separated from it by laborious fractional crystallization. The separation was followed by the increase in intensity of new lines in the ultraviolet spectrum and by a steady increase in the apparent atomic weight of the material until a value of 225.18 was obtained, remarkably close to the accepted value of 226.03.

In modern technology, radium is separated from barium by fractional crystallization of the bromides, followed by purification by ion-exchange techniques for removal of the last 10 percent of the barium. Radium metal may be prepared by electrolytic reduction of its salts.

Radioactivity. The chemistry of radium is what would be expected of the heaviest of the alkaline earths, but the intense radioactivity is its most characteristic property. One gram of radium-226 undergoes 3.7×10^{10} disintegrations per second, producing energy equivalent to 6.8×10^{-3} calories, sufficient to raise the temperature of a well-insulated sample at the rate of 1° C every 10 seconds. The practical energy release is even greater than this, due to the production of a large number of short-lived radioactive decay products. Some radium compounds suffer considerable radiation damage. Radium iodate, $Ra(IO_3)_2$, for example, turns brown from the production of free iodine, and peroxides and oxygen are formed from it in aqueous solutions. The alpha particles emitted by radium may be used to initiate nuclear reactions; a mixture of beryllium (see above *Beryllium*) and radium is used as a

neutron source. Radium mixed with a phosphor that could be excited by the alpha particles was widely used in the manufacture of luminous paints for watch and instrument dials since the early 1950s, but less hazardous alpha emitters have largely replaced it. Although radium-226 is not a very intense gamma emitter, its decay products are; thus radium has been used in medicine as a source of gamma rays for the irradiation of tumours. One of the products of radium decay is radon, the heaviest noble gas; this decay process is the chief source of that element.

Radium tends to concentrate in bone where alpha radiation interferes with red corpuscle production and can lead to bone cancer. The hazards of radium were disastrously emphasized by the deaths of a number of women who had been employed painting luminous dials during World War I and who had ingested considerable amounts of radium through the habit of licking the points of their brushes. The detection of exhaled radon provides a very sensitive test for radium absorption.

BIBLIOGRAPHY. The occurrence, properties, and uses of the individual alkaline-earth elements and their more important compounds are given in C.A. HAMPEL (ed.), *The Encyclopedia of the Chemical Elements* (1968). The general comparative and theoretical aspects of alkaline-earth chemistry are discussed in most modern textbooks of inorganic chemistry and particularly in C.S.G. PHILLIPS and R.J.P. WILLIAMS, *Inorganic Chemistry*, 2 vol. (1965–66).

(C.S.G.P.)

Alkaloids

Alkaloids, a specific class of chemical compounds found in plants and, occasionally, animals, are notable chiefly for their physiological activity; many have long histories as poisons, narcotics, hallucinogens, and medicinal agents. Generally, alkaloids are basic, or alkaline, substances—*i.e.*, they neutralize acids; their molecules contain chiefly atoms of carbon, hydrogen, and nitrogen, which is the source of their basicity.

Historical background. During the long period when the early art of healing involved magical or religious beliefs, the use of minerals, plants, and animal organs as medicine was not necessarily related to any active constituents they might contain. Not until alchemy evolved into chemistry and medicine became more scientific was it apparent that the medicinal virtues of plant and animal tissues had to depend upon specific compounds in the mixture of components contained therein. Chemists were challenged to separate the constituents and determine by experiment which activities could be attributed to them.

In this regard no drug known to the ancients offered modern research more of a challenge than opium. The first recorded isolation of a crystalline constituent from opium, now thought to have been the alkaloid called narcotine, took place in 1803. In the following year another chemist described what was probably morphine, and a few years later a German chemist purified morphine and recognized its nature as a base. Soon a number of similar substances were recognized and grouped into a new class of compounds known as vegetable alkalies; these later became known as alkaloids.

First isolation of an alkaloid

Progress in the study of the pharmacological properties of alkaloids was stimulated by their potent physiological effects that suggested worthwhile medicinal uses. The chemical study of them, however, was slower and depended upon the growth of organic chemistry, for not until simple organic bases (and acid–base reactions generally) were well understood could the intricate structures of the alkaloids be unravelled.

The first major step came with the study of coniine, an alkaloid from hemlock (a poison known to the ancient Greeks). It had been isolated as early as 1827 and recognized as a derivative of pyridine, an organic base obtained from coal tar. The synthesis of coniine was achieved in 1886—the first synthesis of an alkaloid.

General classification and nomenclature. It is difficult to categorize alkaloids in precise terms. In general it is agreed that an alkaloid must be a product of plant or animal metabolism and must contain at least one nitrogen atom that can be active as a base in acid–base reactions. There are exceptions to this, however—the substance called rutaecarpine, which contains three nitrogen atoms, none of which is basic, is nevertheless classified as an alkaloid because its reactions and the method by which it is synthesized in the plant (biosynthesis) are typical of alkaloids. At the same time, the nitrogenous bases known as purines, of which caffeine is a good example, though they used to be, are no longer treated in works on alkaloids. On the other hand, the substance triacanthine, which belongs to the purine family, is frequently classified as an alkaloid. In principle, alkaloids could be classified either by the biological systems in which they occur or by their chemical structures. In practice both methods are used. For example, it is common to speak of the opium alkaloids because they occur in the opium poppy (*Papaver somniferum*); but it is also common practice to speak of those alkaloids that contain a ring system called indole as the indole alkaloids. This dual classification system actually produces little confusion because there is a rough correlation between the chemical types of the alkaloids and their biological distribution (see below *Biological distribution*).

Basis of alkaloid names

Because most alkaloids were isolated from natural sources before their chemical structures had been determined, they generally bear coined names, which do not denote their chemical structures. Such names may reflect the names of the discoverer (*e.g.*, pelletierine, after the 19th-century French chemist Pierre-Joseph Pelletier), the botanical source (strychnine, isolated from *Strychnos* species), or the biological effect (emetine, an emetic). As a rule, alkaloid names end in -ine, a reference to their chemical classification as amines (see above). In certain cases, alkaloid names are derived in whole or in part from systematic nomenclature systems used in organic chemistry (see CHEMICAL COMPOUNDS, ORGANIC). These same systems are encountered in the names of compounds derived from the chemical breakdown, or degradation, products of the alkaloids and synthetic substances related to them. Most alkaloids have one or more of their nitrogen atoms as part of a ring of atoms, frequently called a cyclic system.

In this article alkaloids are grouped into several categories for discussion and, necessarily, many of them appear in each category. These categories are: physiological activity, general chemistry, distribution, classification by structure, and general biology.

GENERAL PHYSIOLOGICAL ACTIVITY

The physiological activity of alkaloids is of importance not only in medicine but also in agriculture and in forensic chemistry. Narcotics addiction and the use of alkaloids as hallucinogens are major social problems. In medicine, alkaloids are employed as narcotics, analgesics; antimalarials; local anesthetics; as cardiac, uterine, and respiratory stimulants; and as materials that raise the blood pressure; cause the pupils to dilate; or bring about relaxation of the skeletal muscles.

Grouping by physiological effects. A comprehensive survey cannot be made here and, in the following sections, only the most important of the alkaloids are discussed.

Poisons. Many alkaloids are of medical importance because they are the cause of occasional poisoning of livestock or man. For example, the senecio alkaloids, and their close relatives isolated from *Crotalaria* and *Heliotropium* species, are only mildly toxic but, in repeated small doses, they cause cirrhosis of the liver. The disease is common in livestock in areas where these plants grow, and it has been variously named in the different areas. Another example is the group of ergot alkaloids, produced by a fungus (ergot) that grows on cereal grains; it finds legitimate use in medicine, but the ingestion of cereals containing ergot was the source of much serious illness (ergotism) until the cause was well understood. Not everything is known about the chelerythrine-type alkaloids and there has been a suggestion that the ingestion of even very small amounts can induce glaucoma. The alkaloids of the horsetails (*Equisetum*) have come under suspicion in instances of livestock poisoning.

Ergot alkaloids

Analgesics. The narcotic of major importance, both in legitimate medicine and in illicit drug traffic, is morphine. Although there are now certain synthetics that substitute for morphine in the relief of pain, this alkaloid is still the most suitable drug for that purpose. Were it not for its addictive properties and the fact that it is sometimes not tolerated well, morphine would be the analgesic and narcotic of choice. Codeine, the methyl ether derivative of morphine, is an excellent analgesic essentially devoid of addictive effects. Diacetylmorphine, best known as heroin, is more powerful as an analgesic but is a strongly addictive drug, which induces in the addict an intense euphoria not obtained with morphine.

Cardiac stimulants. The use of heart stimulants to restore the actions of a malfunctioning heart to normal does not necessarily remedy the causes that underlie the malfunction. The cinchona alkaloid quinidine is the cardiac stimulant used to correct arrhythmias of the auricle (one of the chambers of the heart), and cinchona and rauwolfia alkaloids are used for arrhythmias of the ventricle (the other chamber of the heart). There are no generally acceptable alkaloids for congestive heart failure, in which the pumping action of the heart is inadequate though its rhythm is normal. Generally, another type of material, the digitalis glycosides, is the drug of choice for this kind of heart disease.

Respiratory stimulants. Respiratory stimulants and depressants act on the body by several different mechanisms; owing to the complex nature of these effects, stimulation is often followed by depression, with the result that large doses of stimulants can lead to death by respiratory arrest. As typical respiratory stimulants, nicotine and cytisine show this effect, but they have been used clinically, nonetheless. The alkaloid lobeline is safer in this respect, because eventual arrest of respiration is not as likely, but some alkaloids closely related to lobeline are even more toxic than nicotine. Coniine, too, stimulates respiration, but depression and asphyxia soon follow. Coniine is the main alkaloid from poison hemlock (*Conium maculatum*) and was surely the major ingredient in the hemlock used to kill Socrates. It has been surmised that that poison also contained opium.

Nicotine

Many alkaloids influence respiration; virtually all of them, however, produce other, and often unwanted, effects. Atropine, for example, stimulates respiration in moderate doses, even when respiration has been depressed by morphine, but it also has a number of effects on the brain and dilates the pupils of the eyes.

Certain alkaloids find some use in medicine because they are capable of raising blood pressure. This action can take place in a variety of ways, such as by constriction of the blood vessels or stimulation of the heart or by both at once. Other effects, such as dilation of the pupils, inhibition of gastrointestinal peristalsis, and excitation of the intestines and the uterus, are common.

Blood vessel constrictors. Ergonovine, one of the ergot alkaloids, finds extensive use in obstetrics to reduce uterine hemorrhage following childbirth, its action being primarily constriction of the blood vessels. Ephedrine also causes constriction of blood vessels and, because of this action, it is widely used to alleviate the discomfort of common colds, sinusitis, hay fever, and bronchial asthma. Certain synthetic agents have now partially replaced the drug for this use. The pupil-dilating action of ephedrine, unlike that of atropine, does not abolish the light reflexes and the accommodation reflexes of the eye. There are a number of other useful dilators and, of these, scopolamine is one of the more powerful. Cocaine, a potent local anesthetic, is also a dilator.

Ephedrine

Local anesthetics. Many alkaloids possess local anesthetic characteristics, and some of them do not produce the undesirable effects of cocaine. Extensive experimental modifications of the molecular structure of cocaine have yielded a few compounds that are more effective and less toxic than the parent alkaloid as well as many that are less effective. Modern synthetic local anesthetics, cheaper and often superior to cocaine, have largely, but not entirely, displaced the alkaloid. Although they are not used extensively, the following alkaloids possess acceptable local anesthetic characteristics: derivatives of certain cinchona alkaloids, the erythrophleum alkaloids, an alkaloid from *Fritillaria sewerzowii*, and several derivatives of cotarnine.

It is remarkable that only the cinchona alkaloids, especially quinine, have gained and retained prime importance in the treatment of malaria. Extensive modifications of the molecular structures of these alkaloids also have yielded no superior products. An apparent exception to the first generalization is febrifugine, an alkaloid first isolated from *Dichroa febrifuga*, which is a potent antimalarial, especially against bird and monkey infections. Its curative dose, however, is near its toxic dose, and it is not suitable for use in humans. A number of entirely synthetic compounds, which are now available, however, allow a degree of malaria control that could not be attained with any of the alkaloids.

Cinchona alkaloids

Muscle relaxants. Curariform, or curare-like, activity —that is, relaxation of muscles similar to that produced by the South American arrow poison, curare—is important in medicine, particularly surgery. Useful synthetics are available for this purpose, but the alkaloids of South American species of *Strychnos* and *Menispermum* are much used. These alkaloids, which are the active ingredients of the native curare, are quaternary salts of bisbenzylisoquinolines, the most active being the substances called toxiferine and tubocurarine. They are effective as muscle relaxants only when given subcutaneously or intravenously. On the other hand, dihydro-β-erythroidine, one of the most active tertiary bases from *Erythrina* species, acts as a muscle relaxant only when taken orally.

An important consideration in the therapeutic use of alkaloids is that the physiological actions are generally rather complicated; it is rare that an alkaloid exerts only one type of physiological response. As a result of the study of the relationships between the structural and physiological aspects of alkaloids, it has been possible to direct research into the syntheses of model compounds, which show amplification of one type of activity at the expense of another. It is thus possible in some cases to accentuate desirable effects and minimize unwanted or toxic ones.

Psychedelics. Many of the so-called psychedelic drugs are alkaloids or their derivatives. Those derived from cacti of the various *Anhalonium* species have long been known and are used in the rituals of some American Indian tribes. Mescaline, the chief of these, causes mainly visual hallucinations. More potent is psilocybin, an alkaloid found in *Psilocybe* species of fungi. Still more potent is LSD, the diethylamide of natural lysergic acid—an alkaloid from ergot—which produces a variety of effects. The dimethylamide of lysergic acid, which is found in some morning-glory (*Convolvulus*) species, is less active.

Mescaline and LSD

Fate of alkaloids in the body. When a toxic substance is presented to a body it may be dealt with in one of three ways: (1) it may be eliminated unchanged in the exhaled breath, the urine, or the feces; (2) it may be chemically altered to a less toxic substance; or (3) it may be conjugated—that is, combined with another substance—to give a product more readily eliminated or tolerated. As an example, morphine is excreted in all three ways: unchanged, chemically altered, and in conjugated form. The conjugate—of unknown nature, though possibly a combination with a sugar molecule—can be hydrolyzed (broken down) to morphine by heating with acid, and it can thus be assayed as morphine following such treatment. For reasons not yet understood, addicts excrete a larger proportion of the conjugate than do nonaddicts. When morphine labelled with carbon-14 in the methyl group attached to the nitrogen atom is given, a portion of the radioactivity appears in the carbon dioxide exhaled from the lungs.

Morphine and derivatives

The fate of heroin has not been studied as extensively as morphine, but it is known that there are enzymes in the body capable of removing either or both of the acetyl groups. Since heroin is a more potent narcotic than morphine, it has been assumed that its enzymatic conversion to morphine does not account for its activity. It is pos-

sible, nevertheless, that the acetylated substance (heroin) can achieve easier access to the site of its action, where it is then hydrolyzed to the active material.

Small amounts of ingested or injected cocaine are recoverable from urine, but the fate of most of it is not known. When *N*,*N*-dimethyltryptamine, a naturally occurring psychotomimetic (psychosis-mimicking) drug, is ingested, it is metabolized rapidly and excreted chiefly as indoleacetic acid. At the same time the urine is enriched with 5-hydroxyindoleacetic acid, which may be derived from oxidation of indoleacetic acid or from increased metabolism of serotonin, a substance in the brain that plays a role in the transmission of nerve impulses and may be involved in the action of psychotomimetic drugs.

BIOLOGICAL DISTRIBUTION

The distribution of alkaloids in the plant kingdom is largely confined to the flowering plants, though certain nonflowering forms also produce alkaloids; *Lycopodium*, or club moss, is particularly rich in alkaloids, and a fungus, *Rhizoctonia leguminicola*, produces the alkaloid slaframine. The scent glands of the North American beaver (*Castor canadensis*) contain the alkaloid castoramine, which is structurally related to the nuphar alkaloids found in the yellow pond lilies (*Nuphar* species). The skin glands of a salamander, *Salamandra maculosa*, contain a group of alkaloids that are related to the natural materials called steroids, of which cholesterol is the best known example. It is likely that these alkaloids are generated by the animal from steroids in its diet.

Plant families rich in alkaloids

Plants of a number of families are particularly rich in alkaloids. All plants of the poppy family (Papaveraceae) are thought to contain them. In the buttercup family (Ranunculaceae) alkaloids are confined to a limited number of genera. Among the other families containing alkaloids are barberry (Berberidaceae), amaryllis (Amaryllidaceae), pea (Fabaceae), dogbane (Apocynaceae), moonseed (Menispermaceae), custard-apple (Annonaceae), and nightshade (Solanaceae).

The types of alkaloids produced are often characteristic of a particular family or genus. All alkaloids so far obtained from the poppy and custard-apple families are isoquinoline alkaloids (see *Classes of Alkaloids* below for structures and descriptions of this and other alkaloid classes). Although the alkaloids found in plants of the buttercup family are of several types, those from the monkshoods (*Aconitum*) and larkspurs (*Delphinium*) are all of only one type, the aconite alkaloids, and those from the meadow rues (*Thalictrum*) are mostly of another type, the bisbenzylisoquinolines. The cinchonas (*Cinchona*) in the madder family (Rubiaceae) yield mainly the quinine type of alkaloid, whereas the poison nuts (*Strychnos*) in the related Logania family (Loganiaceae) give rise to the highly complex strychnine type. Almost without exception, the alkaloids from the dogbane family are indole derivatives.

Role of alkaloids in plants

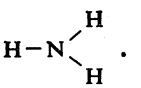
The role that alkaloids play in the life of plants is still largely a matter for speculation. Belladona (*Atropa*) and tobacco (*Nicotiana*) plants produce large quantities of alkaloids, whereas related tomato (*Lycopersicon*) plants do not. When grafts of tomato are grown on stocks of belladona or tobacco, they are almost devoid of alkaloids, whereas the reverse grafts generate fair quantities of them. In the light of these results, the alkaloids would seem to play no essential role in the metabolism of the plants that produce them. That they play some role is suggested by the fact that in many plants prior to seed formation there is a high concentration of alkaloids, which is almost entirely depleted when the seed is ripe. An example of this behaviour is found in the opium poppy (*Papaver somniferum*). In some plants, such as the larkspurs (*Delphinium*), both the seeds and the growing plant are rich in alkaloids. In others (for example, the sweet shrubs [*Calycanthus*]) the seeds are particularly rich, whereas the growing plants contain only traces of alkaloids.

It is thought that the alkaloid content of some plants may protect them from depredation by insects, but there are few, if any plants—even those rich in alkaloids—that are avoided by all insects. *Solanum demissum* seems to be avoided by the larvae of the potato beetle, and several relatively unsuccessful attempts have been made to introduce the genes that control this immunity from this plant into other nightshade (*Solanum*) species, especially the potato. Several species of alkaloid-containing groundsel (*Senecio vulgaris* and *S. jacobaea*) are the preferred and possibly the only food for the larvae of the cinnabar moth (*Callimorpha jacobaea*). These larvae have an even higher concentration of alkaloids than the plants on which they have fed, and they are not acceptable as food to a wide variety of vertebrate predators that prey on larvae of related species. It is presumed that the lack of acceptance by predators is a consequence of the alkaloid content.

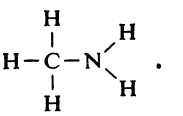
GENERAL CHEMISTRY

Molecular structure. The structures of the various alkaloid molecules can be understood best by comparison with simpler compounds of nitrogen. In its compounds the nitrogen atom generally forms chemical bonds with three other atoms. In chemical terms, this is to say that nitrogen has a valence of three. For example, the valence bonds of nitrogen in the simple substance ammonia are completed by hydrogen atoms (symbol H, valence one). In representations of molecular structure these bonds are symbolized by lines. Thus, ammonia is denoted by

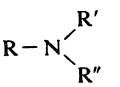
Structures of nitrogen compounds



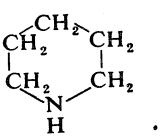
Carbon atoms (symbol C, valence four) also can be joined to the nitrogen atom, giving molecules like that of methylamine, the formula of which is CH_3NH_2:



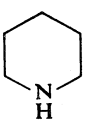
Other combinations of carbon and hydrogen atoms in hydrocarbon chains can be joined to the nitrogen atom, giving the large family of compounds known as amines and represented by the general formula



in which R, R′, and R″ are various hydrocarbon groups that may be considered as replacing the hydrogen atoms of ammonia, NH_3; replacement of one, two, or all three of the hydrogen atoms gives primary, secondary, or tertiary amines, respectively. In some amines the hydrocarbon chains are so joined as to form complete rings of atoms. A simple example of this is the substance piperidine, which has the structure

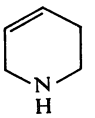


Structures of this type usually are represented diagrammatically by regular polygons in which the corners are understood to signify carbon atoms bearing as many hydrogen atoms as necessary to complete their valences. In this convention, piperidine is represented as



In some hydrocarbons, or hydrocarbon derivatives such as the amines, hydrogen atoms are missing, and the chains or rings are said to be unsaturated or to contain

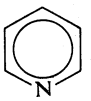
double bonds. Double bonds are indicated in structural diagrams by double lines, as in the formula



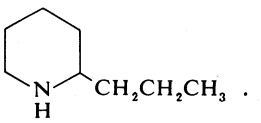
Some unsaturated, cyclic compounds contain series of alternating double and single bonds, which give the molecules special properties. An example is pyridine



Because these compounds, which are called aromatic, do not behave like typical unsaturated compounds, their structures are frequently drawn with circles within the polygons rather than double lines. In this convention pyridine is written



The molecules of the various alkaloids are of a great many different types, which are variations and elaborations of the simple structures discussed above. The alkaloid coniine, for example, is closely related to piperidine (and, therefore, also to pyridine). Its structure is



As indicated, most alkaloids contain at least one basic nitrogen atom—that is, a nitrogen atom in an amine-type structure. Generally, also, alkaloids have this or another nitrogen atom in some type of ring system, either aromatic or nonaromatic (aliphatic). But this is not an absolute requirement, for in the molecules of some alkaloids the nitrogen atom is joined to a simple aliphatic hydrocarbon chain. Besides the obligatory carbon, hydrogen, and nitrogen atoms, many alkaloid molecules also contain oxygen atoms, and a small number contain sulfur or chlorine atoms.

Rings in alkaloid molecules

Isolation, purification, and analysis. Like other bases, most alkaloids yield salts that are soluble in water when treated with acids, and the separation of these alkaloids from neutral and insoluble plant constituents by salt formation is, therefore, relatively straightforward.

Isolation. Special methods have been developed for isolating commercially useful alkaloids such as morphine. In the main, however, plant and animal tissue is processed in one of three ways to obtain aqueous solutions of the alkaloids. (1) The source material, either fresh or dried, is treated (by a method called extraction that involves the dissolving with reagents of only some components of a mixture) directly with a dilute aqueous solution of an acid that reacts with the alkaloid to produce a soluble alkaloid salt. (2) The material is extracted with an alcoholic solvent, with or without the addition of an acid, after which the solvent is removed by evaporation and the residue—consisting of fats, wax, resins, and other material, as well as the alkaloids—is stirred in contact with a dilute aqueous acid solution that forms soluble alkaloid salts. (3) The material is rendered basic by the addition of aqueous ammonia or sodium carbonate; the mixture is extracted with a water-immiscible solvent, such as ether, and finally, this solution is extracted with dilute aqueous acid. The alkaloids are recovered from the aqueous acid solution prepared by one of the above methods by making the solution alkaline and then extracting it with an immiscible solvent. The last process does not yield alkaloids in which the nitrogen is fixed in a saltlike form; such alkaloids must be precipitated from aqueous solution as complex salts.

Since very few plants elaborate only one alkaloid there still remains the problem of separating and purifying the different alkaloids from the mixture of alkaloids obtained by extraction. The number of constituents, and some indication of their relative abundances, can be learned from analyzing the crude mixture using a system of analysis called chromatography. This is a method of separation that takes advantage of the different degrees of adsorption of the substances to be separated on a solid material, such as alumina or silica. When it comes to actual separation of the alkaloids, chromatography again may be used, or attempts may be made to obtain the alkaloids from solution in crystalline form. The list of solvents from which alkaloids can be crystallized is an extensive one. (The following solvents, in order of increasingly favourable solvent action on alkaloids are commonly used: hexane, ether, methanol, acetone and ethyl acetate, carbon tetrachloride, dichloromethane, chloroform, dioxane, and pyridine.) There are, of course, exceptions, in which a particular alkaloid is unusually soluble or insoluble in a given solvent, and these can be used to advantage to separate pure alkaloids from the mixtures.

Methods of separation

Separation. Important methods of separation depend upon formation of salts, which often can be crystallized easily. In this case, also, a wide range of solvents is available, including water, methanol, ethanol, acetone, and various mixtures of these solvents with one another and with others, such as chloroform. Picrate salts (salts formed with picric acid) often are used with low-molecular-weight alkaloids, such as nicotine. Picrates crystallize readily and often have sharp melting points, a property that aids in their identification. Hydrochlorides and hydrobromides (salts formed with hydrochloric and hydrobromic acids, respectively) often are easily crystallized, sparingly soluble, and not difficult to purify. Nitrates and oxalates (formed from nitric and oxalic acids), in a few cases, are crystallized easily, often from alcoholic solvents, but they are frequently hydrated (contain water) and seldom have sharp melting points. Those alkaloids that carry a phenolic hydroxyl group (an –OH attached to an aromatic ring) have acidic properties; they form water-soluble salts in the presence of strong bases and therefore are separated easily from their relatives lacking phenolic groups.

Purification. Distillation and sublimation (vaporization of a solid) in a vacuum are practical methods for separating alkaloids of moderate molecular weight. Countercurrent distribution, a separation technique that depends on repeated partition between an immiscible solvent and an aqueous phase at controlled acidity, is a useful procedure for separating certain alkaloids, but it is laborious unless an automatic instrument for carrying out the process is available.

Determining the amount of individual alkaloids present in a sample is important in the drug trade and in forensic practice. Reliable procedures for the analysis of frequently encountered alkaloids have been developed and are much used.

Structure determination. The molecular formula of an alkaloid, which gives only the kinds and numbers of atoms present in the molecules, can be determined by combustion analysis—that is, by burning the alkaloid to end products, such as carbon dioxide and water, which can be collected and weighed. The functional groups (active groupings of atoms) present must be detected by specific chemical tests, characteristic of each group.

Mass spectrometry. A modern technique known as high-resolution mass spectrometry breaks down the molecule into fragments and accurately measures the mass of each. This not only provides a precise molecular formula but at the same time reveals details of how the atoms are joined in the molecules. Many types of alkaloids have been examined by this technique, and the fragmentation patterns exhibited by the various types are now well understood.

Spectroscopy. Infrared spectra (patterns of absorption of light in the infrared region) give information about the presence or absence of a number of functional groups, chief of which are the carbonyl, hydroxyl, and

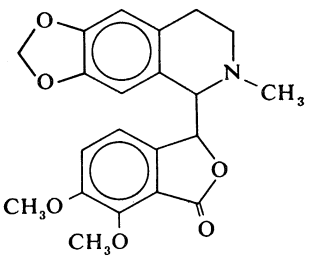
amino groups. Ultraviolet spectra (light absorption in the ultraviolet region), give valuable information about aromatic systems and unsaturated groups conjugated with aromatic rings or carbonyl groups.

Nuclear magnetic resonance. A measure of the nuclear behaviour of the atoms and detailed information on the nature and environment of the hydrogen atoms may be derived by nuclear magnetic resonance (see MAGNETIC RESONANCE); such information frequently tells much about the structure of the molecule as a whole.

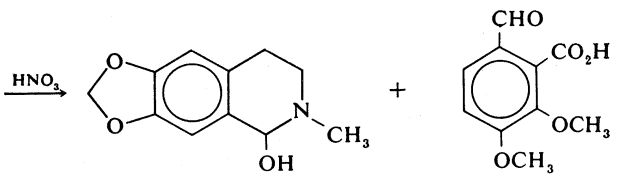
X-ray diffraction. The X-ray is an extremely powerful tool for structure determination. It depends for its success on the interpretation of patterns obtained when X-rays are diffracted by the atoms in crystals. Preferably carried out with a crystalline salt or other derivative containing a heavy atom, such as bromine or iodine, it also has been used with alkaloids containing as elements only carbon, hydrogen, oxygen, and nitrogen. There are several advantages to the method: if suitable crystals can be grown, a complete study can be carried out with very small amounts of material; the method requires no previous knowledge of the functional groups present; and, most important, it can be used in cases in which chemical degradation does not work (see below).

Chemical methods. The structure of an alkaloid can be determined by chemical reactions that break the intact molecule into fragments, the structures of which are already known or can be determined. In ideal cases, all of the carbon and nitrogen atoms can be accounted for in the various fragments, which result either from a single type of chemical assault on the molecule or from the use of a number of breakdown reaction paths. Oxidation of the alkaloid is the most direct attack and, in selected cases, gives enough information to deduce the total structure of the original alkaloid. Upon oxidation with partly diluted nitric acid, the alkaloid hydrastine, for example, generates the known compounds hydrastinine and opianic acid, as shown in the equation:

Oxidation

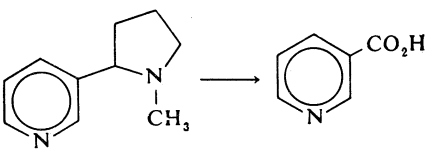


hydrastine



hydrastinine opianic acid

Together these compounds account for all the carbon, nitrogen, and oxygen atoms of the original alkaloid and permit its hypothetical reconstruction. On the other hand, oxidation often destroys much of the molecule. Nicotine, for example, upon oxidation with a variety of oxidizing agents gives mainly nicotinic acid, the remainder of the molecule being converted into small fragments of no value in deducing the structure:

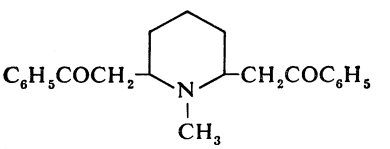


nicotine nicotinic acid

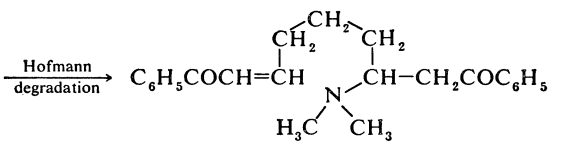
Strychnine (see below) upon oxidation, again with a variety of oxidizing agents, gives a number of oxidation products, all of which provide useful structural information, but only upon further degradation.

To obtain structural information in greater detail, it is convenient to use the method known as the Hofmann degradation. This procedure, developed by a German chemist, August Wilhelm von Hofmann, consists of treating the alkaloid with methyl iodide to convert the basic nitrogen atom into its quaternary, or salt, form; the resulting quaternary iodide is then subjected to strong alkali. Under favourable conditions, any ring containing a nitrogen atom will be opened by this procedure, with the introduction of a double bond into the molecule at the point where the nitrogen atom was attached. At the same time, the nitrogen atom remains part of the molecule as a methylamine derivative. A repetition of the process may generate a second double bond and a new tertiary base, but this occurs only when the nitrogen atom in the original alkaloid is part of two rings. Otherwise the second Hofmann degradation gives a methylamine and a nitrogen-free compound containing two double bonds. These reactions are illustrated by the following equations:

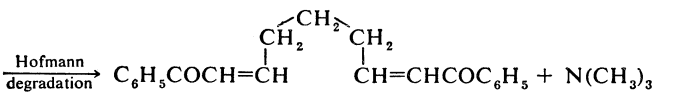
Hofmann degradation



lobelanine



(initial product, one double bond)



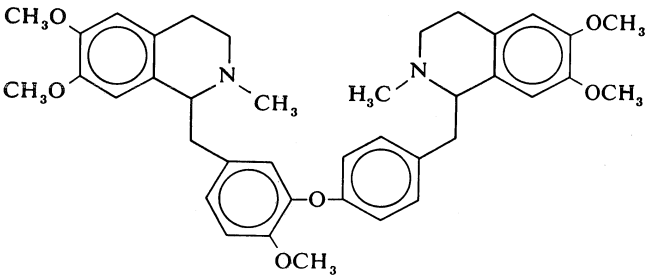
(final product, two double bonds) trimethylamine

If the amine formed in the second-stage Hofmann degradation is trimethylamine, then the original alkaloid must have contained an *N*-methyl group. If the second Hofmann degradation has not eliminated nitrogen from the molecule, a third-stage degradation is sometimes possible, leading to a triply unsaturated carbon compound. There are many alkaloids that do not lend themselves to Hofmann degradation, however, and the reaction may fail at any of the stages. Nevertheless, when it can be employed, this method of degradation is especially useful, because it generates unsaturation in the molecule, which may then open the way to oxidative degradation. A related process achieves the same result as the Hofmann degradation, but with concomitant reduction of the double bond. It is carried out by reacting the methiodide with sodium amalgam or occasionally by reduction with sodium borohydride or with hydrogen in the presence of a catalyst.

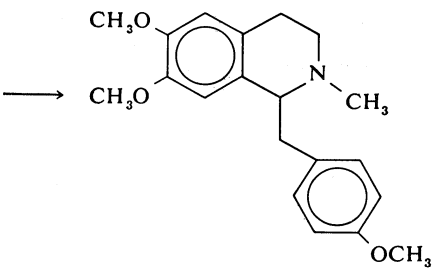
Frequently, Hofmann degradation and subsequent oxidations proceed satisfactorily, but the final product, an acid, is difficult to identify because of its complexity.

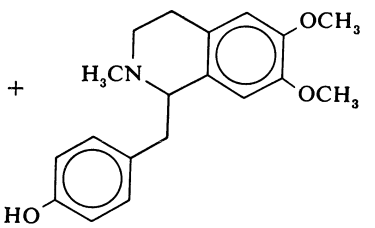
Special cases. *Bisbenzylisoquinolines.* Such is particularly the case with the bisbenzylisoquinoline alkaloids. In addition to relatively simple and identifiable phthalic acids, these alkaloids give rise to extremely complex acids (such as the alkoxyphenylether dibasic acids), which are difficult to identify because their syntheses are so difficult. A reaction is available that eliminates such problems by dividing the molecule into two approximately equal parts at the diphenyl ether junction. For example, when the methyl ether of the bisbenzylisoquinoline

alkaloid dauricine is dissolved in ether or tetrahydrofuran and added to a freshly prepared solution of metallic sodium in ammonia, it gives rise to two simple benzylisoquinolines that are easily identifiable. The cleavage reaction is depicted as follows:



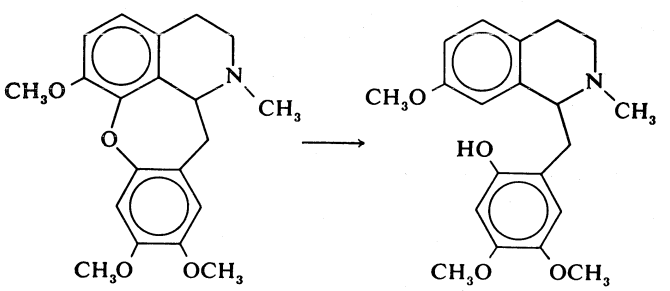
dauricine methyl ether





(easily identifiable benzylisoquinoline products)

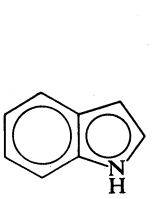
The somewhat simpler alkaloid cularine was the first to be subjected to this reaction, and the product, also a benzylisoquinoline, was relatively easy to identify:



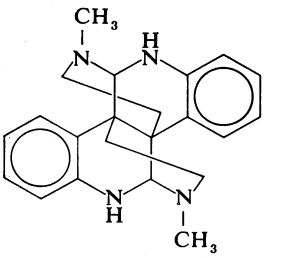
cularine (benzylisoquinoline product)

The results of degradation reactions must always be interpreted with caution.

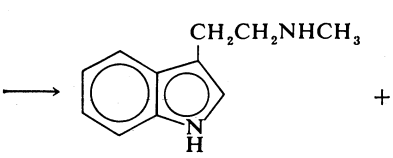
Calycanthine. An example of misleading information obtained by degradation is found with the alkaloid calycanthine. Attempts to degrade the alkaloid by the Hofmann procedure were unsuccessful, but, by chance, it was found that benzoylation of the alkaloid yielded a product that was easily identified as a benzoyl derivative of *N*-methyltryptamine. This strengthened the suspicion that calycanthine was an indole derivative, like *N*-methyltryptamine, as already suggested by a colour reaction, which is given usually only by indole compounds. In a further experiment, however, calycanthine was heated to a high temperature with zinc dust in an inert atmosphere, giving *N*-methyltryptamine, as well as a product that was later named calycanine:



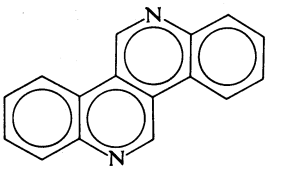
indole



calycanthine



N-methyltryptamine

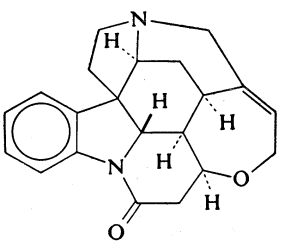


calycanine

Surprisingly enough, the structure of calycanine was shown by synthesis not to be an indole derivative. An X-ray diffraction study of the calycanthine molecule showed that its structure did not include indole nuclei, and it was concluded that the extreme conditions of the pyrolysis, which could have induced great structural changes, had only eliminated molecular fragments, whereas the mild benzoylation reaction had actually brought about deep-seated molecular rearrangements.

Determining the structure of strychnine

Strychnine. A similar, though still more complicated, example of misleading results of degradative experiments is presented by investigations of the strychnine molecule. Elucidation of this extraordinary complex structure of overlapping rings drawn as

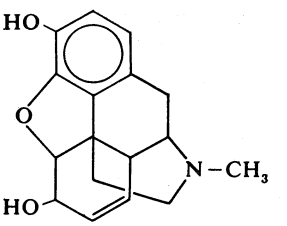


strychnine

(bonds shown - - - extend below the plane of the paper, – above it)

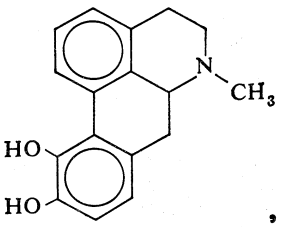
is often cited as a major triumph of the classical techniques of degradative chemistry. For more than a century the structure remained a puzzle. That the structure yielded finally to attack was due chiefly to the skill of a British chemist, Sir Robert Robinson, and an American chemist, Robert Burns Woodward. Strychnine had been subjected at an early date to pyrolysis under a variety of conditions, from which a number of degradation products had been identified. When the structure was finally known, it was recognized that proper juxtaposition of these fragments was sufficient to write the complete structure of the alkaloid.

Morphine. The determination of the structure of morphine, which is



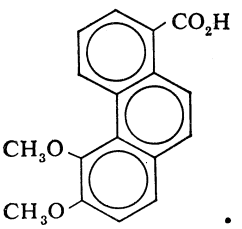
morphine

was thwarted by rearrangements that occurred when the molecule was subjected to chemical attacks. A partly reduced phenanthrene skeleton (composed of three fused six-membered rings) was postulated for the alkaloid because phenanthrene was obtained when the alkaloid was heated with zinc dust. Other degradations led to hydroxyphenanthrenes, and heating the alkaloid to 130° C with strong hydrochloric acid generated the substance called apomorphine, the structure of which,



apomorphine

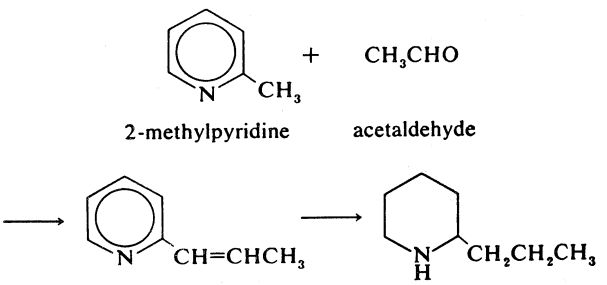
was ascertained by reactions that have since become standard. Furthermore, on Hofmann degradation, the dimethyl ether of the alkaloid gave a vinylphenanthrene, which on oxidation was converted to the corresponding acid, the structure of which is



(acid product)

The structure of this compound was proved by synthesis. Nonetheless, structures of morphine, based upon that of these products—apomorphine, in particular—were untenable because the nature of the rearrangement that gave apomorphine was not understood. A great deal of degradative and synthetic chemistry was necessary before these misleading results could be reinterpreted and the finally accepted structure could be written.

Synthesis. As mentioned above, the first synthesis of an alkaloid was that of coniine. In this synthesis, acetaldehyde was condensed with 2-methylpyridine to give 2-propenylpyridine; on reduction with sodium in ethanol this substance gave synthetic coniine:



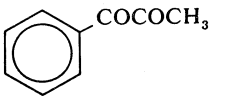
2-methylpyridine acetaldehyde

2-propenylpyridine coniine

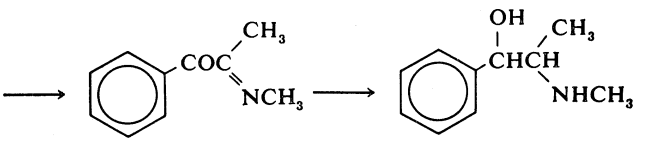
The synthetic product is not quite identical with the natural alkaloid because it does not exhibit the optical activity of that substance. Optical activity is a property of materials that is only detected when plane polarized light (see LIGHT) is passed through a crystal or solution of the substance; the observed effect is the rotation of the plane of polarization of the light (as determined in an instrument called a polarimeter). Optical activity results when molecules are capable of existing in two isomeric forms (see ISOMERISM), which are related to one another like mirror images, or right- and left-hand gloves. These forms are designated *dextro* (*d*) or *levo* (*l*), according to whether they rotate plane polarized light to the right or left, respectively. The molecules of many alkaloids, like those of other biological materials, exist in nature in only one of the optically active forms. Synthetic processes, however, usually produce mixtures of both isomers, which are called *dl*– forms. Separation of the forms, a process called resolution, often can be achieved by crystallizing the *dl*– mixture with another substance already optically active. In the case of coniine, the natural material is the *d*– form, and resolution of the synthetic mixture is carried out by crystallizing the salt formed with *d*–tartaric acid.

Resolutions of *dl*– mixtures

The alkaloid ephedrine, used as a nasal decongestant, has been synthesized by several methods, the most direct of these, a process used commercially, consists in condensation of methylphenylglyoxal with methylamine and reduction without isolation of the initial condensation product. The sequence of reactions is as follows:



methylphenylglyoxal methylamine

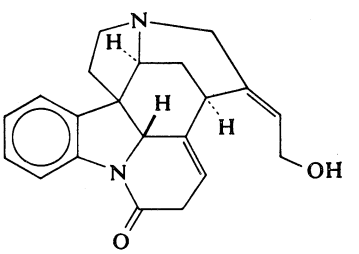


(condensation product) ephedrine

In this case, too, it is necessary to resolve the product into *d* and *l* forms, only one of which (in this case, the *l* form) corresponds to the natural substance. The *l*– and *d*–mandelic acids, respectively, proved to be satisfactory for this purpose, each combining with the oppositely rotating form of the mixture to produce a salt that can be purified by crystallization. Conversely, the *d*– and *l*–ephedrines can be used to prepare pure *l*– and *d*–mandelic acids, respectively. The examples of coniine and ephedrine illustrate the general principle that the synthesis of an alkaloid usually necessitates a resolution of the forms, either as a final or intermediate step.

Synthesis of strychnine

The synthesis of the poisonous alkaloid strychnine posed as great a challenge to chemists as did the determination of its structure and, for the same reason, the extreme complexity of the molecule. The final synthesis of the alkaloid was reported in 1954. Complicating the problem are six asymmetric carbon atoms in the strychnine molecule (the spatial arrangement of which is shown above). No less than 15 sequential reactions were necessary to achieve the synthesis of isostrychnine of the correct configuration:



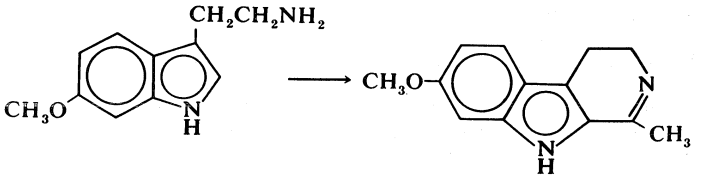
isostrychnine

bonds shown --- extend below the plane of the paper, —above it.

It was necessary to insure that the correct stereochemistry was achieved at each step; in one instance optical resolution of an intermediate was needed. Since isostrychnine can be prepared from strychnine, and since it can be reconverted to the alkaloid, synthesis of it represents a synthesis of strychnine itself.

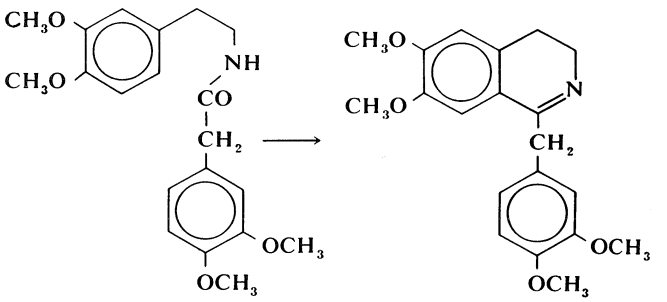
Of chiefly historical interest is the relatively simple alkaloid harmaline; it was the first alkaloid of the indole

type the structure of which was determined, and the first indole alkaloid to be synthesized. Since the initial synthesis, there have been a number of others. Typically these begin with the substance 6-methoxytryptamine:



6-methoxytryptamine harmaline

Quite analogous syntheses of 1-substituted isoquinolines can be carried out beginning with β-arylphenethylamines. A typical example is the preparation of dihydropapaverine from *N*-(3,4-dimethoxyphenethyl)-3,4-dimethoxyphenylacetamide:



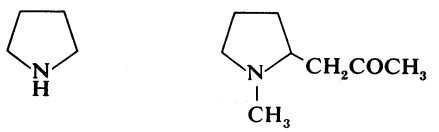
N-(3,4-dimethoxyphenethyl)-3,4-dimethoxyphenylacetamide dihydropapaverine

Conversion of the dihydro- derivative to the alkaloid papaverine by dehydrogenation can be accomplished by a variety of procedures.

CLASSES OF ALKALOIDS

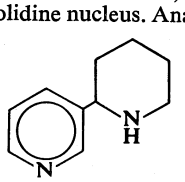
It is convenient to classify alkaloids according to their structural types, taking into consideration the nature and number of their nuclei (rings of atoms) and the substituents attached to these. The simplest, aside from purely aliphatic bases that contain no rings, are the alkaloids the molecules of which include the five-membered pyrrole ring composed of one nitrogen and four carbon atoms. The more complex structures vary considerably in the types of ring systems present and in the number of nitrogen atoms they contain.

Pyrrolidines. The unsubstituted compound, pyrrolidine itself, has been found in the carrot, *Daucus carota*, and in tobacco; its *N*-methyl derivative has also been isolated from tobacco and from belladonna. Hygrine, found in the leaves of the cocoa plant, is among the few other pyrrolidine alkaloids.



pyrrolidine hygrine

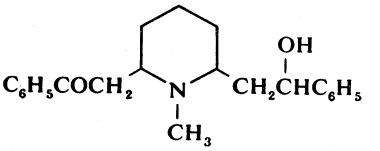
Pyridines. The pyridine alkaloids constitute a large group in which the pyridine nucleus is present in fully aromatic, partly reduced, or fully reduced form. Those alkaloids in which the pyridine ring is fused to a second ring are discussed under other classes. The best known of the pyridine alkaloids is nicotine, which has both a pyridine and a pyrrolidine nucleus. Anabasine



anabasine

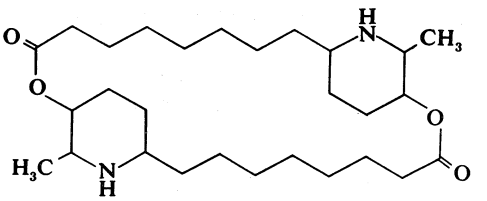
is formed by the union of two pyridine nuclei. The important lobelia and sedum alkaloids, of which there are well over 20, are mostly 2-monosubstituted and 2,6-disubstituted piperidines, with or without a methyl group on the nitrogen atom. The substituents are of the type $-CH_2COR$ or $-CH_2CH(OH)R$, in which R may be methyl ($-CH_3$), ethyl ($-C_2H_5$), or phenyl ($-C_6H_5$). Lobeline

Lobeline



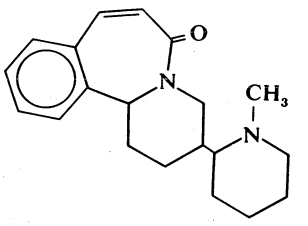
lobeline

is the chief alkaloid of the family used in medicine. The alkaloid carpaine from papaya (*Carica papaya*) had long been given the formula $C_{14}H_{25}O_2N$ and assigned a structure that included a 13-membered lactone ring. Mass spectral data, however, showed that the correct formula is $C_{28}H_{50}O_4N_2$ and that the alkaloid contains a double lactone ring of 26 atoms, as shown in the structural formula:



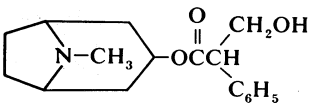
carpaine

There are many possible ways of combining several pyridine nuclei, as well as forming extra rings and adding substituents (see below). Astrocasine from *Astrocasia phyllanthoides* is one example.



astrocasine

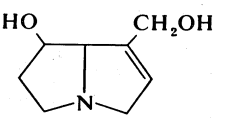
Tropanes. Bridging of the 2,6- position of piperidine (fully reduced pyridine) by a two-carbon unit gives the nucleus of the tropanes. Tropinone (see below: *Emulation of biosynthesis*) is the keto derivative of this nucleus, and tropine is the corresponding alcohol. Esterification of tropine with tropic acid gives the alkaloid atropine, one optically active form of which is hyoscyamine:



atropine (hyoscyamine)

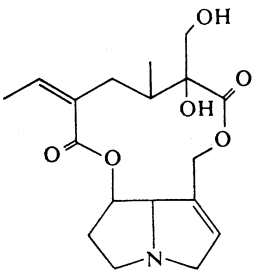
Pyrrolizidines. The ring system formed by the fusion of two five-membered rings with a common nitrogen atom is known as pyrrolizidine. The base itself has not been found to occur naturally but some important derivatives are widespread. The most common and most important of these is retronecine, which occurs in an esterified form in a number of genera in the families Compositae, Leguminosae, and Boraginaceae. These esters are

Retronecine



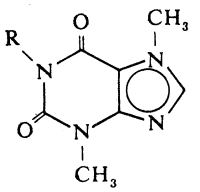
retronecine

commonly referred to as senecio alkaloids because of their abundance in *Senecio* and closely related genera. More often than not esterification is by a dibasic acid, often contaning isoprene units, so that large ring systems may be present. Those alkaloids, such as retrorsine, that are cyclic diesters of acids with branched chains are more toxic than straight-chain diesters. These alkaloids are mutagenic in fruit flies (*Drosophila*) and can cause liver tumours in rats. A number of senecio alkaloids occur in plants as amine oxides, and these are equally toxic. Retrorsine has the structure:



retrorsine

Purines. Although the purine bases (with two rings in the molecules, five- and six-membered, respectively, each including two nitrogen atoms) are nitrogenous compounds elaborated by many different plant species and genera, many of them are not considered alkaloids because of their major biochemical roles. Caffeine, having the following structure,



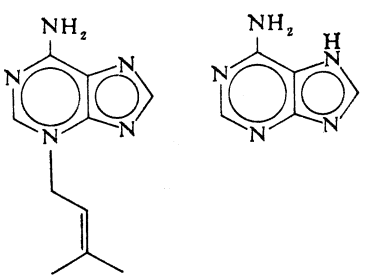
caffeine (R = CH_3)
theobromine (R = H)

however, is generally classified as an alkaloid and is the most widespread of the purine alkaloids. Aside from its presence in tea and coffee, it occurs in a number of species of holly (*Ilex*) and in many other plants.

Caffeine

The stimulating effects of caffeine are well known and doses larger than one gram in man may cause alarming symptoms, though there seem to be no recorded fatalities from overdoses. Dogs are more sensitive to this alkaloid, and poisoning by theobromine, a closely related alkaloid, has been noted often when cacao meal has been used as food for pigs and chickens. Such meal may contain up to 5 percent of theobromine, which can be recovered and converted commercially into caffeine by methylation with methyl chloride in the presence of caustic soda.

The purine nucleus also is present in the alkaloid triacanthine, which occurs in *Gleditsia triacanthos*, and which is remarkable because of its high nitrogen content. It is cleaved by strong acids to adenine, an extremely important component of living cells. Triacanthine and adenine are depicted as:



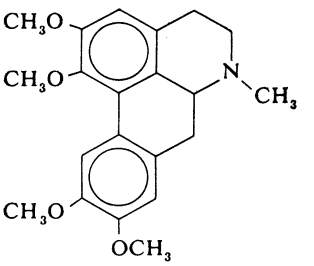
triacanthine adenine

Isoquinolines. The isoquinoline nucleus consists of two fused six-membered rings, one of which contains a nitrogen atom. In most isoquinoline alkaloids there are substituents on the 1- position of the isoquinoline nucleus and the pyridine ring is in its tetrahydro- form. Most frequently the 1- substituent is a benzyl group, although it may be other alkyl, aryl, or arylethyl groups. Furthermore, the 1-benzylisoquinolines undergo a series of transformations, with or without the addition of carbon fragments, to generate various polycyclic types.

The isoquinoline alkaloid classes represented by papaverine, protoberberine, and protopine have already been mentioned. Morphine (see above) and codeine do not, on superficial examination, belong here, but alkaloids found with them in the plants are benzylisoquinolines; they themselves may be transformed into benzylisoquinolines; and they arise biogenetically from benzylisoquinolines. The structures of morphine and codeine were long obscured by the observation that morphine is converted into apomorphine when heated with strong acid. The alkaloid sinomenine has a similar nuclear structure but lacks the oxygen bridge.

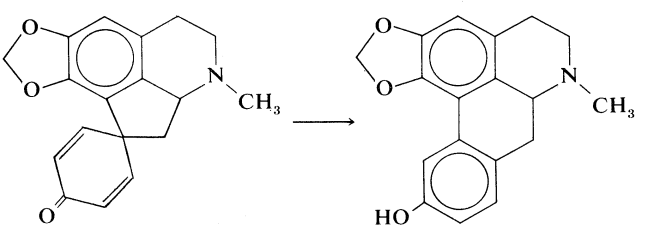
Morphine and codeine

Apomorphine is only one representative of a large variety of alkaloids known as aporphines. These differ from apomorphine only in the position, number, and type of substituents on the nucleus. Glaucine, first isolated from the horn poppy (*Glaucium flavum*) and subsequently found in many other plants, was the first aporphine to be synthesized. Pharmacologically the aporphines are not important. Glaucine has the structure:



glaucine

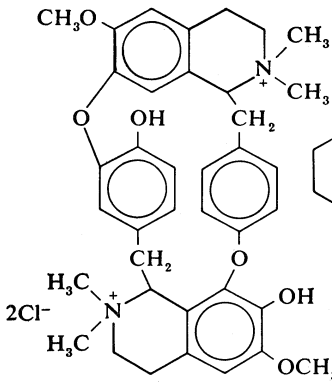
The proaporphine group of alkaloids was more recently discovered. They occur in *Croton* (Euphorbiaceae), *Nelumbo* (Nymphaeaceae), *Stephania* (Menispermaceae), *Ocotea* and *Neolitsea* (Lauraceae), and *Papaver* and *Meconopsis* (Papaveraceae). They are prepared synthetically by phenolic oxidation of benzylisoquinolines, and their biogenetic origin seems to be similar. There is evidence that they are precursors of some aporphines. Mecambrine has the typical nuclear structure of these alkaloids, which differ from one another in the number and position of substituents and the degree of unsaturation of the quinone nucleus. When mecambrine is heated in dilute hydrochloric acid it undergoes rearrangement to mecambroline.



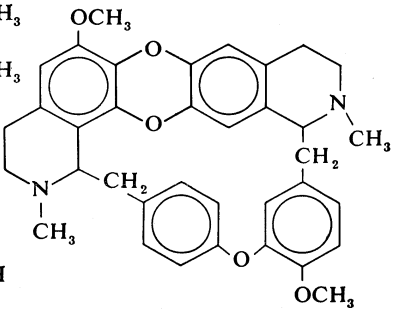
mecambrine mecambroline

The bisbenzylisoquinolines also are generated biologically by oxidation of benzylisoquinolines—in this case, two being incorporated per molecule. These alkaloids also occur in a number of genera belonging to several plant families. The simplest representative is dauricine, found in some *Menispermum* species. There is much variation in the structures of this group of alkaloids. Tubocurarine, already mentioned for its use in surgery, is a quaternary base containing two diphenyl ether linkages in a large ring; isotrilobine is an example of the alkaloids of this family that have three diphenyl ether linkages.

Bisbenzylisoquinolines

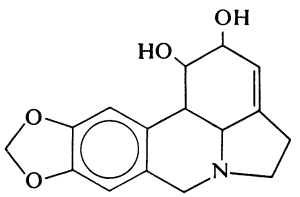


tubocurarine chloride

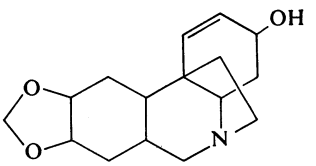


isotrilobine

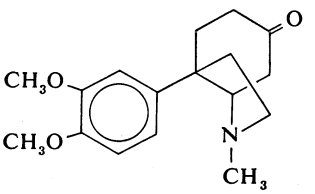
A rather distinct class of isoquinoline alkaloids is represented by the numerous alkaloids of the Amaryllidaceae. Included are a variety of nuclear structures, that of lycorine being representative of one group. Crinine is a variant of this type and an "open" form of this structure is represented by mesembrine, an alkaloid from *Sceletium* species (Ficoidaceae). The last named base still retains the polyhydroindole system present, sometimes in disguised form, in all of these alkaloids, the structural formulas of which are:



lycorine



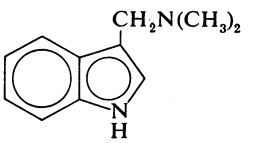
crinine



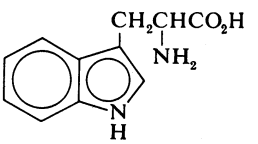
mesembrine

Gramine and strychnine

Indoles. Gramine, a simple indole derivative, and strychnine, a far more complex one, already have been mentioned. The alkaloids of the indole family, without exception, are derived biogenetically from the amino acid tryptophan, and with only one exception, namely that of gramine, they retain the aminoethyl chain of tryptamine.



gramine



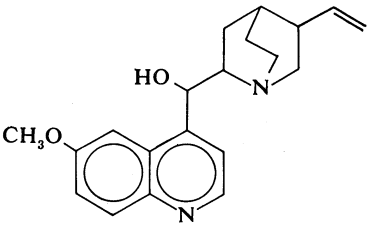
tryptophan

These alkaloids have been isolated from plants representing well over 30 families, the Apocynaceae, Rubiaceae, and Loganiaceae being particularly rich sources. These alkaloids often evince powerful physiological action, and one particularly complex group has shown promise in the control of leukemia. Vincaleucoblastine, one of some ten alkaloids present in *Catharanthus roseus* (*Vinca rosea*) being used clinically to treat leukemia, is a representative of one of the indoline-indole dimeric alkaloids.

The determination of the structure of such a complex molecule required the combined techniques of organic chemistry and spectroscopy, the study of light absorption and its relation to molecular structure. Even alkaloids containing only one indole nucleus have presented challenges to the structural chemist.

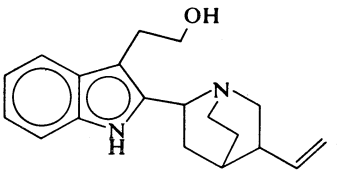
Quinolines. The quinolines, too, constitute a large group of alkaloids. Remarkably enough, they are also derived from tryptophan. The biogenesis of quinine, which has the following structures,

Quinine



quinine

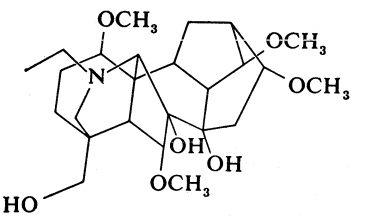
is not understood in detail but it is known that the indole nitrogen of tryptophan remains in the quinoline portion, and that the side chain nitrogen appears in the quinuclidine moiety. There is some evidence that cinchonamine, which is present along with quinine, is an intermediate in the synthesis of quinine.



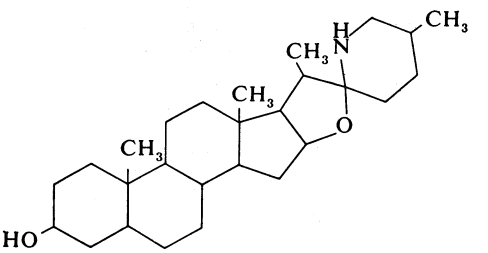
cinchonamine

The alkaloid calycanthine (see above), which for long had been regarded as an indole alkaloid and can be readily converted into indole derivatives, also has its origin in tryptophan.

Terpenoids and steroids. Most of the terpenoid and steroid alkaloids can be viewed as derivatives of the nonnitrogenous terpenoids and steroids. These are common constituents of many living systems, and are derived from mevalonic acid. In these alkaloids a nitrogen atom from ammonia presumably has been incorporated into the terpenoid or steroid molecule. The alkaloids also are distributed widely in plants (and are found occasionally in animals); the family Ranunculaceae yields many terpenoid bases, such as lycoctonine, and the Solanaceae yield largely steroid alkaloids, as represented by tomatidine.



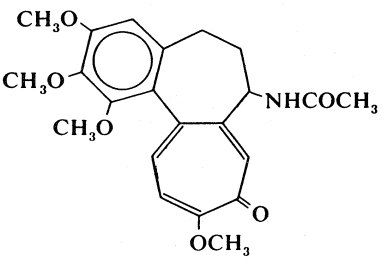
lycoctonine



tomatidine

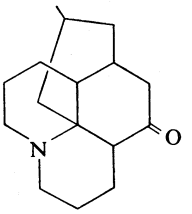
Miscellaneous. Since plants do not elaborate alkaloids for the convenience of classification, many alkaloids do not fit neatly into classes. All of those that have been subjected to scrutiny with labelled precursors, however, fall into one or other of the recognized pathways of biosynthesis.

Among the miscellaneous alkaloids, colchicine, from



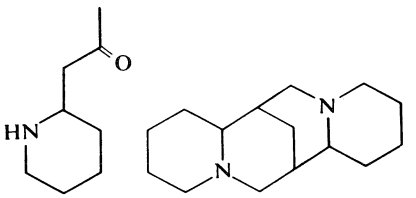
colchicine

the autumn crocus (*Colchicum autumnale*), is of importance not only because it has been used to cure gout but because it has been the cause of many deaths following accidental ingestion. Colchicine was one of the first compounds known to bring about profound chromosomal changes in plant cells. The lycopodium alkaloids, typified by lycopodine,



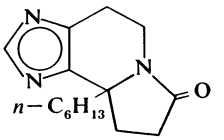
lycopodine

are represented by over 30 known examples. They appear to be derived from two isopelletierine units. The quinolizidine alkaloids, commonly found in *Lupinus* species and other leguminous genera, are typified by sparteine.



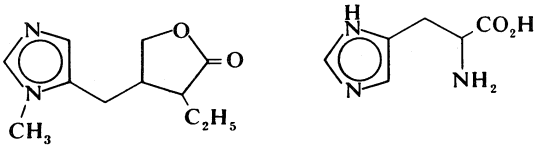
isopelletierine sparteine

Glochidicine is one of three closely related alkaloids found in *Glochidion philippicum* (Euphorbiaceae).



glochidicine

It is of interest because, like pilocarpine, which is used in medicine as an agent to induce perspiration, it is one of the few alkaloids derived from the amino acid histidine:



pilocarpine histidine

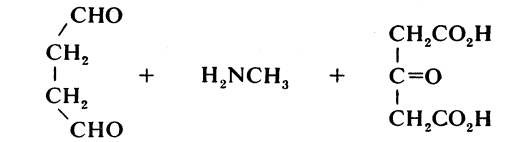
In addition to the alkaloids mentioned here there are many others that are of interest because of their molecular structure and physiological activity.

GENERAL BIOLOGY

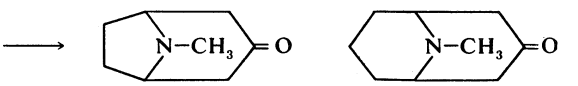
Biosynthesis. It has long been suspected that the nitrogen atoms and some of the carbon atoms of alkaloids have their origin in amino acids, nitrogen-containing organic compounds that are constituents of proteins and other biological materials. A proposal describing in detail how certain alkaloids could have been derived from amino acids was made in 1917. When the radioactive isotope of carbon, carbon-14, was made generally available, it became possible to introduce radioactively labelled amino acids into plants and verify that the label appeared in the alkaloid under investigation, confirming the incorporation of the amino acid into the alkaloid. Even earlier, indirect evidence had been obtained to explain the structures of the various alkaloid types on the basis of their origin from different amino acids. It became feasible to select the more plausible of alternate structures of a new alkaloid on the basis of the possibility of origin of the structure from known amino acids.

Emulation of biosynthesis. Attempts to emulate the presumed biological synthesis (biogenesis) of an alkaloid by in vitro experiments often met with unqualified success. These experiments were conducted with compounds known or thought to be present in the plant and were carried out under conditions that might prevail in the plant cell, namely, aqueous neutral solutions maintained at room temperature. The first such successful experiment was a spectacular synthesis of the alkaloid tropinone achieved simply by bringing together succindialdehyde, methylamine, and acetone in aqueous solution. In vitro processes

The yield was improved when acetonedicarboxylic acid was substituted for acetone, decarboxylation of the product taking place either spontaneously or by heating in acid solution. Extension of this synthetic scheme to pseudopelletierine required the replacement of succindialdehyde by its homologue glutardialdehyde.

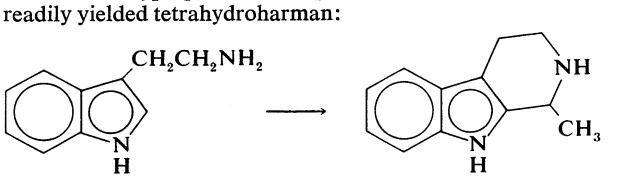


succindialdehyde methylamine acetonedicarboxylic acid



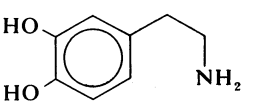
tropinone pseudopelletierine

Subsequently, many successful syntheses of alkaloids of varied structural types were achieved under equally mild conditions. Tryptamine, an indole base derived from the amino acid tryptophan, in the presence of acetaldehyde readily yielded tetrahydroharman:

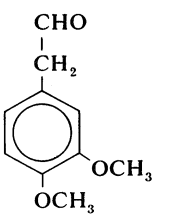


tryptamine tetrahydroharman

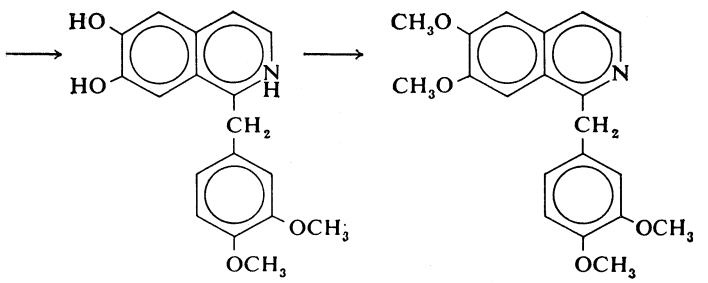
Similarly, 3,4-dihydroxy-β-phenethylamine and 3,4-dimethoxyphenylacetaldehyde yielded a tetrahydroisoquinoline, which on methylation of the oxygen atoms and dehydrogenation was converted to papaverine:



3,4-dihydroxy-β-phenethylamine



3,4-dimethoxyphenylacetaldehyde



(a tetrahydroisoquinoline) papaverine

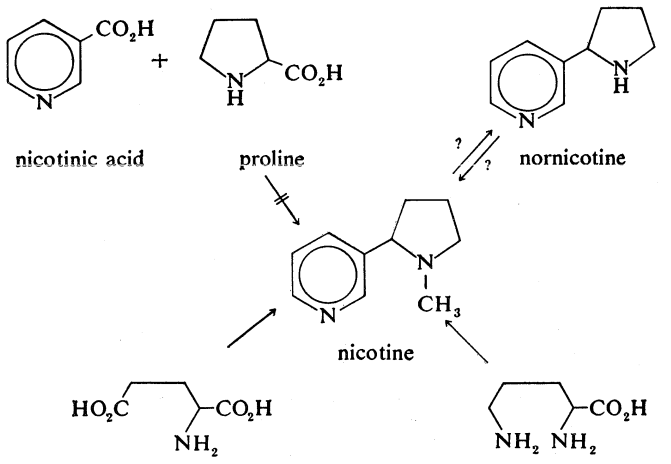
The pyruvic acid, of which the above aldehyde is a decarboxylation product, underwent the same reaction. This acid could be prepared from the amino acid tyrosine. The phenethylamine involved in the reaction is a decarboxylation product of dihydroxyphenylalanine, which also is known to be derived ultimately from tyrosine.

Radioactive labelling and biosynthesis

Use of tracers. With the aid of carbon-14-labelled substances, it became possible to test biogenetic hypotheses by direct experiments and, with a few exceptions, results were as expected. In many cases labelling experiments provided additional detail lacking in the original biogenetic hypotheses. In the meantime the processes by which the precursor amino acids are synthesized in the cell from carbon dioxide and ammonia also were being determined.

As early as 1912 it had been surmised that nicotine arose biologically from a combination of nicotinic acid with the amino acid proline. Experiments involving living tissues of *Nicotiana* species, however, have shown that two other amino acids, glutamic acid and ornithine, are incorporated into the five-membered ring of nicotine instead of proline. It proved difficult to trace the exact sequence of reactions leading to nicotine, and it is still unclear whether the nornicotine (nicotine missing its methyl group) is a precursor or a conversion product of nicotine.

The relationships among these compounds are shown in the following scheme:

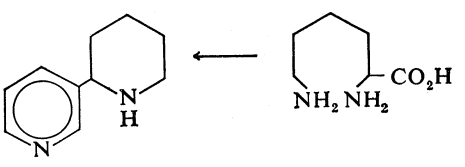


nicotinic acid proline nornicotine

nicotine

glutamic acid ornithine

The biogenesis of a structurally related alkaloid, anabasine, also is subject to some uncertainty, although the amino acid lysine is known to be incorporated into the saturated six-membered ring of the compound:

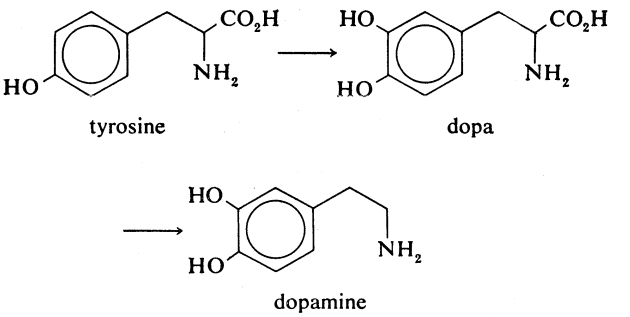


anabasine lysine

A particular advantage of labelling techniques is that they permit one to determine not only what precursors are involved in the biosynthesis of a given product but also the way in which the precursors are used—that is, the specific atoms of precursor molecules that become atoms in the molecules of the products. This information is secured by synthesizing precursor molecules bearing labelled atoms at known positions in the molecule, introducing them into the organism and, after an appropriate time period, isolating the product from the organism. If, as expected, the product bears the radioactive label, it is carefully degraded chemically, each atom being isolated in turn so that it can be determined whether or not it is radioactive. The correspondence reveals precisely what molecular transformations have taken place. Such experiments are, of course, extremely laborious; however, they yield information that can be obtained in no other manner.

Radioactive labelling experiments also have been useful in unravelling the tangled situations that arise when two molecules of a single amino acid are involved in the biosynthesis of a given alkaloid molecule. In such cases the two units may go through separate pathways of transformation, with the ultimate result that they are incorporated in different forms, and possibly to different extents, in the alkaloid molecule. On the other hand, both units may be incorporated into the alkaloid in an identical form or, after incorporation, they may form a symmetrical intermediate; in either of these latter cases the alkaloid will show an identical degree of incorporation of the two units. The relative degree of incorporation of two amino acid units into an alkaloid structure is best shown by radioactive labelling. An example of this situation and the procedure used to clarify it is given by the amino acid tyrosine, two molecules of which are believed to be incorporated into the molecules of a number of alkaloids (papaverine, morphine, narcotine, thebaine, hydrastine, berberine, and chelidonine, to name a few). Furthermore it is known that tyrosine is transformed in plant cells into a substance called dopa (for dihydroxyphenylalanine), which also is believed to be incorporated into these alkaloids to the extent of two units for every alkaloid molecule. A further transformation product of dopa (called dopamine), however, is incorporated only into one part of the molecules of these alkaloids. The progression from tyrosine to dopamine is represented as follows:

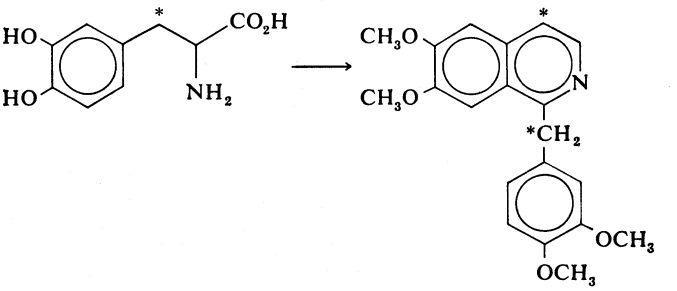
Biosynthesis with duplicate amino acids



tyrosine dopa

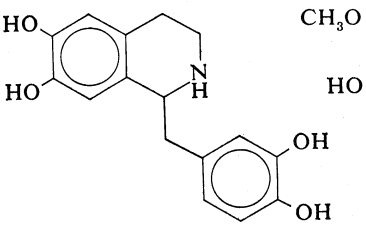
dopamine

It seems, therefore, that there must be two separate routes by which dopa is incorporated into the alkaloid molecules. This conclusion was verified by the finding that radioactively labelled dopa was incorporated into papaverine in such a way that the two labelled carbons were both radioactive but to different extents.

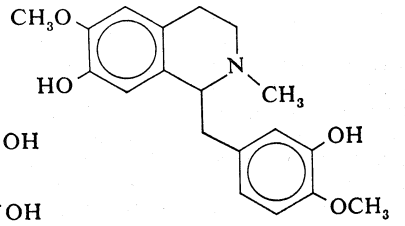


labeled dopa labeled papaverine

It seems likely that the substance norlaudanosoline,



norlaudanosoline

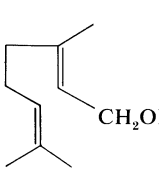


reticuline

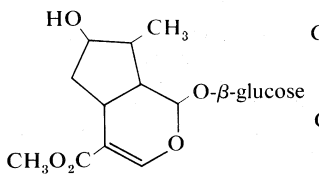
or a close relative of it, is the primary condensation product of two dopa-derived units, and the related base reticuline is actually known to occur in a number of plants that elaborate alkaloids of this type. With the aid of labelling experiments, it has been verified that reticuline is indeed the precursor of the simple benzylisoquinolines and their subsequent biological descendants, which are many of the alkaloids under discussion. Indeed, it has been shown that some of these alkaloids are intermediates in the synthesis of others.

It should be emphasized that the incorporation of amino acids into biosynthetic pathways involves more often than not a partly broken down (catabolized) amino acid in which one or more of the functional groups had been either eliminated or altered. Long before biosynthetic studies with labelled compounds became possible the degradation of amino acids into a variety of catabolites by living cells had been recognized.

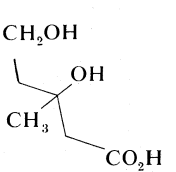
Special cases. There are, however, a large number of alkaloids whose biosynthesis cannot be explained solely by using the degradation products of amino acids. Many of the indole alkaloids, for example, such as ajmalicine and reserpinine, include in their molecules a 10-carbon fragment derived from naturally occurring hydrocarbons, the monoterpenes, like geraniol or loganin. Loganin itself occurs in the Madagascar periwinkle *Catharanthus roseus* plant (*Vinca rosea*). Mevalonic acid, the source of the five-carbon-atom (isoprene) unit in such other natural products as the terpenes and steroids, is the precursor of geraniol and loganin. Radioactive labelling experiments have shown that two units of mevalonic acid are incorporated into skytanthine, an alkaloid from the South American plant *Skytanthus acutus*. The structures of these related compounds are shown below.



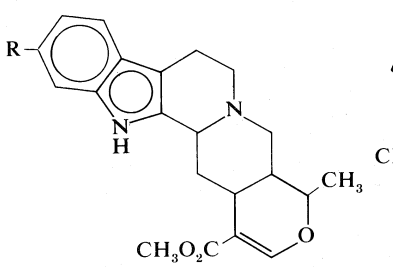
geraniol



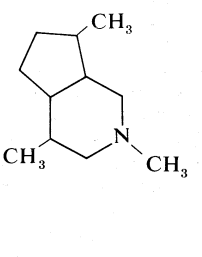
loganin



mevalonic acid



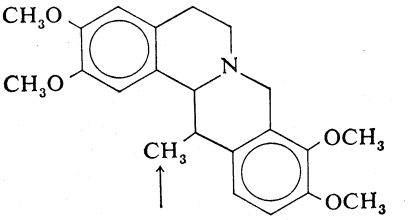
ajmalicine (R = H)
reserpinine (R = OCH_3)



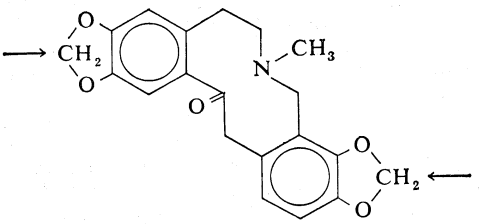
skytanthine

Origin of methyl groups

Many alkaloids contain methyl ($-CH_3$) groups joined to oxygen or nitrogen atoms and, in virtually all instances studied so far, this methyl group has originated (joined to a sulfur atom, as $S-CH_3$) in methionine [$CH_3-S-CH_2-CH_2-CH(NH_2)-CO_2H$], an amino acid that occurs in free form in plant sap and in combined form in many proteins. In addition, methyl groups attached to carbon atoms also come from methionine in certain alkaloids, corydaline, for example. The carbon atoms of the methylenedioxy ($-O-CH_2-O-$) groups of protopine

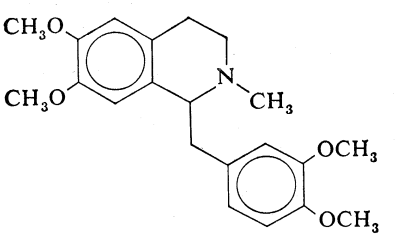


corydaline
C-methyl group, coming from methionine, indicated by arrow

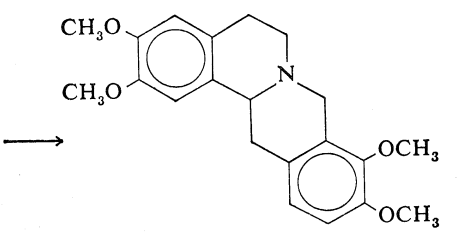


protopine
methylene groups coming from methionine indicated by arrows

also originate from methionine. It has been shown that the methylenedioxy group is formed by a process in which one of the oxygen atoms is methylated first and then, with removal of two hydrogens, ring closure with the adjacent oxygen atom takes place. The mechanism of this reaction is unknown, and it has not been possible to duplicate it in the test tube. A somewhat similar example of the removal of hydrogen atoms with consequent ring closure, also not duplicated in the laboratory, takes place in the biosynthesis of tetrahydropalmatine from *N*-methyltetrahydropapaverine, represented in the formulas:



N-methyltetrahydropapaverine



tetrahydropalmatine

BIBLIOGRAPHY. R.H.F. MANSKE and H.L. HOLMES (eds.), *The Alkaloids: Chemistry and Physiology*, 14 vol. (1950–73), a comprehensive series by authorities in the many fields of alkaloid chemistry, updated at intervals; T. KAMETANI, *The Isoquinoline Alkaloids* (1968), a study of 23 classes of isoquinoline alkaloids with physical properties, structures, and references to discovery and syntheses in tabular form; T. SWAIN (ed.), *Chemical Plant Taxonomy*, ch. 14 (1963), taxonomy from the point of view of molecular relations; R.E. ALSTON and B.L. TURNER, *Biochemical Systematics* (1963), chemical taxonomy from the botanical point of view; R. GOUTAREL, *Les alcaloïdes stéroïdiques des apocynacées* (1964),

on isolation, structure, and synthesis; I.D. SPENSER, "The Biosynthesis of Alkaloids and Other Nitrogenous Secondary Metabolites," in M. FLORKIN and E.H. STOTZ (eds.), *Comprehensive Biochemistry*, vol. 20 (1968).

(R.H.M.)

Allergy and Anaphylactic Shock

An allergy is a tendency of the body to react adversely to substances that are usually from outside itself and are typically proteins. Hay fever, some forms of hives, anaphylactic shock, and some types of asthma are among the more common forms of allergy in man.

The proteins to which an allergic reaction is possible are present in the environment in countless forms. They may be in pollen or dust that enters the nose or eyes; they may be in the serum of a vaccine, or in an antibiotic, such as penicillin; or they may be in some article of food.

Sensitization, antigens, and antibodies

An allergic reaction is in effect a matter of two contacts with the foreign protein. Reaction at the first contact is not apparent, though a change takes place within the body. It is this change, called sensitization, that causes a strong adverse reaction to later contacts with the protein. The protein causing the reaction is called an antigen, and substances formed within the body during the processes of sensitization are known as antibodies.

The antibodies may be carried in the bloodstream as immunoglobulins (*i.e.*, modified forms of a blood protein known as globulin). Allergies may also be mediated by sensitized cells—lymphocytes or monocytes—in the bloodstream and in the spleen and lymph nodes (see BLOOD, HUMAN; LYMPHATIC SYSTEM, HUMAN). Allergic reactions involving sensitized cells are ordinarily several hours longer in developing than those related to antibodies in the bloodstream, which usually occur promptly after contact has been made with the antigen. The two types of allergy are sometimes referred to as delayed and immediate allergies, respectively.

In an allergic reaction another element besides antigen and antibody often plays a role. Cells affected by the antigen-antibody interaction give off certain chemicals—histamine is the most familiar example—that are directly responsible for the visible effects of the allergic attack: hives, runny nose, itching eyes, and so forth. Drugs designed to counter the effects of histamine—antihistaminics—are sometimes used to alleviate these conditions in persons suffering attacks of allergy.

The acquisition of immunity to infections and toxins is a favourable bodily response to a foreign protein analogous to the unfavourable allergic response. Both types of reaction are referred to as immunologic (see IMMUNITY).

"IMMEDIATE" ALLERGIES

Anaphylactic shock. Anaphylactic shock, a violent reaction to the injection of a foreign protein, is marked by sudden collapse, a dramatic fall in blood pressure, difficulty in breathing, and a bluish tinge in the skin. The foreign protein may be any of those listed previously as causing an allergic reaction.

The perceptible reactions are directly due to histamine and other substances released by cells affected by the antibody-antigen reaction. The difficulty in breathing results from spasms in the muscular walls of the bronchi, with consequent narrowing of the air passages, while the fall in blood pressure is from dilation of the small blood vessels. The histamine and similar substances may also cause extreme itching, particularly about the face.

Detection of susceptibility to anaphylactic shock is sometimes difficult. Skin testing to determine whether there is a reaction to the application of small samples of the suspected substance to the surface of the skin or to scratches in the skin are useful in uncovering sensitivity to such substances as horse serum, pollen, and food, but may be unreliable in testing for susceptibility to drugs. Moreover, the tests themselves can cause anaphylactic shock. A person is more likely to be susceptible to anaphylaxis if the substance has caused the reaction previously, or if he has had hay fever, asthma, or hives.

Treatment of anaphylactic shock

The treatment of anaphylactic shock is directed principally toward restoration of normal blood pressure and breathing. Epinephrine has been found most reliable for this purpose. Antihistaminic drugs may relieve itching but do not control the shock, and corticosteroids—adrenal cortical hormones—are useful only in controlling the accompanying minor symptoms. Emergency measures—mouth-to-mouth breathing and heavy massage of the chest over the heart—may be necessary to resuscitate the affected person. Recovery from anaphylactic shock is sometimess spontaneous. If a foreign substance must be injected into a person known to be highly sensitized to the substance, desensitization by repeated small but increasingly larger amounts of the material may prevent the reaction. This process of desensitization must be repeated before each new series of injections of the substance.

Hay fever and allergic asthma. Hay fever is characterized by recurrent, usually seasonal rhinitis ("runny-stuffy nose") with swelling of the mucous membranes of the nose, itching of the eyes and nose, sneezing, and cough. Inhaled particles of pollen, animal dander, or dust are usually responsible, and the severity of the effects on a susceptible person is proportional to the quantity of the antigen in the air.

If an extract of the antigen is injected into the sensitized person it may cause a reaction in the skin or in the body as a whole; it may even cause anaphylactic shock. The skin reaction is a reddened wheal, which is like a scratched mosquito bite; it appears a few minutes after the injection and lasts for about an hour. This reaction can be induced in the skin of a person not ordinarily subject to hay fever by a two-stage process: first, serum (the blood minus its cells and clotting factors) from an affected person is injected; a short time later the antigen is injected into the same skin area.

About one person in ten is a potential sufferer from hay fever. The tendency is inherited. The direct cause of the hay fever symptoms appears to be histamine, given off by white blood cells in the presence of the antigen and antibody. Antihistamines, adrenal hormone, and desensitization with the antigen are used to deal with attacks.

Allergic asthma usually begins in childhood. It may disappear in adolescence; less frequently, it develops into chronic bronchial disease in adult life. A child with allergic asthma has difficulty in breathing, especially in breathing out, and coughs and sneezes. He usually has hay fever, also, but his asthma, unlike his hay fever, is not helped by antihistamines; corticosteroids and desensitization with the antigen responsible are of benefit. Drugs that relieve spasms of the bronchi stop attacks.

Persons who have hay fever or allergic asthma are subject to infections of the respiratory tract. Hay fever victims may have inflamed sinuses and nasal polyps (projecting lumps of swollen mucous membrane). Asthmatic persons tend to have attacks of bronchitis and pneumonia; chronic asthma and emphysema are likely to be aftermaths of these infections rather than direct consequences of the asthma itself.

Atopic dermatitis and hives

Other types of immediate allergies. *Skin allergies.* Allergic diseases of the skin may be of the immediate or the delayed variety. Allergy of the latter type, including contact dermatitis, is discussed in a later section. Atopic dermatitis and hives, or urticaria, may be manifestations of the immediate type of allergic reaction. Both may be attributable to foods, drugs, or chemicals.

The skin inflammation called atopic dermatitis tends to occur in the hollows of the elbows and knees of adults and older children. On infants it appears on the face, in the diaper area, and on the arms and legs.

Hives, or urticaria, characterized by patches of wheals, is often attributed to eating of berries or fish, or to being treated with a drug such as penicillin. Atopic dermatitis and hives tend to disappear when the affected person stops coming into contact with the antigen, although the irritated skin of infants often becomes infected, with consequent difficulties in restoring a normal condition.

Gastrointestinal allergy. Recurrent attacks of abdominal pain, vomiting, and diarrhea have occasionally been associated with an allergic reaction. The attacks, which will disappear without treatment, may be accompanied

by hives or rashes. Affected persons may have an inherited tendency to hay fever, asthma, or allergic dermatitis.

DELAYED ALLERGY

Delayed allergy has been implicated in contact dermatitis, in the bodily reactions to some infections, in autoimmune diseases, and in transplantation and tumour rejection. (Autoimmune diseases involve an individual's allergic reaction to his own cells or antibodies.)

Contact dermatitis. The manifestation of contact dermatitis may be acute or lasting. In the acute phase, papules (small, raised patches) and vesicles (small blisters) predominate, along with redness and swelling of the affected skin. Crusting of the skin may develop. In the lasting or chronic phase the skin becomes thick, scarred, and scaly; there may be no vesicles or papules. The course of the illness waxes and wanes with repeated exposures to the material responsible for the reaction. Generally, the substances responsible for contact dermatitis are simple chemicals such as dyes, cosmetics, plant resins, topical medications, plastics, and metals. The skin site affected often gives a clue as to the substance that is probably responsible; the scalp: lotions, shampoos, hair dyes, wave sets; the ears: earrings, perfumes; the trunk: clothing, bathing materials, perfumes; the legs, ankles, and feet: shoes, leather dyes, clothing, and metallic objects.

There is usually an interval of 12 to 72 hours between exposure to the substance and development of a reaction. The substance responsible for the reaction may be identified by means of patch tests—application of a succession of substances to the skin for several hours under adhesive tape or a bandage, until a substance is found that causes a skin reaction.

Allergic reaction to infection. It is probable that delayed allergy is a general occurrence with infection by viruses, bacteria, and fungi. Largely in chronic infections, however, does delayed allergy seem to have a role in the causation of disease reactions or in the acquisition of resistance to infection. The involvement of delayed allergy in tuberculosis has been studied more intensively than in any other infection.

The presence of delayed hypersensitivity can be detected by injection of a killed micro-organism or a protein extract from the micro-organism into the skin of the person affected. This produces a local reaction, such as that produced by the patch test, in 24 to 48 hours.

Manifestations of delayed allergy during chronic infection include both local and systemic reactions. Severe inflammation and tissue damage with death of tissue may develop at the site of bacterial growth when delayed allergy appears during the course of infection. This reaction probably accounts largely for the tissue destruction and inflammation characteristic of tuberculosis in the lung and in other organs. In tuberculosis of the lungs, when infected material spills over into previously uninvolved tissue, an acute inflammatory reaction producing pneumonia may develop, attributable entirely to delayed allergy to protein of the tubercle bacillus.

Systemic reactions

Systemic allergic reactions are particularly likely to occur in chronic infections when the infection is disseminated through the blood. Then fever, headache, prostration, and even shock may ensue. Delayed allergy is involved in the fever of tuberculosis.

The rose spots of the skin over the abdomen and chest in typhoid, the fever of brucellosis, and manifestations of other infections have been related to delayed allergic reactions to microbial antigens but this interpretation has not been as fully documented for other diseases as for tuberculosis.

Macrophages, large cells that attack and consume foreign bodies and debris in a process called phagocytosis or phagocytic activity, play a secondary role in delayed allergy but are important in resistance to infection. With development of delayed allergy during infection the macrophages begin multiplying rapidly and demonstrate a burst of phagocytic activity. It is probable that this effect of delayed allergy contributes in part to acquired resistance to infection. It has been demonstrated also that macrophages from persons with delayed allergy to specific micro-organisms may be better equipped to kill bacteria after phagocytosis. Certain viral infections, particularly vaccinia or cowpox, also appear to be controlled in part by delayed allergic reactions to the virus.

Autoimmune diseases. Allergic encephalomyelitis (see below), chronic inflammation of the thyroid, and possibly other diseases associated with autoimmune reactions may be related to delayed allergy. Allergic encephalomyelitis was first observed after an administration of Pasteur vaccine made of rabbit spinal cord and inactive virus. A similar disease subsequently developed after vaccination for smallpox and some viral infections. Manifested by a staggering gait, weakness or paralysis, and occasionally by lethargy, confusion, stupor or coma, the disease can also be produced experimentally in several animal species by the administration of brain tissue.

Chronic thyroiditis—inflammation of the thyroid—occurs spontaneously in some persons. The thyroid is enlarged and may be tender and hard. Initially there is some evidence of increased thyroid activity; later, in some instances, there is decreased thyroid function. Lymphoid cells collect in the gland. Delayed autoallergy (self-sensitization) is probably largely responsible for this disease.

ALLERGENIC REACTIONS OF BLOOD

The important allergic diseases affecting blood cells are erythroblastosis fetalis (hemolytic—red blood cell destroying—disease of the newborn infant), autoallergic acute acquired hemolytic anemia, autoallergic thrombocytopenic purpura, cold hemoglobinuria, allergic blood transfusion reactions, and certain allergic reactions to drugs.

Erythroblastosis fetalis, a disease of newborn infants characterized by jaundice, pallor, bruising, occasionally severe edema, and heart failure, is attributable to destruction of the infant's red blood cells by an antibody transferred from the mother to the fetus across the placenta. The antibody is produced by the mother. This antibody reacts to an isoantigen (Rhesus factor) possessed by the red blood cells of the fetus and not present in the red blood cells of the mother. (Isoantigens are antigens possessed by some and not all members of a species. Blood groups, for example, are established on the basis of the isoantigen shared by the persons of each blood type.) During some pregnancies fetal red blood cells appear to enter the maternal blood through the placenta; and if these cells behave as foreign antigens, an immune reaction develops in the mother, which may result in transfer to the fetus of antibody-causing hemolytic (red-cell destroying) disease.

Rh factor

The Rhesus (Rh) isoantigens are now known to be complex but for all practical purposes about 80 percent of human beings possess the Rh isoantigen (Rh+), and 20 percent do not (Rh−). When the father is Rh+ and the mother Rh−, their offspring may be Rh+ or Rh−; if the fetus is Rh+, the infant may develop erythroblastosis fetalis. In Rh+ babies with Rh− mothers, hemolytic disease occurs only in about 1 in 100 births. The likelihood that babies will develop the disease under these conditions, however, increases with each additional pregnancy.

Recent evidence confirms that administration of antibody to Rh isoantigens to the pregnant Rh− woman can prevent development of erythroblastosis in the infant; this measure probably removes or renders less antigenic the Rh+ red cells of the fetus that cross the placenta in the mother.

Some blood transfusion reactions are accounted for by immunologic reactions between isoantigens of transfused blood cells (white cells and red cells) and the recipient's isoantibodies or antibodies acquired because of previous blood transfusions. These reactions may be characterized by fever, rash, hemolysis, or shock.

Autoallergic hemolytic anemia may affect persons of all ages and both sexes. The disease varies in severity and duration and may be relapsing. Anemia, the principal feature of the disease, may be responsible for weakness, tiredness, and lethargy. Persons with this disease are found to have antibodies that react directly with their own red blood cells.

There is no failure of red blood cell production in persons with allergic hemolytic disease. This is demonstrated by increased bone marrow activity and the presence of young, new red blood cells (reticulocytes and erythroblasts) in the blood of affected persons. Red blood cells acted upon by the antibody may be destroyed or removed by the spleen and liver; both processes result in anemia. The antibodies leading to the destruction of red cells (lysis) and those leading to the removal of red blood cells by the spleen and liver are probably different types of immunoglobulin. In addition, the antibodies acting on red blood cells may react maximally at 37° C (98.6° F) or below 37° C ("warm" or "cold" antibodies). When "cold" antibodies are responsible for hemolytic anemia, the patient's clinical illness often is precipitated by exposure to cold and is characterized by jaundice, pain, and pallor of fingers and toes (Raynaud's phenomenon), and the presence of hemoglobin in the urine.

Some forms of thrombocytopenic purpura, a condition characterized by platelet deficiency with bleeding, have been attributed to autoallergy, but the actual demonstration of platelet autoantibodies has been very difficult. The most important evidence of the immunological basis for the disease has been low level of platelets in the blood after transfusion into normal persons of plasma or blood from a person with the disease.

DRUG ALLERGIES

Adverse reactions to drugs have become a major cause of illness, the full extent of which has not yet been completely delineated. Only about 15 percent of the diseases caused by drugs appear to be allergic. Usually the identification of the disease as allergic is made by recognition of an allergic type of reaction rather than by laboratory tests.

These allergic manifestations include such signs as hives and rash and also include the following characteristics: (1) a sensitization period is required before administration of the drug produces a reaction; (2) once the sensitization has developed, readministration of the responsible drug may result in an "accelerated reaction"; that is, the reaction appears abruptly rather than slowly; (3) the development of a reaction is not specifically related to the quantity of the dose.

The method of administration of medications may influence the likelihood and type of allergic reactions that may occur. Penicillin given orally, for example, is less likely to induce sensitization than when given subcutaneously. Penicillin applied to the bodily surface is associated with such a high rate of sensitization that this method of administration is no longer employed. Application of drugs to skin may also cause allergic contact dermatitis.

Manifestations of drug allergy

The common manifestations of allergy to drugs do not differ from allergic reactions of other types. Rashes of almost every variety are common. Itching, pruritus, is frequent. Certain types of skin manifestations are particularly associated with certain drugs.

Fever may be an isolated manifestation of drug allergy. Penicillin and other antibacterial drugs are particularly common causes of drug fever. Occasionally drug fever is also associated with rash, pain in joints, and swelling of lymph nodes ("serum sickness"). Discontinuance of the drug terminates the fever unless the fever is accompanied by other manifestations, such as rash, when it may persist for several days.

Nephritis, acute inflammation of the kidneys, has been particularly related to allergy to sulfonamide drugs. A diffuse disease of the blood vessels, polyarteritis nodosa, probably due to an allergic reaction, has been attributed to several drugs, including sulfonamides, chlorpromazine, and penicillin. Liver disease with jaundice has been associated with reactions to many drugs. Anaphylactic shock is usually caused by allergic reactions to penicillin in the United States, although other drugs also are blamed.

The only effective way to control an allergic reaction to a drug is to discontinue use of the drug. Antihistamine and corticosteroid drugs are sometimes used to suppress symptoms if the reactions persist.

ORGAN AND TISSUE REJECTION REACTION

The transplanting of organs from healthy into sick persons has received impetus in recent years, but the possibility has intrigued man for centuries. Certain forms of transplantation (of the cornea of the eye to remedy some forms of blindness, for example), have become commonplace and are successful. Other forms of organ transplantation, however, pose difficulties, which have been shown to be attributable to allergic reactions.

Types of grafts

An important determinant in the survival of transplanted tissue is the relationship between the donor and the recipient. For this reason, the following terms have been used to describe tissue grafts: (1) autografts are transplants from one region of the same individual to another; (2) isografts are transplants from one individual to a genetically identical individual (identical twins); (3) homografts or allografts are transplants from one individual to a genetically nonidentical individual of the same species; (4) heterografts are transplants from one species to another. The donors of tissues or organs, therefore, are referred to as autologous, isologous, homologous, or heterologous. Among vertebrates, autografts and isografts usually produce no reactions and transplants are accepted and endure. In contrast, allografts and heterografts are regularly rejected by an allergic reaction.

When an allograft is placed on an animal the graft at first develops blood vessels and appears to be healthy, but after 10 to 14 days or longer a severe inflammatory reaction develops and the graft withers and sloughs. If another graft from the same donor is then made into the recipient animal it quickly is involved in a severe inflammatory reaction, and in less than seven days is rejected. This accelerated reaction to the second allograft is attributable to allergic sensitization. The reaction is specific for the tissues or organs transplanted from a specific donor, and the tissue antigens responsible for the allergy are called transplantation or histocompatibility antigens. These antigens are determined genetically by histocompatibility genes. The number of different combinations of the genetically determined histocompatibility antigens is astronomical, a fact that accounts for the uniqueness of individuals with respect to their transplantation antigens and for the allergic rejection reaction regularly seen with allografts.

Virtually all cells of the body except red blood cells possess transplantation or histocompatibility antigens involved in graft rejection. The reaction to an allograft is generally considered to be an expression of delayed allergy. Susceptibility to the reaction is transferable from one individual to another by means of lymphoid cells. Serum from individuals demonstrating allograft rejection usually is ineffective in transfer of the allergic reaction.

In certain circumstances allografts are not rejected. Some sites of the body (the brain coverings and the anterior chamber of the eye, for example) are "privileged" sites and accept allografts without an allergic reaction. This probably is related to the lack of lymphatic drainage at these sites, as draining lymphatics may be essential to development of allergic sensitization. The fetus in the uterus is actually an allograft, possessing histocompatibility antigens foreign to the mother, but allergic rejection generally does not occur. This probably is related to the separation between maternal and fetal circulatory systems in the placenta.

Tolerance of certain allografts by chimeras

It was found in 1945 that nonidentical twin calves with interconnecting (anastomosing) placental blood vessels in the uterus had persistent mixtures of each other's red blood cells in their circulation after birth. This was taken to indicate that in each calf's bone marrow there were cells of the other calf producing red cells that, although foreign, were accepted, persisted, and did not act as antigens to produce allergic reactions. It was then demonstrated that such nonidentical twins accepted allografts from each other without allergic reactions or rejection. The calves, therefore, had become permanently tolerant of each other's tissues. They were mixtures of genetically different cells; and such individuals are called "chimeras," after the monster in Greek mythology with a lion's head,

a goat's body, and a serpent's tail. It has since been shown that individuals can be made tolerant to allografts from a genetically alien individual if embryonic or newborn animals are inoculated with living cells from the donor. When mature, such animals accept allografts permanently from the donor, although they reject in usual fashion allografts from other individuals.

Adult animals have been made tolerant to allografts by inoculation of cells from particular donors when the recipient's ability to respond immunologically has been temporarily rendered incompetent by intense treatment with X-rays or with cytotoxic drugs (drugs that poison cells), followed by multiple injections of the donor cells. If lymphoid or spleen cells are the allograft given to an animal with immunological tolerance to a specific donor's tissue, the grafted cells, which are able to react immunologically, may react allergically with the recipient's own tissues, producing weakness and death.

Antigenic differences between tumour and normal cells

Tumour cells commonly differ antigenically from normal cells in either (1) having fewer or no histocompatibility antigens or (2) having new antigens. The animal developing a tumour with new histocompatibility antigens may react to it as if it were an allograft. Tumours induced in animals with chemical substances or viruses have regularly been found to have new histocompatibility antigens not represented in normal cells of the host. Such antigens in tumours have been found in spontaneously developing tumours of man, particularly illustrated by the Burkitt lymphoma. (This malignant tumour, possibly caused by a virus, is found in a well-defined area in equatorial Africa and also has been identified in the United States.)

Allogenic transplantation of organs from one person to another (allografts) continues largely to be experimental. Recent development of methods for typing human histocompatibility antigens of cells, however, has made it possible to secure greater success than was previously possible except between genetically identical or closely related individuals. Drugs to suppress the allergic reactions and use of means to abolish delayed allergy with antilymphocyte serum continue to be useful but produce severe complications, which themselves may cause illness and death.

RECENT THERAPEUTIC ADVANCES

Allergic disease attributable to antigens in the environment are best controlled by avoiding or eliminating the offending substance. This may be impossible if the antigen cannot be avoided or identified precisely or if the allergic reaction involves autosensitization or results from a procedure such as transplantation. Under these conditions two methods for control of the allergic reaction are possible: (1) interference with or alteration of the immunological process; and (2) counteracting or interfering with the action of chemical mediators of the allergic reaction (*i.e.*, histamine and the other similar substances given off by affected cells).

Interference with the immunological process. Once sensitization has occurred, as in pollen hay fever or drug allergy, desensitization or hyposensitization can occasionally be accomplished by the administration of gradually increasing doses of antigen. Weekly or twice-weekly injections of increasing amounts of the antigen are given before natural or full exposure to the antigen is expected.

Desensitization treatment carries some risk, and unpleasant local or systemic allergic reactions may be induced. Furthermore, the amount of antigen needed to give adequate therapeutic results is not predictable in the individual patient. The success of the treatment is not assured and its use is controversial except in carefully selected instances, such as in childhood hay fever and insect allergic reactions.

Immunosuppressive therapy

Suppression of the immunological responses, or immunosuppressive therapy, has become important in control of transplantation allergic reactions and in treatment of certain chronic diseases associated with autoallergy. Cytotoxic drugs have been shown experimentally to interfere with the immunological mechanisms of allergic reactions and have been used in these clinical situations.

A problem in the use of cytotoxic drugs is that their effect is to suppress not only immunological reactions to allergic disease but protection against infection as well. Consequently severe and uncontrollable infections may develop in persons in whom the drug's effect upon phagocytes and upon immune cells is excessive.

Antiserum to lymphocytes will reduce or eliminate lymphoid cells and their function and will suppress delayed allergic reactions. Persons treated in this way will accept allografts without allergic rejection. Such treatment, however, has also been associated with the appearance of tumours. This mode of treatment, therefore, as with cytotoxic drugs, suffers from not affecting the allergic disease specifically and from interfering with other immune responses.

The need is for safe methods that will regularly induce immunological tolerance, even in sensitized individuals. Its effects must be specific and durable. Such methods are not now available.

Inhibition of chemical mediators. As has been said above, certain chemicals released by the cells are involved in producing some allergic manifestations. These chemical mediators of tissue reactions include histamine, "slow reactive substances," 5-hydroxytryptamine (serotonin), bradykinin, acetylcholine, and a few others. Drugs that interfere with the action or effects of these chemical substances can suppress some of the manifestations of allergic reactions.

Antihistamine drugs

Antihistamine drugs are effective in alleviating several allergic reactions, but their action is not specific, and the antihistamine activity of these drugs does not always parallel their clinical effectiveness. There are no reasonably potent specific inhibitors of slow reacting substance, but drugs relaxing bronchial smooth muscle can counteract its effect. Calcium has some effect in limiting the action of the substance. Development of new drugs highly effective in inhibiting slow reacting substance will probably have significant use in suppressing some of the manifestations of allergic disease.

Drugs with strong inhibitory action on 5-hydroxytryptamine are available, but appear to have little effect on allergic disease, particularly asthma. No drugs are available to inhibit the action of bradykinin. Atropine is an effective antagonist of acetylcholine and relieves the excess bronchial secretions and reduces bronchial tone, but acetylcholine has never been implicated directly in any allergic reaction in tissue.

A wider knowledge of the biochemistry of acute allergic reactions has not yet provided any significant advance in therapy. Development of specific antagonists of chemical mediators of the allergic reactions is required for rational therapy of these diseases.

Adrenal corticosteroids are useful in alleviating many allergic reactions, but their effect is highly unspecific and probably related to their inhibition of inflammation and their maintenance of blood vessel integrity.

BIBLIOGRAPHY. P.A. MIESCHER and H.J. MULLER-EBERHARD (eds.), *Textbook of Immunopathology* (1968), a comprehensive text covering the fundamental mechanisms involved in the development of immunological diseases; M. SAMTER *et al.* (eds.), *Immunological Diseases* (1965), a complete presentation of the clinical and pathophysiological features of immunological processes; L.E. CLUFF and J.E. JOHNSON, "Drug Fever," *Prog. Allerg.*, 8:149–194 (1964), the only definitive review of the clinical features, circumstances, and mechanisms involved in drug fever, including the physiological and immunological processes concerned; K.F. AUSTEN, "Systemic Anaphylaxis in Man," *JAMA*, 192:108–110 (1965), an accurate and critical study of anaphylaxis in man, attributable to drugs and other antigenic substances, including known mechanisms; L.E. CLUFF *et al.*, "Epidemiological Study of Adverse Drug Reactions," *Trans. Ass. Amer. Physicians*, 78:255–268 (1965), a review of the overall problem of adverse reactions to drugs including those due to drug allergy; F.M. BURNET, *Cellular Immunology* (1969), a comprehensive, delightfully written, speculative text on modern immunology.

(L.E.C.)

Alluvial Fans

Alluvial fans are sedimentary deposits at the mouths of canyons or valleys, where emerging streams are free of

lateral constraints and can spread their transported loads. The deposits are fan-shaped in plan view and approximate a half-buried cone in their three-dimensional aspect. The term alluvial cone is sometimes used synonymously, derived perhaps from the fact that the French geologist Alexandre Surell (1813–87) described an alpine fan as a *cône dé déjection* in 1841.

Fans can develop under a wide range of climatic conditions and have been studied in the Canadian Arctic, Swedish Lappland, Japan, the Alps and Himalayas, and other temperate to humid areas. They tend to be larger and more prominent in arid and semi-arid regions, however, and generally are regarded as characteristic desert landforms. This is particularly true in the basin-and-range type areas of parts of Iran, Afghanistan, Pakistan, the western United States, Chile and Peru, Sinai and western Arabia, and Central Asia, where the basic landscape configuration consists of mountains set against adjacent basins. Fans commonly occupy 30 percent of the total surface area in such regions and may occupy 70 percent or more of the area of a particular desert basin.

Alluvial fans border the mountain fronts with the apex of each fan just within a canyon mouth that serves as the outlet for a mountain drainage system. Sediment derived by erosion within the mountains is transported by these drainage systems to the adjacent basin; in arid regions this principally occurs during relatively infrequent flash floods that may involve mudflows. The fans, which are the main sites of deposition, are therefore an intrinsic part of an erosional–depositional system in which mountains tend slowly to wear away and basins to fill with sediment through geological time.

Analogies with ancient fans and submarine fans

Several important analogies exist with regard to both space and time. River deltas (*q.v.*) are an obvious example of an erosional–depositional system, and their fan-shaped deposits differ in origin from alluvial fan deposits largely because the deltas are formed under water, where a river enters the sea or another body of water. Deep-water submarine fans are being formed today on the ocean floor; sediment for these marine analogs is derived from the submarine canyons that cut the continental shelves and slopes of every continent (see CANYONS, SUBMARINE). The transportation of sediment from these canyons is accomplished by turbidity currents, a variety of density currents (*q.v.*); this mechanism can be considered analogous to the terrestrial mudflows that occur on alluvial fans, although the submarine flows are much less viscous.

Mudstones and cobble conglomerates of Silurian age (395,000,000 to 430,000,000 years ago) that occur in western Connacht, Ireland, are thought to have been deposited in a submarine fan, on the floor of an ancient sea. Many conglomerates and arkosic (feldspar-rich) sandstones in the geological record, however, have been interpreted as ancient alluvial fan deposits that formed in surficial mountain–basin systems not unlike those that exist today. The term fanglomerate is sometimes applied to individual conglomerate units to indicate their terrestrial fan ancestry. The Torridonian Beds of northwestern Scotland and the Newark Series of the eastern United States are examples of Late Precambrian (about 600,000,000 to 800,000,000 years ago) and Triassic (190,000,000 to 225,000,000 years ago) fan formation, respectively.

The study of modern alluvial fans provides some clues that permit discrimination of ancient fan deposits and, hence, earth history. For this reason, and because certain aspects of modern fan formation are not yet entirely clear, alluvial fans are the subject of scientific interest.

Importance to man

Fans also are of practical and economic importance to man, particularly in arid and semi-arid areas where they may be the principal groundwater source for irrigation farming and the sustenance of life. In some instances, entire cities have been built on alluvial fans (*e.g.*, Los Angeles).

Many fans in Iran, southern and Central Asia, and North Africa are dotted with debris mounds that were produced during the hand excavation of networks of tunnels (horizontal water wells) and shafts. These works are variously termed *qanāts, ghanats, karez*, or *foggaras*, and their purpose was to tap water within the upper reaches of fans near the mountains and obtain flow for irrigation needs at the bases, or toes, of the fans in the valleys below. The practice is an ancient one, and the city of Persepolis may have obtained its water supply in this way around 500 BC. In Iran, the use of *qanāt*s predates the Islāmic era. The ancient Chinese bored primitive wells in fans of the Gobi Desert that are reputed to have attained depths of about 1,500 metres. In addition to their potential use as water sources, alluvial fans and their associated floods are of concern to those involved in the planning, construction, and maintenance of highways, bridges, pipelines, canals, and other modern man-made structures in fan terrain.

This article treats the physical properties of fans, aspects of sediment transport, and the factors involved in fan formation. For further information of relevance on these points see FLUVIAL PROCESSES; SEDIMENT YIELD OF DRAINAGE SYSTEMS; DESERTS; PEDIMENTS; and LANDFORM EVOLUTION. For details on the nature of the rocks formed in ancient fan environments see CONGLOMERATES AND BRECCIAS; and SANDSTONES.

PHYSICAL ASPECTS OF ALLUVIAL FANS

The basic physical aspects of alluvial fans are (1) fan morphology, namely the size, shape, and gradients of fans; (2) such surficial features as active and abandoned washes, desert pavements (see below *Surficial features*), and mudflow levees and deposits; and (3) the nature of fan sediments and their stratigraphic arrangement. The available data permit some broad generalizations to be made about each of these physical aspects, but they are not sufficient to permit precise mean values or ranges of values to be set forth. This results from the fact that detailed and reliable studies in both humid and arid regions have centred on a few selected fans within a given basin or area or on a few basins or areas within a given region. Moreover, published data occasionally are misleading because an investigator may have studied a pediment (bedrock surface adjacent to a mountain mass) and confused its thin veneer of alluvium with a true alluvial fan. These qualifications are clearly of relevance with respect to discussion of the physical aspects of fans.

Size, shape, and gradients

Fan morphology and empirical relationships. The size of an alluvial fan is related to many factors, among which sediment supply, climatic variables, and the physiography of the source area and the depositional basin are important. There is, essentially, no true lower limit for fan size; microfans can form beneath a runoff rill in a fresh road cut, from a lump of sod in a newly plowed field, or in similar situations. The upper size limit is similarly vague, largely because fans whose lengths exceed 20 kilometres, approximately, tend to flatten, spread, and merge with adjacent sediments to so great an extent that their boundaries cannot be discerned and they are described as alluvial plains. Size frequency studies in the western United States and in parts of Pakistan suggest that six to ten kilometres is the most common length class. In more humid regions, particularly those that have been deglaciated during the Holocene Epoch (the last 10,000 years), fans commonly are located at the junctions of tributary and mainstream valleys. Sediment derived from glacial deposits at the heads of tributary valleys is continually added to the fans, but their size is restricted because the stream in the main valley continually removes sediment at their lower ends. This situation exists, for example, in Ladtjovagge Valley, Sweden, and along some reaches of the Copper River, Alaska (Figure 1). In general, the high frequency of rainfall and stream flooding in temperate or humid regions tends to inhibit fan growth by erosion; small fans (no longer than two to three kilometres) at valley junctions are the rule rather than the exception. In addition, sediment tends to be transported beyond fan boundaries by through-flowing streams in humid areas; this factor also mitigates against large fans.

The shapes of fans approximate a segment of a cone

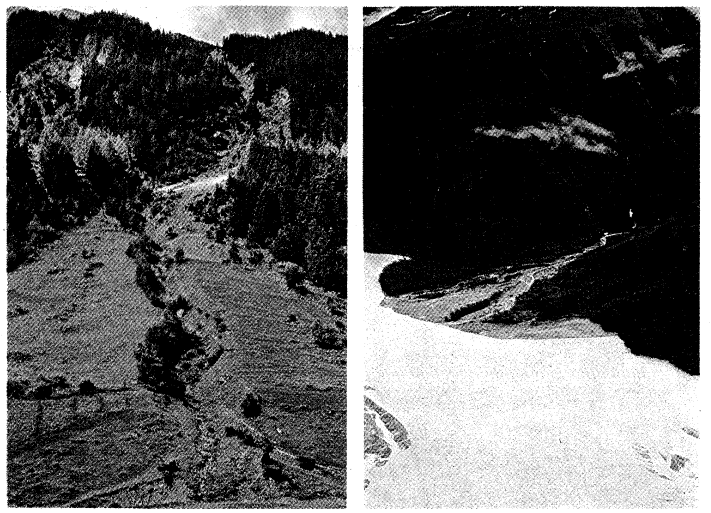


Figure 1: (Left) Grass-covered alluvial fan in the Dolomite Alps, Austria. (Right) Alluvial fan at the junction of tributary and main valleys, Copper River, Alaska.
By courtesy of L.K. Lustig

where not recently deformed by tectonism (movements of the earth's surface) or restricted by the growth of adjacent fans. Cross-fan profiles are convex, and radial profiles (paths from the fan apex to its lower margin) tend to be slightly concave upward. Locally, however, the midfan termination of some former mudflow or debris flow along a radial path may introduce a slight convexity not yet removed by erosion. The gradients of radial profiles generally range from 2° to 5° but, again, average values are based on limited samples. Moreover, the gradients are a function of fan size, dominant rock type of the source area, and depositional and tectonic history. Fans may be quite steeply sloping in recently uplifted areas.

Importance of particle size and drainage area

A fairly good correlation exists between fan gradients and particle size. Gravels as coarse as boulders are associated with steep slopes, whereas fans derived from shale and mudstone terrains tend to be flatter and more subdued, and much clay-size material is present. This same relation exists in river channels, where stream gradients are adjusted to the particle size of the sediment present in a given reach, among other factors. Additionally, field and laboratory investigations suggest that small fans are steep when they are associated with mudflow deposition and large fans are steep when associated with high rates of sediment supply from their source areas.

Gradients on some fans in Canada, Afghanistan, New Zealand, and elsewhere have been shown to be segmented. Precise slope determinations along radial lines of some fans in California, for example, generally reveal the presence of three distinct segments, the steepest of which is near the fan apex and the gentlest of which is near the toe. This configuration has been interpreted in terms of two successive uplifts of the mountain source areas, producing steeper channel gradients and, essentially, the deposition of successively steeper fans on preexisting fan surfaces. Other factors, such as channel deepening in the fan apex regions, however, has led to successively gentler depositional areas through time.

The most widely used empirical relationship is that between the area of a fan and its source area or mountain catchment. First established by the American geologist W.B. Bull, this can be expressed in the form of the general power function $A_f = cA_s^m$, in which A_f is fan area, A_s is source area, and c and m are the coefficient and exponent of the relation, respectively. The latter should be constants for a given region or specific rock type in the source areas. Various applications of the relationship have had mixed success. It is clear that the volume of sediment present in a given fan is directly related to the volume removed from the fan's source area by erosion. Two fans of differing steepness can exhibit identical surface areas despite a drastic difference in sediment volume. The use of area measurements for the unobtainable volume measurements is therefore at least partly responsible for the large variation of values of the coefficient c obtained to date.

Surficial features. The relationships among surficial features of alluvial fans in arid regions are well illustrated by the map of the Bat Mountain Fan, Death Valley (Figure 2). This fan is typical of many in the general area, and it consists of four main surficial features, namely, active washes, abandoned washes, desert pavement, and undifferentiated gravel and sand areas. As shown by the map and by the view of one fan farther north, along the east side of the valley (Figure 3), there is a harmony among these features that lends to each fan in the area the appearance of a mosaic.

Desert varnish

The darker hues are generally attributable to the presence of desert varnish on rock fragments. This brownish-black veneer, consisting chiefly of iron and manganese oxides, occurs in many areas of the world only on items of considerable antiquity. Egyptian monuments, arrowheads, and other man-made objects that are known to be about 2,000 years or more in age are examples. In one or two instances, it has been suggested that varnish may start to form within 50 years, but most authorities accept the presence of varnish as evidence that the rock fragment bearing this veneer has not been subjected to much water inundation (because varnish is soluble in water) in the last 2,000 years.

The matter is of some importance with regard to the history of a fan because the presence or absence of varnish must be related to the presence or absence of water flooding. This, in turn, is related to a fundamental distinction between active and abandoned washes: active washes contain flowing water when storm runoff occurs today, whereas abandoned washes do not.

Both active and abandoned washes tend to contain much coarse gravel and sand, interspersed with blocks and boulders that may be as large as ten metres in diameter. Such material is derived from the erosion of crystalline rocks such as granites and quartzites and, in general, from any strongly cemented or bonded rock species. Shales and other weak or easily crumbled rocks cannot yield boulders because any large fragment or block will be broken up during transportation from source area to fan. Whenever coarse material is present in active and abandoned washes, however, it is invariably coated by varnish in the abandoned washes and is slightly larger in average size than the coarse material in active washes. These two distinctions are most suggestive.

The size difference suggests that when flooding did occur in the abandoned washes—at some former time, prior to their hydrological abandonment—the average water discharge must have been greater than that associated with modern runoff events. This inference stems from the basic principle that greater discharges tend to move larger particles and to move particles of any given size farther downfan. Additionally, it seems probable that the climate must have been more humid than the prevailing one because greater precipitation (or lower evaporation rates) usually is associated with greater discharge.

With regard to the presence of varnish, this suggests that the wetter climate responsible for greater water discharge and the movement of coarser gravels may have occurred as long ago as 2,000 years. Phrased differently, it may mean that little runoff has occurred in abandoned washes during the last 2,000 years, which would be the period of varnish formation. This conclusion may well be justified in Death Valley, for example, where other kinds of evidence suggest that a lake nearly 200 metres deep was sustained at the beginning of the Christian Era. Today, in contrast, the lake is gone, ground temperatures as high as 88° C (190° F) have been observed, and water-soluble desert varnish is widespread.

Other surficial features of note may include mudflow levees, which are sinuous bouldery ridges that mark the lateral boundaries of former mudflows on fans; fan-head trenches (Figure 3), which are canyon-like gorges that represent the walls of the active channels near the fan

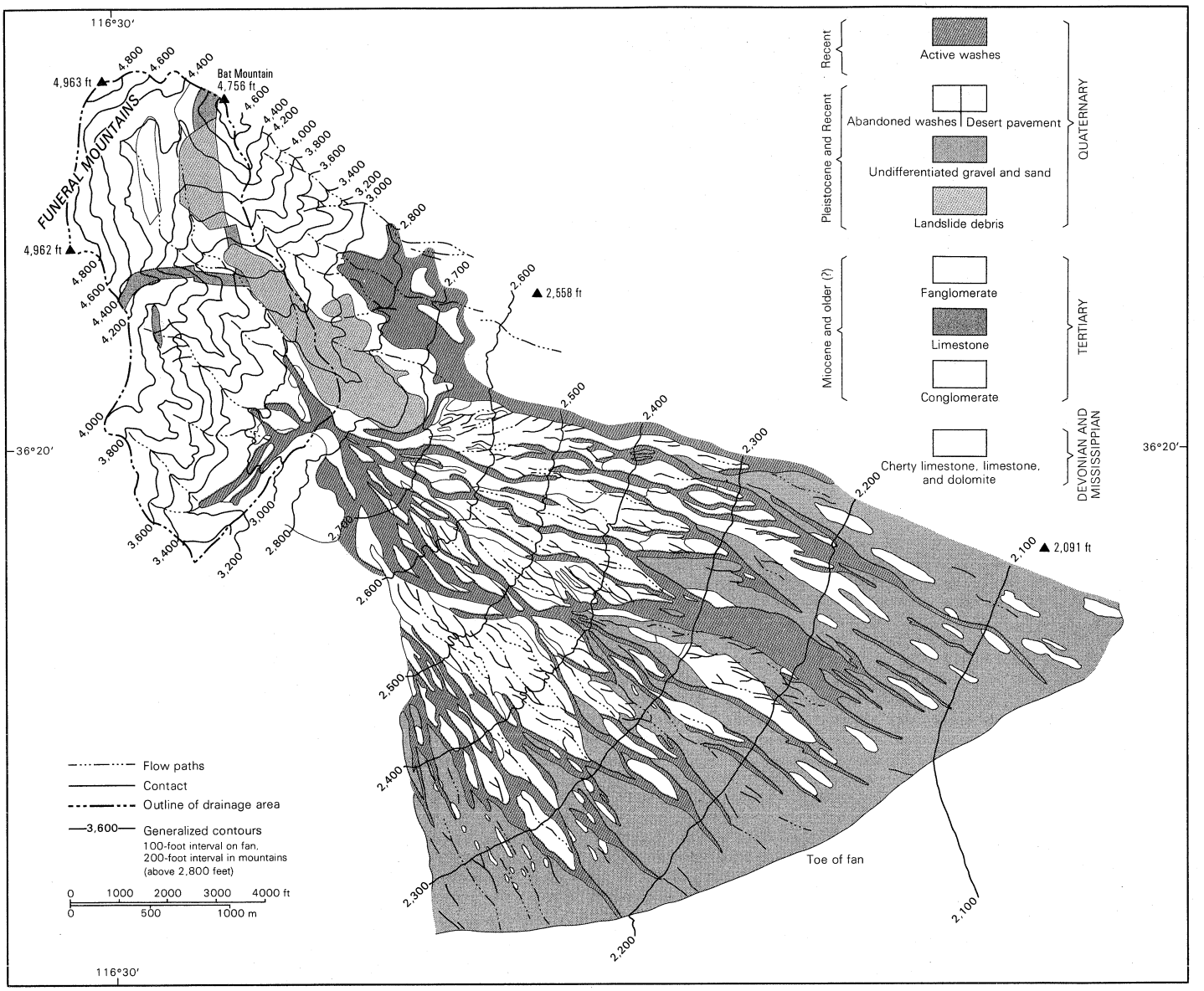


Figure 2: Bat Mountain alluvial fan and its source area, Ash Meadows Quadrangle, California.
By courtesy of United States Geological Survey

Desert pavements

apexes; and desert pavements. The latter are literally pavements that consist of angular fragments of pebbles, cobbles, and boulders. These fit together like a jigsaw puzzle and form an armour on the flat or gently rolling parts of fan surfaces. In the classical view, desert pavements were thought to be the product of wind action (*q.v.*) because fine-grained material is generally absent from pavement surfaces and because the wind, undeniably, does blow across them.

Several modern investigators have suggested that pavements may form as a result of precipitation and runoff on fan surfaces. The basic mechanism invoked involves first the segregation of silt beneath coarser material to form a silt layer. Precipitation of modest amount saturates this layer with water and, under the influence of gravity, it tends to flow in lateral and downfan directions. This leads to the observed interlocking of the larger fragments that form the pavement proper. A number of ancillary observations support this theory, and it seems likely that the wind serves principally to prevent the accumulation of fine material on desert pavements after they have been formed by fluvial and gravitational processes.

Fan sediments and their stratigraphy. The nature of fan sediments is generally variable because, as stated above, particle size is controlled by the rock type in source areas. Alluvial fans that rim the Richardson Mountains in the Northwest Territories of Canada, for example, are largely composed of silt-sized sediment; core drilling has revealed a general absence of sand and gravel. The breakdown of larger particles by frost action may partly be responsible in this Arctic area, but the presence of soft shales and sandstones in the Richardson Mountains is the most important factor. The maximum particle size that may occur in fan sediments is basically limited only by the spacing of fractures in the mountain bedrock, especially wherever tough, coherent rocks are involved. Blocks larger than a small truck may occur in the upper reaches of some fans, and boulders as large as a metre or so in diameter may be transported to lower fan boundaries or the flat valley floors beyond. The nature of fan sediments, therefore, is heterogeneous in most instances: a size distribution that encompasses both boulders and clay in the same fan deposit is not an uncommon occurrence.

Complexity of the stratigraphic arrangement

The stratigraphic arrangement of fan sediments reflects the character of the sediments, climatic factors, and modes of deposition. Fans in humid regions tend to be characterized by annual or seasonal flooding and the presence of a vegetative cover. This may result in a vertical series of buried organic zones that exhibit considerable lateral continuity; sediment is successively de-



Figure 3: (Left) Alluvial fan on the east side of Death Valley, California. (Right) The active wash and fan-head trench of the same fan, viewed from its apex.
By courtesy of L.K. Lustig

posited over each previous year's grass cover. In arid areas, however, flooding occurs much less frequently and is much more likely to take the form of mudflows or to be otherwise laterally restricted by channel walls or fan-head trenches. Depositional sequences, therefore, include linear-to-sinuous lobate tongues of debris that extend from fan apexes downslope to midfan or beyond. The fans are built up by successive shifts in the loci of deposition in a lateral and downfan sense. Thus, nowhere is there likely to occur a vertical series of beds or layers that exhibit lateral continuity of deposition over the entire fan surface. The stratigraphic arrangement consists of a series of lenses or wedges of sediment because successive flow paths may cross and intertwine and because the fan stratigraphy is usually exposed in the wall of a trench (Figure 3) or some other section that cuts across former depositional paths. Even where viewed parallel to a former depositional path, the sedimentary layers generally appear as lenses or wedges. Two successive floods along the same path may be of different magnitude, for example, and this will produce vertical alternation of coarse and fine sedimentary deposits at any particular distance from the source area. In summary, the internal fan structure consists of a complex series of coarse, fine, and mixed-size deposits that alternate, interleave, and wedge out both vertically and laterally. The entire sedimentary assemblage reflects the fluvial processes (*q.v.*) that have led to the deposition of alluvial fans.

SEDIMENT TRANSPORT ON FANS

The nature of sediment transport on fans is principally determined by the characteristics of the drainage basin from which the sediment is derived, the rock types present in this source area, and a number of hydrologic variables that are related to climatic conditions. The effect of rock type has been discussed above (see above *Physical aspects of alluvial fans*), but other drainage basin characteristics of importance include the drainage area, overall relief, steepness and lengths of valley sides and stream channels, abundance and type of vegetation and soil cover, and the permeability of the ground surface with respect to the water that falls upon it or flows over it. A large number of additional factors exist or have been developed empirically, but they are largely derivable from the factors already cited. The drainage density of a basin, for example, is often considered an important factor in treating sediment transport or streamflow; it is computed by dividing total stream channel length by the drainage area of a particular basin. For a description of the several geometric parameters of drainage basins and of the laws that govern drainage networks, see RIVERS AND RIVER SYSTEMS; and SEDIMENT YIELD OF DRAINAGE SYSTEMS.

Basin hydrology and basin geometry

Precipitation–runoff relations. The climatic variables of moment chiefly involve precipitation and evaporation. A large number of indices that incorporate one or both of these factors have been used to provide some quantitative indication of the hydrologic conditions in a drainage basin prior to, and during, a storm-runoff event. Of concern, for example, are soil moisture conditions before the occurrence of a storm, absolute amounts of precipitation, intensity and frequency of precipitation, and precipitation–runoff relations. One widely used index of precipitation effectiveness is the cumulative sum of the monthly precipitation and evaporation ratios. This index provides some general sense of overall moisture availability in a basin, but average values of precipitation are not necessarily the values most closely related to sediment transport on fans, particularly in semi-arid and arid regions. Two areas might receive an average of 360 centimetres of precipitation per year, for example; but if one of these areas received its allotment in equal monthly increments of 30 centimetres, whereas the other area received the entire annual amount in only one or two storm events, both the runoff and sediment transport characteristics of the two areas would differ markedly. Flash floods, peak discharges, and erosion and sediment transport on a catastrophic scale would occur in the area receiving less frequent precipitation.

Many complex statistical relations between drainage basin characteristics and hydrological conditions have been developed, particularly for temperate and humid areas. In a general sense, this would be expected. Larger drainage areas, greater relief, steeper slopes, minimal vegetation, and relatively impermeable ground cover all suggest—and indeed tend to produce—greater runoff or stream discharge under any particular set of hydrologic conditions. Multiple-regression analysis and the advent of modern computers have permitted investigation of the interdependence of all the variables involved. One of the most reliable relationships of simple nature, however, is that between flood discharge and drainage area. This is expressed by the general power function $Q_t = gA^d$, in which Q_t is the flood discharge of any particular return period or frequency of occurrence t, A is drainage area, and g and d are constants for a given region. The value of the exponent d ranges from 0.5 to 1.0, and for a semi-arid area, such as central New Mexico, it typically equals 0.75, approximately.

Flow regimes on fans. Sediment transport on fans is, therefore, intimately related to a host of variable factors and conditions in the source area or drainage basin. If a supply of highly erodible sediment is present in a par-

ticular source area, and if the basin characteristics and hydrologic input favour sediment transport, then the fan that is fed by this source area will receive much sediment. Actual types of flow onto fan surfaces, however, may range from relatively clear water to mudflows or debris flows.

Clear water flows and mudflows

Clear water flows on alluvial fans are similar to those on other landforms or in other environments. The same physical and hydraulic laws apply and the sedimentary deposits that result tend to be well sorted according to size. A basic difference in flow characteristics does exist, however, because a drainage inversion occurs between flow in the source area and flow on its alluvial fan. Tributary streams in the source area continually enter and join the trunk channel along its course, and for this reason water discharge increases between drainage divide at the mountain crest and the channel outlet at the mountain front. At this point the source area terminates and the fan essentially begins. Once this channellized flow emerges onto a fan it is free to subdivide at some point downfan where it becomes unconfined. Thus, the tributary–trunk stream role in the mountain drainage basin becomes reversed; on the fan surface the trunk stream feeds distributary channels and water discharge diminishes along its course. Moreover, most fans consist of relatively permeable sediments and additional discharge is lost because of downward water percolation. These discharge losses in a downfan direction promote an increased sediment concentration within the water–sediment mixture during flow of any type; this leads to sediment deposition on fan surfaces.

When the source area contains much fine-grained sediment, the runoff from an intense rainstorm may incorporate this material and flows can attain very high densities. One published value is 2.5 grams per cubic centimetre for a mudflow that occurred at Wrightwood, California in 1941. In such dense and viscous flows, the weight of a submerged boulder may be reduced by as much as 60 percent, compared to its weight in water, and for this reason the transporting competency of mudflows is very great. Typically, mudflows result from cloudburst floods in arid regions that may occur in a particular source area only once in 50 to 100 years or even less. In humid areas mudflows generally are produced by precipitation amounts and intensities that greatly exceed average values, as in the Guil basin, France, in 1957; the Grand Teton, Wyoming, in 1941; and elsewhere. Deposits in all cases tend to consist of poorly sorted mixtures of cobbles and boulders, transported together with sand, silt, and clay, depending upon the material present in the source area.

Relatively clear water flows commonly follow on the heels of mudflows, possibly because most of the readily erodible material is flushed out of a mountain source area by the generation of a mudflow but also because rain may continue to fall during and after the event. The lobate tongues and sheets of ill-sorted mudflow debris may be dissected by these subsequent clear water flows and the eroded sediment transported farther downfan. This erosion also may lead to the creation of new channels on the fan and to cobble and boulder deposits that have been winnowed of their clay to sand matrix. Such openwork accumulations resemble a sieve. They are essentially identical in appearance to openwork gravels that accumulate when no fine grained matrix material is available; such gravels have been called sieve deposits. The described erosion generally makes it difficult to determine the proportion of any particular fan that may be ascribable to mudflows, clear water flows, and any other flow of intermediate density and viscosity.

Rates of sedimentation. Volumetric data on fan sediments are generally unavailable in any case, and this has led to great difficulty in assessing rates of fan erosion or deposition. In addition, a time datum or reference surface clearly is required and this too usually is lacking.

A 700,000-year-old example

The Milner Creek Fan, on the west side of the White Mountains, California, is nearly unique in these respects, however. It has been deposited upon the Bishop Tuff, a relatively flat-lying volcanic rock formation of known age—700,000 years old according to potassium–argon radiometric dates. This means that the bulk of the fan formed during the last 700,000 years, and one study of the fan and its probable volume suggests that an average sedimentation rate of about 7.5 to 20 centimetres per 1,000 years was responsible. Knowledge of the volume of material added to this fan during one flood event of recent vintage further suggested that a flood of such magnitude need occur only once in 350 years to produce the fan within the allotted time span.

Several approximations are involved in such calculations of sedimentation rates. Intermittent erosion of fan materials and the effects of the several climatic changes on depositional rates during the last 700,000 years, for example, are not considered. The values provided do serve to indicate, however, that alluvial fans are landforms that originated during the Pleistocene Epoch (from about 10,000 to 2,500,000 years ago) in many instances and that relatively infrequent floods are capable of yielding the volumes of sediment involved.

FORMATION OF FANS

The formation of fans would seem obvious in light of the foregoing sections of this article, and in the broad sense this is quite true. Alluvial fans form at the base of any highland mass where emerging streams become unconfined and where hydraulic conditions require deposition of the sediment being transported. Although agreement on this proposition is universal, the evolution of alluvial fans, like landform evolution in general, is subject to some debate by those who have considered the matter in detail. Specific points of contention involve the relative importance of tectonism (earth movements) and climate and the question of whether alluvial fans are equilibrium landforms that are in adjustment with presently operative processes. These matters remain unresolved save in specific instances or areas, but some generalizations can be made here.

Tectonism and climate. The relative influence of tectonism and climate can best be discussed in relation to the drainage networks in a particular source area and on its associated fan; these networks are part of a single fluvial system. The trunk channel in a source area tends to continue smoothly onto the related fan, where it is designated an active wash or active channel. The physical relationships that obtain suggest that the entire channel system reflects a balance or adjustment: sediment calibre and load, channel slope, form, and dimensions, and the water discharge are all mutually interrelated. Taken together, these factors constitute the hydraulic geometry of the system; and because they tend toward mutual adjustment, change in any one of the variables must produce change in one or more of the others.

If uplift of the mountain source area results in steepening of the trunk channel, then one possible consequence would be the construction of a new fan segment at the apex of the existing fan. Deposition would continue until the gradient of the new segment became adjusted to the steepened gradient in the source area. More commonly, however, authorities who consider tectonism a dominant factor in fan formation correlate steepening of channel gradients—or of entire fans in some instances—with downcutting and channel entrenchment. The basis for this view is evident from some simplified expressions for sediment transport and streamflow.

Tractive force, velocity of flow, and gradient

The stress placed by a fluid medium upon a particle resting on a fan surface or channel bottom is equal to the product of the specific weight of the fluid, the depth of flow, and the gradient or slope—namely, $\tau = \gamma d S$, in which τ is tractive force or applied stress, γ is specific weight of the medium, d is depth of flow, and S is slope. It follows from this relation that the tractive force, which is the force that will move particles and thus promote erosion or downcutting, is directly proportional to slope. Moreover, it is well known that particle motion is related to the velocity of flow; greater velocities tend to move larger particles and a greater number of particles and should also promote downcutting in many instances. The relation between velocity of flow and channel slope

is given as follows: $v = (1.486/n)R^{2/3}S^{1/2}$, in which v is velocity of flow, R is the hydraulic radius of the channel (the cross sectional area divided by the wetted perimeter), S is slope, and n is the roughness, a number that indicates the departure of the channel bed and banks from an approximately smooth condition. Again, it is evident that an increase in slope S will be related to increased sediment transport, in this case by reason of an associated increase in the velocity of flow.

The hydraulic radius R also must be considered, however, particularly because the equation shows that a given increase in this factor will increase flow velocity to a greater degree than will an increase in slope (velocity increases as the two-thirds power of R and as the one-half power of S). This is significant because the hydraulic radius of an alluvial channel is closely related to water discharge (actually, to the bank-full discharge) and water discharge is clearly related to climate, in the broad sense. The question of the relative influence of tectonic movements and climate on fan development must therefore be given an answer that is a compromise: both factors are of demonstrable importance and neither can be ignored in most areas. This follows from the fact that both climatic changes (*q.v.*) and tectonic events have occurred during the period of fan formation in most parts of the world.

Models of fan development. Granting that both tectonism and climate affect sediment transport, erosion and deposition, and thus fan formation, there remains the question of which of three differing models of evolution best fits and describes the development of fans through time. The models involved may best be called the erosion cycle model, the dynamic equilibrium model, and the intermediate, climatically linked model. These models and their distinctions as related to alluvial fan development essentially are as follows.

The erosion cycle model emphasizes the importance of initial uplift of a mountain mass. Given this uplift, and the geographic setting of mountain block and adjacent basin, fan formation is assumed to proceed through a series of distinct and inexorable stages of growth. A particular group of fans is considered to grow and enlarge continuously from the time of initial uplift until the basin or depositional site is completely filled with sediment. This depth of filling would equal the height of the eroded mountain remnants and would produce a relatively plane surface—the ultimate stage of this sequence of events.

The model makes no allowance for climatic or tectonic events subsequent to initial uplift, and, even aside from this fact, it would appear unlikely that fan growth can proceed in a totally uninterrupted fashion. If deposition is occurring upon a fan apex during some particular time interval, for example, this situation cannot proceed indefinitely. Sooner or later the material deposited will produce a change in channel profile or gradient such that one of two alternate possibilities will occur. These are lateral channel shifting and marginal deposition in the apex region or erosion and downcutting in the apex region and a migration of the principal deposition site farther downfan. In either case, changes in the loci of deposition and erosion are involved; such changes do occur because they are reflected by the complexities of fan stratigraphy (see above *Physical aspects of alluvial fans*).

Fan equilibrium in humid and arid regions

Both remaining models of fan development are concerned with these and other changes through time. The dynamic equilibrium model postulates that fans tend to attain a steady state or equilibrium condition in which the sediment eroded from a fan balances nicely the increment of sediment that is added from its source area. Hence, a fan in equilibrium would neither increase nor decrease in bulk at any instant in time. Its general configuration as a landform would be a reflection of the balance between materials and operative processes, and the historical imprint of time or evidence of former processes and events would be obliterated.

There is a large body of evidence to support the dynamic equilibrium model in nature, but with respect to alluvial fans it seems probable that the theory best fits fans in temperate or humid areas. The fans in river valleys in Sweden, Alaska (Figure 1), and similar regions, for example, clearly are subject to a balance between erosion and deposition today. If their sediment supply were cut off, the streams in the trunk valleys would soon remove these fans by erosion. Similarly, if streamflow should cease in the trunk valleys, the fans would soon build across the empty channels to the opposite valley walls.

In arid regions, however, there is some reason to doubt that alluvial fans are at or near a steady state today. The presence of desert varnish, elevated beach deposits, playa lakes, and kindred phenomena all provide testimony to former climatic conditions that were cooler and wetter at times than are prevailing climates. It might be said, in fact, that arid areas best preserve the historical imprint of time, or the evidence of past events. This being true, it follows that the alluvial fans in these regions are likely to exhibit some degree of departure from equilibrium conditions. Indeed, the third model of fan development simply holds that climatic conditions strongly influence rates of fan growth and the loci of deposition on fans. When climates change from arid to temperate or humid conditions, equilibrium may be attained quite rapidly, because of an increase in the magnitude and frequency of discharge, for example. When the reverse occurs, and climates become arid, however, there is likely to be a time lag of several thousands of years before a balance is struck between form and processes; this is the hallmark of the steady state (see further LANDFORM EVOLUTION).

BIBLIOGRAPHY. Support may be found for each of the described viewpoints both in nature and in the literature on fans. The latter is somewhat specialized and the interested reader should consult the following papers, each of which contains references to other relevant works: W.B. BULL, "Geomorphology of Segmented Alluvial Fans in Western Fresno County, California," *Prof. Pap. U.S. Geol. Surv. 352-E* (1964); C.S. DENNY, "Alluvial Fans in the Death Valley Region, California and Nevada," *ibid., 466* (1965); L.K. LUSTIG, "Clastic Sedimentation in Deep Springs Valley, California," *ibid., 352-F* (1965); and R.L. HOOKE, "Processes on Arid-Region Alluvial Fans," *J. Geol.*, 75:438–460 (1967).

(L.K.L.)

Alp-Arslan

Alp-Arslan was the second of the three first great sultans of the Seljuqs, the ruling family of the Oğuz Turkmen tribes who invaded western Asia in the 11th century. His fame rests mainly on his victories against the Byzantine Empire. He followed his uncle, Toghrïl Beg, who, through the conquest of Iran and Baghdad, had created the empire that the third sultan, Alp-Arslan's son Malik-Shāh, was later to organize. Alp-Arslan's military exploits are regarded as laying the foundation for the modern state of Turkey.

Early life

Born about 1030, Alp-Arslan was the son of Chaghri Beg, the ruler of Khorāsān in Iran, and the nephew of Toghrïl, the governor of western Iran, the base of Seljuq expansion. In 1061 his father died. When, in 1063, his uncle died without issue, Alp-Arslan became sole heir to all the possessions of the dynasty except Kerman, in southern Iran, which was held by one of his brothers, whom he promptly reduced to vassalage. He likewise easily eliminated the son of one of Toghrïl's widows, as well as Qutlumush, a cousin and rival.

Born outside the traditional Muslim countries that he was later to govern, he left their administration to his vizier, Niẓām al-Mulk, who continued as administrator under Malik-Shāh. While maintaining control of Iraq, Alp-Arslan nevertheless shunned that country in order to avoid such clashes of interests with the caliphate, the seat of which was there, as had complicated Toghrïl's last days.

Alp-Arslan's political activity was based on the ideas that inspired all the three great Seljuq sovereigns. In Central Asia, peace was maintained with the Ghaznavid rulers who were hard to track down in their mountain strongholds in India, whereas against the Qarakhanids of Transoxania, force was used. In the west, where Alp-Arslan was to gain all his glory, he was faced with a

more complicated situation. On the one hand, he decided to go to Egypt to crush the Ismāʿīlī Fāṭimid heresy, which the ʿAbbāsid Sunnite caliphate at Baghdad, whose protector he was, would not accept. On the other hand, he was aware of the necessity of keeping his influence over the Turkmens, which was essential to his military strength. The Turkmens were interested above all in the success of the holy war against the infidels and in raids on Christian territory. Against the Byzantines and their Armenian and Georgian neighbours, Alp-Arslan conducted a series of campaigns, which were extended by attacks from autonomous Turkmen bands. In 1064 he seized Ani, the former Armenian capital, and Kars. These operations resulted only in some consolidation of boundaries, which assured the Turkmens control over pastureland on the Aras River. Nevertheless, although the Turkmens returned to Muslim territory to store away their booty, these expeditions upset the Byzantine defense system and prepared the subsequent Turkish conquest of Asia Minor. They resulted in Byzantine reactions in Syria and Armenia, after which the two empires began to negotiate.

Campaigns against the Armenians and Byzantines

Alp-Arslan then judged himself sufficiently protected on the Byzantine side to undertake, at the request of Egyptian rebels, the great anti-Fāṭimid expedition that had been asked for by the orthodox ʿAbbāsid caliphate. As he was about to attack Aleppo, whose prince was too late in siding with the ʿAbbāsids, and was preparing to occupy Syria, Alp-Arslan learned that the Byzantine emperor Romanus IV Diogenes, with a formidable army, was assaulting his rear army in Armenia. Quickly retracing his steps, he faced his adversary near Manzikert in August 1071. The Byzantine army, powerful in numbers but weak in morale, fell before the outnumbered but dedicated Turks. By evening the Byzantine army was defeated, and, for the first time in history, a Byzantine emperor had become the prisoner of a Muslim sovereign. Alp-Arslan's goal was not to destroy the Byzantine Empire: he was content with the rectification of boundaries, the promise of tribute, and an alliance. But the Battle of Manzikert opened Asia Minor to Turkmen conquest. Later, every princely family in Asia Minor was to claim an ancestor who had fought on that prestigious day.

Victory of Manzikert

The Sultan's triumph was followed by a commonplace death, permitting moralists to recall that power rests only in God: at the end of 1072, he had returned to the Qarakhanid frontier, and during a quarrel, was mortally wounded by a prisoner. He had designated as his heir his son Malik-Shāh, 13 years old, under the guardianship of Niẓām al-Mulk.

Alp-Arslan's personality, in spite of the glory surrounding his name, is not easy to evaluate. Muslims see in him a great captain, a trainer of men, an honest man, an enemy of all treachery. Christians, contrasting his reputation with that of his son Malik-Shāh, paint him in harsher colours. There is no doubt that conquest seems to have been his favourite pastime. Although an anonymous writer dedicated to him the *Malek-nāmeh*, an attempt to trace the origins of his family and the empire, Alp-Arslan appears to have shown little interest in intellectual matters, leaving them, like the administration of his empire, to his vizier.

Personality

BIBLIOGRAPHY. Most works on Alp-Arslan are in Turkish. In English, W. BARTHOLD, *Turkestan Down to the Mongol Invasion*, 2nd ed. translated from the Russian by H.A.R. GIBB (1928; new rev. ed. by C.E. BOSWORTH, 1968); CLAUDE CAHEN, *Pre-Ottoman Turkey* (1968); and the article on Alp-Arslan in the *Encyclopaedia of Islam*, new ed., vol. 1 (1960), may all be recommended.

(C.C.)

Alphabets

An alphabet is a set of symbols or characters used to represent the sounds of a language; in most alphabets the symbols are arranged in a definite order, or sequence (*e.g.*, *A*, *B*, *C*, etc.). Each of the characters usually represents either a consonant or a vowel, rather than a syllable or group of consonants and vowels. As a result, the number of characters required can be held to a minimum. A language that has 30 consonant sounds and five vowels, for example, needs only 35 separate letters. In a syllabary, on the other hand, the same language would require 30 times five symbols to represent each syllable (*e.g.*, separate forms for *ba, be, bi, bo, bu; da, de, di,* and so on), and an additional five symbols for the vowels, thereby making a total of 155 individual characters. Both syllabaries and alphabets are phonetic symbolizations; that is, they are symbols to record the sounds of words rather than to represent objects or ideas being expressed. (With the passage of time some languages, notably English, failed to synchronize spelling changes with changes in pronunciation, so that these languages can no longer be said to have totally phonetic spelling.)

Pictography, or picture writing, was the first step toward true writing. Ideography, or ideographic writing, extended the possibilities of pictography. The earliest fully developed systems of writing, those of the ancient Near East (Mesopotamia and Egypt), China, and pre-Columbian Central America and Mexico, along with others, were not, however, purely pictographic, ideographic, or phonetic. The oldest of these developed scripts was the cuneiform; the second oldest was the Egyptian hieroglyphic style.

The word alphabet, from the first two letters of the Greek alphabet—alpha and beta—was first referred to in its Latin form, *alphabetum*, by Tertullian (died AD 230), a Latin ecclesiastical writer and Church Father, and by St. Jerome. The classical Greeks customarily used the plural of *to gramma*, "the letter," while later the form *alphabētos* was adopted, probably under Latin influence (see also WRITING, FORMS OF; HIEROGLYPHIC WRITING).

Derivation of the term alphabet

THEORIES OF THE ORIGINS OF THE ALPHABET

Evidence for the original alphabet is scarce, but such as it is, it comes exclusively from the lands bordering the eastern shores of the Mediterranean, including ancient Canaan and Phoenicia, and belongs to the period between 1700 and 1500 BC. This alphabet is known as the North Semitic.

Over the centuries, various theories have been advanced to explain the origin of alphabetic writing, and, since classical times, the problem has been a matter of serious study. The Greeks and Romans considered five different peoples as the possible inventors of the alphabet—the Phoenicians, Egyptians, Assyrians, Cretans, and Hebrews. Modern theories include some that are not very different from those of ancient days. Every country situated in or more or less near to the east Mediterranean has been singled out for the honour. Egyptian writing, cuneiform, Cretan, Hieroglyphic Hittite, the Cypriot syllabary, and other scripts have all been called prototypes of the alphabet. The Egyptian theory actually subdivides into three separate theories, according to whether the Egyptian hieroglyphic, the hieratic, or the demotic script is regarded as the true parent of alphabetic writing. Similarly, the idea that cuneiform was the precursor of the alphabet may also be subdivided into Sumerian, Babylonian, and Assyrian cuneiform.

Egyptian theory of the origin of the alphabet

Among the various other theories concerning the alphabet are the hypotheses that the alphabet was brought by the Philistines from Crete to Palestine, that the various ancient scripts of the Mediterranean countries developed from prehistoric geometric symbols employed throughout the Mediterranean area from the earliest times, and that the Proto-Sinaitic inscriptions (discovered since 1905 in the Sinai Peninsula) represent a stage of writing intermediate between the Egyptian hieroglyphics and the North Semitic alphabet. Another hypothesis, the Ugaritic theory, evolved after an epoch-making discovery in 1929 (and the years following) at the site of the ancient Ugarit, on the Syrian coast opposite the most easterly cape of Cyprus. Thousands of clay tablets were found there, documents of inestimable value in many fields of research (including epigraphy, philology, and the history of religion). Dating from the 15th and 14th centuries BC, they were written in a cuneiform alphabet of 30 letters.

The Early Canaanite theory was based on several un-

deciphered inscriptions also discovered since 1929 at various Palestinian sites; the writings belong in part to *c.* 1700 BC and are thus the earliest preserved documents in an alphabetic writing.

Despite the conflict in theories, leading scholars are generally agreed that for about 200 years before the middle of the 2nd millennium BC, alphabet making was in the air in the Syro-Palestinian region. It is idle to speculate on the meaning of the various discoveries referred to. That they manifest closely related efforts is certain; what the exact relationship between these efforts was, and what their relationship with the North Semitic alphabet was, cannot as yet be said with certainty.

Conditions favourable for alphabetic writing

It can, however, be ascertained that the period from 1730 to 1580 BC in Syria, Palestine, and Egypt, during which there was an uprooting of established cultural and ethnic patterns in the Fertile Crescent, provided conditions favourable to the conception of an alphabetic script, a kind of writing that would be more accessible to larger groups of people, in contrast to the scripts of the old states of Mesopotamia and Egypt, which were confined largely to the priestly class. In default of other direct evidence, it is reasonable to suppose that the actual prototype of the alphabet was not very different from the writing of the earliest North Semitic inscriptions now extant, which belong to the last two or three centuries of the 2nd millennium BC. The North Semitic alphabet was so constant for many centuries that it is impossible to think that there had been any material changes in the preceding two to three centuries. Moreover, the North Semitic languages, based as they are on a consonantal root (*i.e.*, a system in which the vowels serve mainly to indicate grammatical or similar changes), were clearly suitable for the creation of a consonantal alphabet.

The inventor or inventors of the alphabet were, no doubt, influenced by Egyptian writing—perhaps also by other scripts. Indeed, it is probable that the man or men who invented the alphabet were acquainted with most of the scripts current in the eastern Mediterranean lands at the time. Though the nationality of the inventor or inventors of the alphabet is unknown, it is now generally agreed that he or they belonged to the Northwest Semitic linguistic group, which includes the ancient Canaanites, Phoenicians, and Hebrews.

Letters as signs

It seems probable that the original letters were conventional signs and not, as is still held by some, pictures used as ideograms (*e.g.*, the letter 'alef representing an ox, bet a house, gimmel a camel, and so on). Indeed, the great achievement in the creation of the alphabet was not the invention of signs but the inner working principle; that is, the production of a system in which each sound is represented by one symbol and each symbol generally represents one sound. The principle governing the conventional names of the letters is known as acrophony—that is, the value of each consonant is the value of the first letter of its name: *b* of "bet," *g* of "gimmel," *d* of "dalet," and so on. These names were not derived from pictographic representations of the letters but were an artificial, mnemonic device similar to those used in modern ABC books for children.

The North Semitic alphabet remained almost unaltered for many centuries. If the signs' external form (which, it must be emphasized, had no particular significance) is ignored, and only their phonetic value, number, and order are considered, the modern Hebrew alphabet may be regarded as a continuation of the original alphabet created more than 3,500 years ago. The Hebrew order of the letters seems to be the oldest. The earliest evidence that the Hebrew alphabet was learned systematically was a schoolboy's scribbling on the vertical face of the upper step of a staircase leading up to the palace at Lachish, in southern Israel. It included the scratching of the first five letters of the early Hebrew alphabet in their conventional order, and it belonged to the 7th or 8th century BC.

DEVELOPMENT AND DIFFUSION

At the end of the 2nd millennium BC, with the political decay of the great nations of the Bronze Age—*i.e.*, the Egyptians, Babylonians, Assyrians, Hittites, and Cretans—a new historical world began. In Syria and Palestine, the geographical centre of the Fertile Crescent, three nations—Israel, Phoenicia, and Aram—played an increasingly important political role. To the south of the Fertile Crescent, the Sabaeans, a South Arabian people (also Semites, though South Semites), attained a position of wealth and importance as commercial intermediaries between the East and the Mediterranean. To the west, seeds were sown among the peoples who later constituted the nation of Hellas—the Greeks. As a result, an alphabet developed with four main branches: (1) the so-called Canaanite, or main branch, subdivided into Early Hebrew and Phoenician varieties; (2) the Aramaic branch; (3) the South Semitic or Sabaean branch; and (4) the Greek alphabet, which became the progenitor of the Western alphabets, including the Etruscan and the Latin. The Canaanite and Aramaic branches constitute the North Semitic main branch.

Development of four main branches of the alphabet

The Canaanite alphabet. The two Canaanite branches may be subdivided into several secondary branches. First, Early Hebrew had three secondary branches—Moabite, Edomite, and Ammonite—and two offshoots—the script of Jewish coins and the Samaritan script, still in use today, for liturgical purposes only. Second, Phoenician can be divided into Phoenician proper and "colonial" Phoenician. Out of the latter developed the Punic and neo-Punic scripts and probably also the Libyan and Iberian scripts.

Early Hebrew. The term Early Hebrew is used to distinguish this branch from the later so-called Square Hebrew. The Early Hebrew alphabet had already begun to acquire its distinctive character by the 11th century BC. It was used officially until the 6th century BC and then lingered on for several centuries. In a stylized form, it was used on Jewish coins from 135 BC to AD 132–135. The most ancient example of Early Hebrew writing is that of the Gezer Calendar of the period of Saul or David (*i.e.*, *c.* 1000 BC). The oldest extant example of the Early Hebrew ABC is the inscription of the first five letters of the Hebrew alphabet, faintly scratched on the steps of the palace in Lachish, perhaps dating from the 8th century BC. A cursive style reached its climax in the inscriptions at Lachish, dating from the beginning of the 6th century BC. The Leviticus and other small Early Hebrew fragments found in the Dead Sea caves, which are probably from the 3rd century BC, are the only remains of what is considered to be the Early Hebrew book, or literary, hand.

Paramount importance of the Phoenician alphabet

Phoenician. It is difficult to overestimate the importance of the Phoenician alphabet in the history of writing. The earliest definitely readable inscription in the North Semitic alphabet is the so-called Ahiram inscription found at Byblos in Phoenicia (now Lebanon), which probably dates from the 11th century BC. There is, however, no doubt that the Phoenician use of the North Semitic alphabet went further back. By being adopted and then adapted by the Greeks, the North Semitic, or Phoenician, alphabet became the direct ancestor of all Western alphabets. Only very few inscriptions have been found in Phoenicia proper. This rarity of indigenous documents is in contrast to the numbers of Phoenician inscriptions found elsewhere—on Cyprus, Malta, Sicily, and Sardinia, and in Greece, North Africa, Marseille, Spain, and other places.

The Aramaic alphabet. The adaptation of the North Semitic alphabet to the Aramaic language took place at some time in the 10th century BC, when Aramaic was spoken in several petty kingdoms in northern Mesopotamia and Syria, the most important of them being Dammeshek (Damascus). The process of the re-establishment of the Assyrian Empire and its hegemony over a good part of the Near East began in the 9th century. One after another, the Aramaean states gave way under Assyrian onslaught. Dammeshek, the last survivor, fell in 732 BC. The end of Aramaean political independence marked the beginning of Aramaean cultural and economic supremacy in western Asia. The transplantation of masses of Aramaeans by the Assyrians, a political measure designed to break up military alliances, bore remarkable

fruit. By the end of the 8th century BC, the use of the Aramaic language and alphabet had become very widespread in Assyria itself; by the end of the following century all of Syria and a large part of Mesopotamia had become thoroughly Aramaized.

Aramaic inscriptions

On the whole, the few early Aramaic inscriptions that have been found belong to the 9th, 8th, and 7th centuries BC. Inscriptions from the 6th and later centuries are more numerous; the increase reflects the rapid spread of the Aramaic alphabet throughout the Near East. Numerous Aramaic papyri and ostraca (inscribed pottery fragments) have been found in Egypt; the earliest of these can be dated to *c.* 515 BC, while the most famous are the Elephantine papyri, containing information of a religious and economic nature about a 5th-century Jewish military colony in Egypt. Aramaic inscriptions have been found in northern Arabia, Palestine, Lycia, Cappadocia, Lydia, Cilicia, Assyria, and as far afield as Greece, Afghanistan, and India.

Almost as if by pre-arrangement, all of the alphabetic scripts west of Syria seem to have been derived, directly or indirectly, from the Canaanite alphabet, whereas the hundreds of alphabetic writings of the East apparently have sprung from the offshoots of the Aramaic alphabet. On the whole, the direct and indirect descendants of the Aramaic alphabet can be divided into two main groups: the scripts employed for Semitic languages, and those adapted to non-Semitic tongues. With regard to the Semitic offshoots, six separate alphabets may be discerned: the Hebrew, the Nabataean-Sinaitic-Arabic, the Palmyrene, the Syriac-Nestorian, the Mandaean, and the Manichaean. Some of these alphabets became links between the Aramaic alphabet and the numerous scripts used for the non-Semitic languages of central, southern, and southeastern Asia.

Adaptations of the Aramaic alphabet

Among these scripts, which were directly or mainly indirectly adapted to non-Semitic languages from the Aramaic alphabet, are: (1) the Persian (Iranian) scripts known as Pahlavi, which were used for writing the sacred (pre-Islāmic) Persian literature, etc.; (2) Sogdian, a script and language that constituted the lingua franca of Central Asia in the second half of the first millennium AD; (3) Kök Turki, a script used from the 6th to the 8th century AD by Turkish tribes living in the southern part of central Siberia, in northwestern Mongolia, and in northeastern Turkistan (this alphabet was the prototype of the early Hungarian alphabet); (4) the alphabet of the Uighurs, a Turkic-speaking people who lived in Mongolia and eastern Turkistan in the early 13th century; this script was adapted, with Tibetan influence, and adopted as the writing of the Mongol Empire (the so-called Kalika script); (5) the early scripts of the Mongols, including Kalmuck, Buryat (Buriat), Mongolian proper, and the allied Manchu alphabet.

The Aramaic alphabet was probably also the prototype of the Brāhmī script of India, a script that became the parent of nearly all Indian writings. Derived from the Aramaic alphabet, it came into being in northwest India. The Armenian and Georgian alphabets, created by St. Mashtots (Mesrob) in the early 5th century AD, were also based on the Aramaic alphabet.

The South Semitic alphabet. The South Semitic, or Sabaean, branch remained within the confines of the Arabian Peninsula for most of its history. It was in use at the beginning of the 1st millennium BC. The most that can be said about its origins is that it neither developed from, nor directly depended upon, the North Semitic alphabet. It may have been derived, ultimately, from the Proto-Sinaitic script, with some influence from the North Semitic. Offshoots from the South Semitic branch included the Minaean, Himyaritic, Qatabanic, and Hadhramautic alphabets in southern Arabia, and Thamudene, Dedanite, and Safaitic alphabets in the northern part of the peninsula. Numerous inscriptions in these alphabets are the principal source for the study of those once-flourishing kingdoms, including the biblical Sheba, relegated by the rise of Islām to the backwaters of history.

The Sabaean offshoot, a graceful and elegant script consisting of 29 letters, spread into Africa, where it became the progenitor of the Ethiopic alphabet; this in turn gave birth to the modern Amharic, Tigré, Tigriña, and other alphabets of modern Ethiopia. These are the only South Semitic scripts still in use today.

The Greek alphabet. As with so many other things, the importance of the ancient Greeks in the history of the alphabet is paramount. All of the alphabets in use in Europe today are directly or indirectly related to the Greek. Although the Greek alphabet was an adaptation rather than an invention, it was such an improvement that it has remained for 3,000 years—with only slight modifications—an unparalleled vehicle of expression and communication for men of the most diverse nationalities and languages. The Greek alphabet, created early in the 1st millennium BC, spread in various directions in Asia Minor, Egypt, Italy, and other places, but far and away the most important developments of it were the Etruscan-Latin and the Cyrillic alphabets.

From D. Diringer, *The Alphabet: A Key to the History of Mankind*

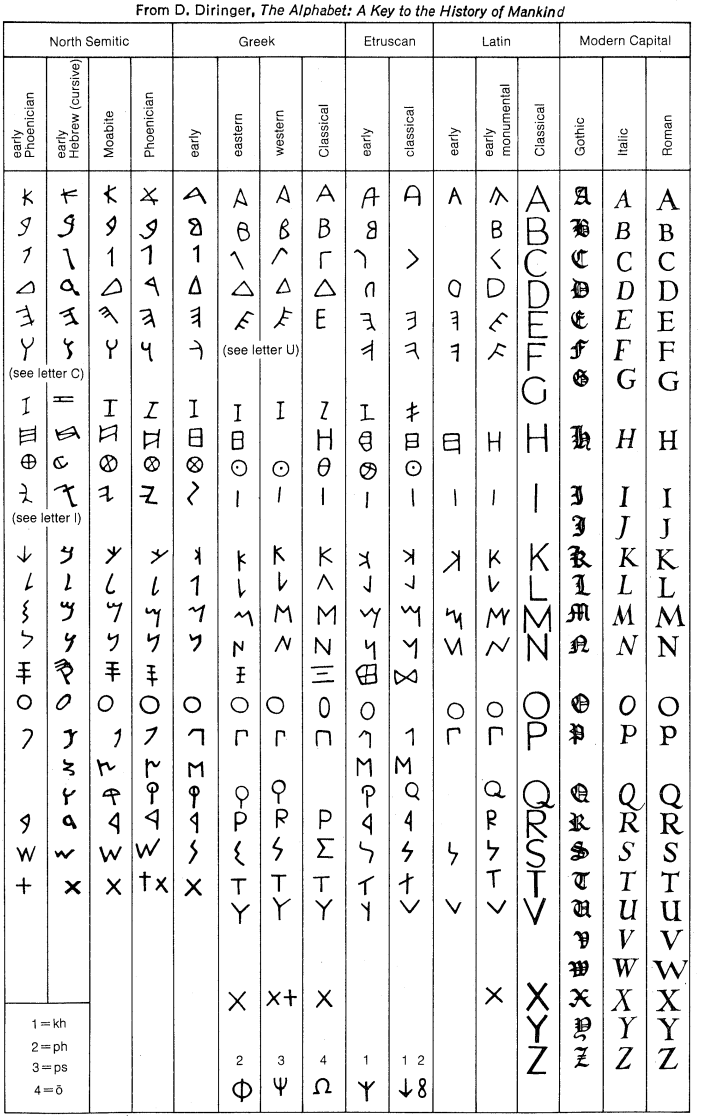


Figure 1: Alphabet characters from North Semitic to modern Roman capital letters. The North Semitic, Greek, and Etruscan symbols do not always designate the same sound as the modern letters.

Theories explaining diffusion. There is no complete agreement as to how or why certain alphabets have come to dominate the world. Some scholars believe that the alphabet follows the flag; that is, that the diffusion of the alphabet results from political and military conquests. Others hold that the alphabet follows trade. More accurate, perhaps, is the theory that the alphabet follows religion. A few examples may illustrate the point: (1)

The theory that the alphabet follows religion

The Latin language and script were carried by Roman legionaries and Imperial officers to all parts of the vast Roman Empire, particularly to the regions that were not Hellenized. In later centuries, however, churchmen and missionaries carried the Latin language and script still further afield. The ascendancy of Latin led to the adoption of the Latin (Roman) alphabet by a large majority of nations; it became used for tongues of the most diverse linguistic groups, not only in Europe but in all other parts of the world as well. (2) Two alphabets, the Cyrillic and the Latin, are used for writing Slavic languages. Cyrillic is used by those Slavic peoples who accepted their religion from Byzantium, whereas Roman Christianity brought the use of the Latin alphabet to the Poles, Lusatians, Wends, Czechs, Slovaks, Slovenes, and Croats. Particularly interesting is the case of Yugoslavia, in which a single language is written differently by the Catholic Croats and the Greek Orthodox Serbs. (3) The Arabic alphabet is, after Latin, the most generally used in Asia and Africa. The rise of Islām in the 7th century AD and the tremendous Islāmic expansion and conquest carried the Islāmic holy book, the Qur'ān, written in the Arabic alphabet, over a vast area: the Near and Middle East, North and Central Africa, South and Southeast Asia, and even southern Europe. The Arabic alphabet was, therefore, adapted to Semitic and Indo-European forms of speech, to Tatar-Turkish, Iranian, and Austronesian (Malayo-Polynesian) tongues, and to several African languages. (4) The movement eastward from India of the Indian Brāhmī-Buddhist alphabets was much more peaceful than that of the Arabic alphabet. These offshoots, which took root in Ceylon, Burma, Thailand, Cambodia, Laos, Vietnam, Indonesia, and the Philippines, were again the result of the spreading of a religion —Buddhism, but by missionaries and not by armies.

MAJOR ALPHABETS OF THE WORLD

Hebrew alphabet. It is generally believed, in accordance with Jewish tradition, that the Early Hebrew alphabet was superseded in the Holy Land by the Aramaic alphabet during the period of the Babylonian Exile (586–516 BC), and that the Aramaic script therefore became the parent of the Square Hebrew (in Hebrew *ketav meruba'* "square script," or *ketav ashuri* "Assyrian writing"). The theory may be only partly right, for in the Holy Land the Early Hebrew alphabet was an object of such local attachment that for several centuries it was used side by side with the Aramaic script.

At any rate, there is little doubt that the Square Hebrew did derive from the Aramaic alphabet. A distinctive Jewish variety of the Aramaic alphabet that can be regarded as the Square Hebrew script can be traced from the 3rd century BC. It became standardized just before the Christian Era, and it was from this script that the modern Hebrew alphabet, in all its styles, eventually developed. The development was gradual and purely external (*i.e.*, in the shapes of the single letters); from the internal standpoint (*i.e.*, considering the phonetic values of the letters), there has been no development, though it must be borne in mind that for several letters (*e.g.*, waw, ḥet, tzade, qof, shin, sin, etc.) the exact original phonetic value is still uncertain.

Standardization of Square Hebrew

When the Square Hebrew alphabet became standardized, it took (at least, in its formal style and, much later, in its printed form) the form that, with insignificant changes, it now has. Minute rules laid down by the Talmud as to calligraphy and orthography made further development of the Square Hebrew impossible.

In the Square Hebrew alphabet, there are five letters—kaf, mem, nun, pe, and tzade—that have dual forms. That is, there is one character for initial or medial position and another for final position.

The Hebrew alphabet consists of 22 letters, all consonants, though four of them—alef, he, waw, and yod—are also employed to represent long vowels. The absence of vowel letters was not at first a problem, because Hebrew, like other Semitic languages, has consonantal roots, with vowels serving principally to denote inflections in nouns, moods of verbs, and other grammatical variations. As Hebrew speech passed out of daily use (being superseded by Aramaic, which became the vernacular of the Jews) and the knowledge of biblical Hebrew steadily declined, it became necessary to introduce some form of vocalic distinction so that the Bible could be read and explained correctly. The three main vowel systems now extant are the Babylonian, the Palestinian, and the Tiberiadic; of these the latter is the most important, and indeed, the only one still in use. The Tiberiadic system consists of dots, combinations of dots, and little dashes.

Before the discovery of the celebrated Dead Sea Scrolls, several Square Hebrew inscriptions belonging mainly to the 1st century BC and the succeeding centuries were known; they were found in Palestine, Syria, North Africa, and Italy, on rocks, tombs, or ossuaries (depositories for the bones of the dead), in synagogues and catacombs. The biblical manuscripts, except for some fragments written on papyrus, belong to a much later date. The earliest fragment is the Nash papyrus of approximately the 1st century BC, now in the Cambridge University Library. Many thousands of fragments of Hebrew biblical and other manuscripts, partly of the 7th and 8th centuries AD, were discovered in the Geniza, an archive in the old synagogue in Cairo.

The Dead Sea Scrolls

The focus of world interest during the late 1940s and the successive years has been the sensational discovery of Hebrew biblical and nonbiblical scrolls in caves near the Dead Sea. The tens of thousands of fragmentary manuscripts, composing what are popularly called Dead Sea Scrolls, may be divided into several groups, the oldest being a collection of biblical and other Hebrew manuscripts, dating approximately from the 3rd century BC.

In the more than bimillennial development of the Square Hebrew alphabet, four fundamental types can be noticed: (1) the square script, which evolved into the well-proportioned printing type of modern Hebrew (the majority of Dead Sea Scrolls are in this Square Hebrew script); (2) the medieval formal styles; (3) the rabbinic, also known as Rashi writing, which was the medieval book or literary hand; and (4) a cursive script or daily handwriting, which gave rise to many local varieties (Oriental, Spanish, Italian, Franco-German, and so on), of which the Polish-German became the current Hebrew handwriting of today. The Hebrew script has been adapted to some other languages, such as Arabic, Turkish (for the Karaite people of Crimea), and so forth, but particularly to German—hence, Yiddish—and Spanish —hence, Judeo-Spanish.

Arabic alphabet. The Arabic script descended from the Aramaic through the Nabataean and the neo-Sinaitic alphabets. After the Latin script, it is the most widely used form of alphabetic writing in the modern world. The Arab conquests of the 7th and 8th centuries AD, which carried with them the religion of Islām and its holy book—the Qur'ān—written in Arabic, brought the language and the script to the vast expanse of territory extending from India to the Atlantic Ocean. The Arabic alphabet was adapted, with some necessary modifications, to such diverse languages as the Slavic tongues, Spanish, Persian, Urdu, Turkish, Hebrew, Berber, Swahili, Malay, Sudanese, and others.

Kūfic and Naskhi styles

The Arabic alphabet probably originated at some time in the 4th century AD, but the earliest extant Arabic writing is a trilingual inscription—Greek-Syriac-Arabic —of AD 512. The two principal types of Arabic writing, which developed quite early in the Muslim period, were the Kūfic, from the town of Kūfah in Mesopotamia, seat of a famous Muslim academy, and the Naskhi, or Mecca-Medina script. Kūfic, a heavy, bold, and lapidary style, appeared toward the end of the 7th century AD. It was particularly suitable for writing on stone or metal, for painting or carving inscriptions on the walls of mosques, and for lettering on coins. Its letters are generally thick, squat, and unslanted. With the high development of Arabic calligraphy, Kūfic writing became an exceptionally beautiful script. From it there were derived a number of other styles, chiefly medieval, in North and Central Africa, Spain, and northern Arabia.

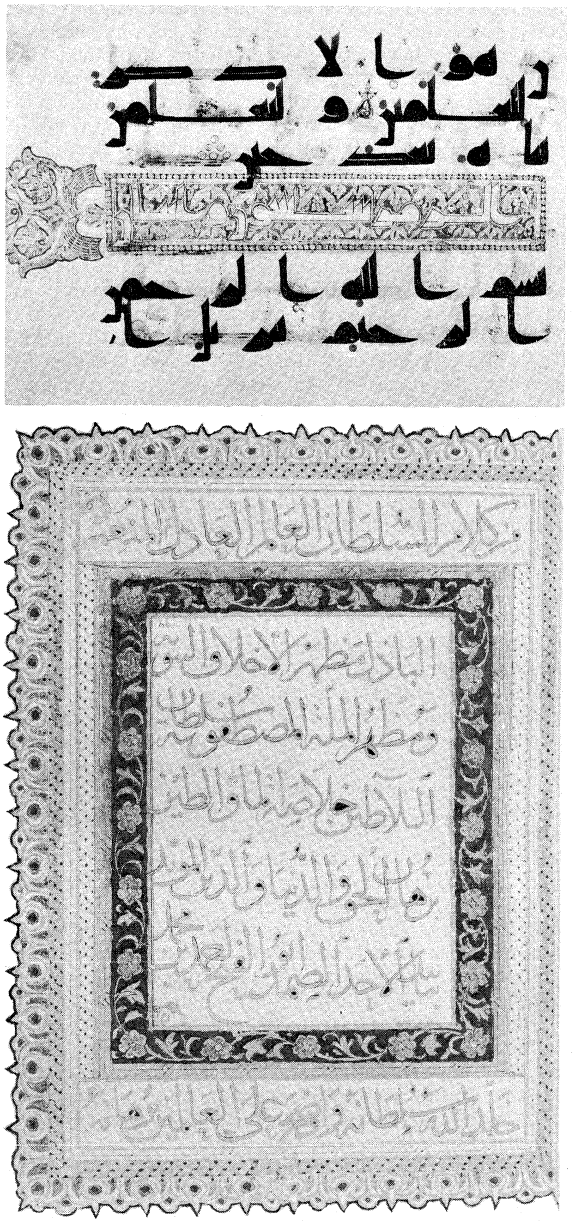


Figure 2: *Early Arabic Writing.*
(Top) Early Kūfic book style, leaf from a Qur'ān, 8th–9th century. In the Freer Gallery of Art, Smithsonian Institution, Washington, D.C. (Bottom) Old Ottoman *naskhī*, opening of the Qur'ān, 1394. In the British Museum (Ms. OR 4126).
By courtesy of the (top) Freer Gallery of Art, Smithsonian Institution, Washington, D.C., (bottom) trustees of the British Museum

Thereafter, it was virtually discontinued except for formal and monumental writing. Nevertheless, it was also used for writing precious manuscripts of the Qur'ān, many of which are extant today.

The Naskhi style was from the very outset a more cursive form. It was always employed chiefly for writing on papyrus. In time, it evolved into innumerable styles and varieties, including the *taʿliq*, the *riqaʿ*, the *divani*, the *thuluth*, and the *syakat*, and became the parent of the modern Arabic writing.

Like other Semitic scripts, Arabic is written from right to left. Its alphabet contains 28 consonantal letters, 22 being directly derived from the Aramaic-Nabataean branch of the North Semitic alphabet, and six being new additions; three of the letters—alif, wāw, and yā'—are also used as long vowels.

The written letters undergo a slight external change according to their position within a word. When they stand alone or occur at the end of a word, they ordinarily terminate in a bold stroke; when they appear in the middle of a word, they are ordinarily joined to the letter following by a small, upward curved stroke. With the exception of six letters, which can be joined only to the preceding ones, the initial and medial letters are much abbreviated, while the final form consists of the initial form with a triumphant flourish. The essential part of the characters, however, remains unchanged. On the whole, the evolution of the forms of the Arabic letters was the most rapid of all the branches of alphabetic writing.

Vowel marks in Arabic

Although the absence of vowel letters was not strongly felt in Arabic (as in Hebrew and other Semitic languages), for teaching purposes and for correct reading of the Qur'ān, the use of diacritical marks (including signs for short vowels, which are sometimes used in conjunction with the letters alif, wāw, and yā') was introduced in Basra in the early 8th century. The practice was probably borrowed from the Syriac script. It not only provides vowel sounds but also distinguishes different consonants; diacritical points are also used as endings in the inflection of nouns and the moods of verbs. These marks—there are three of them—are written above or below the consonants (preceding the vowel), while a sign called *sukūn* indicates the absence of a vowel. Thus, there are, on the whole, a great number of diacritical points; these form a peculiar characteristic of this writing form.

The Indian branch. The Aramaic alphabet was probably the prototype of the Brāhmī script of India, the ancestor of all Indian scripts. The transmission probably took place in the 7th century BC. Adapting the Aramaic script to the Indo-Aryan tongue of India was by no means simple or straightforward. The shapes of many Brāhmī letters show clear Semitic influence; moreover, the Brāhmī script was originally written from right to left. It is obvious, however, that on the whole it was the idea of alphabetic writing that was transmitted and that the fully developed Brāhmī writing was the outcome of the brilliant philological and phonological elaboration of the scientific Indian school.

During the 5th century BC, the second of the prototypal Indian alphabets—the Kharoṣṭī script—came into being in northwest India (which was then under Persian rule). Although the origin of Brāhmī is still uncertain and hotly discussed, it is commonly accepted that the Kharoṣṭī alphabet is a direct descendant from the Aramaic alphabet. Moreover, the direction of writing in Kharoṣṭī script is as in Aramaic, from right to left, and there is also a likeness of many signs having similar phonetic value.

In the later centuries of its existence, Brāhmī gave birth to eight varieties of scripts. Three of them—the early and late Maurya and the Śuṅga—became the prototypes of the North Indian subdivision of the Brāhmī script in the 1st centuries BC and AD. Out of this North Indian subdivision there arose the Gupta, a monumental form of writing, which was employed from the 4th to the 6th century AD and became the ancestor of the great majority of Indian scripts.

Gupta writing in India

The western variety of the Gupta spread into eastern (or Chinese) Turkistan, where it was adopted for a number of languages, including the recently discovered Turfanian and Kuchean (Tocharian A and B), and where it strongly influenced the invention or revision of the Tibetan script (AD 639). There were two main offshoots of the Tibetan writing: the 'Phags-pa, adapted to the Chinese and Mongolian languages in 1272; and the Lepcha, which arose in the beginning of the 18th century.

Much more important was the Siddhamatrka script, developed during the 6th century AD from the western branch of the eastern Gupta character. The Siddhamatrka became the ancestor of the Nāgarī or Devanāgarī script (Sanskrit *deva* "divine," *nāgarī* "script of the city"), which is the script used for Sanskrit. It is, therefore, the most important Indian script. Consisting of 48 signs (14 vowels and diphthongs and 34 basic consonants), it is the common means of communication among learned men throughout India. The Devanāgarī developed in the 7th to 9th centuries and has remained, since the 9th century, essentially unaltered even today.

From the Devanāgarī writing as used in eastern India

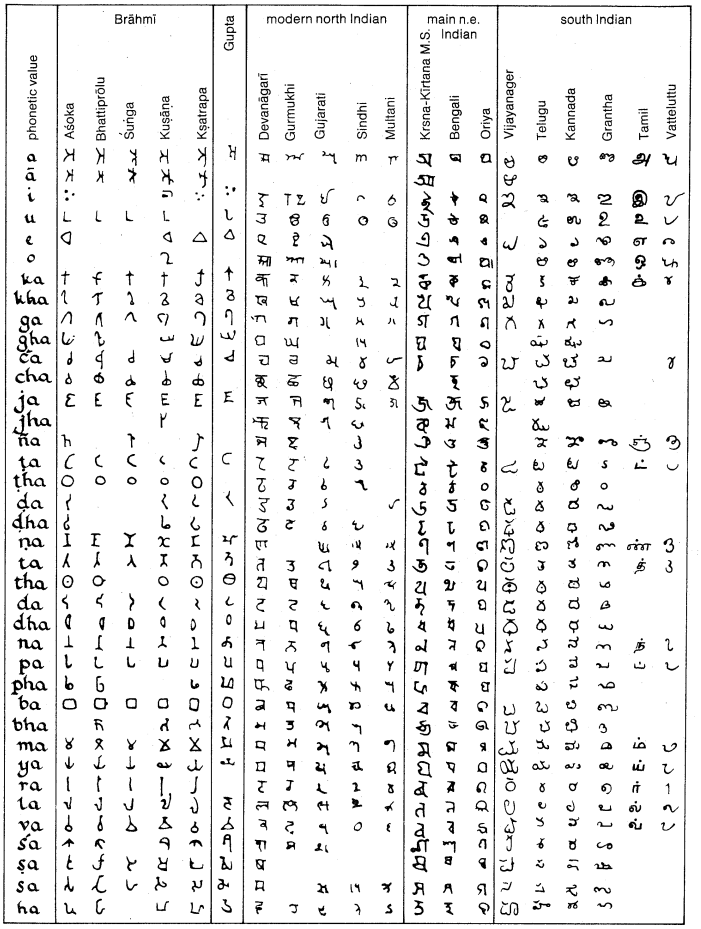


Figure 3: Comparison of several Indian scripts.
From D. Diringer, *The Alphabet: A Key to the History of Mankind*

in the 11th century, there developed the proto-Bengali and the early Nepali, or Newari, scripts, from which the many scripts employed nowadays in northern India and Bangladesh descended; *e.g.*, the Bengali, Oriya, Manipuri, Assamese, Gujarati, and Bihari scripts, and the various Eastern Hindi local scripts.

In northwestern India several other scripts are employed. The Sarada script, a descendant of the western type of the Gupta character, originated in the 8th century and is still employed for Kashmiri. In addition, there are the several varieties of the Takri, used over the lower ranges of the Himalayas north of Punjab; the Dogri, used for a dialect of Punjabi; the Landa, the national alphabet of Punjabi, which has many varieties and is used mainly by shopkeepers all over Punjab and Sindh; and the Gurmukhi script, the characters of the Sikh scriptures.

In South India, which is inhabited by peoples speaking Dravidian languages, several other scripts are used, of which the Kannada, or Kanarese, the Telugu, the Grantha, the Tulu-Malayalam, the Tamil, and the Vaṭṭeḷuttu are the most important.

Long before the existence of the Gupta script, the Brāhmī script had already begun its eastward movement. The Indo-Aryan migration to the island of Ceylon in the 5th century BC had set the stage there, and the earliest Brāhmī inscriptions in Ceylon can be dated in the 3rd century BC. Most dramatic of all, however, was the expansion of Buddhism from India into Ceylon, Burma, Thailand, Cambodia, Laos, Vietnam, Malaya, and Indonesia. As already mentioned, unlike the conquests of Islām, this was a peaceful movement; its "soldiers" were Buddhist monks, political independents who built an empire founded on the cultural and spiritual community of peoples. Among their many achievements, these monks brought into being offshoots from the Brāhmī script, principally from its South Indian varieties, throughout the vast extent of territory from India itself to the Philippines. Thus arose the many scripts of Southeast Asia, from the Cham writing of Cambodia to the Kavi character of Java and its Sumatran offshoots, and the Tagalog writing of the Philippines.

Spread of Brāhmī to Southeast Asia

All these Indian and Southeast Asian scripts involve types of semisyllabaries rather than alphabets. They consist of vowels and diphthongs and basic consonants (*i.e.*, consonants followed by a short *a*); there are no pure consonants (*i.e.*, consonants written by themselves).

Greek alphabet. The Greek alphabet derived from the North Semitic script between *c.* 1000 BC and the 9th century BC. The direction of writing in the oldest Greek inscriptions—as in the Semitic scripts—is from right to left, a style that was superseded by the boustrophedon (meaning, in Greek, "as the ox draws the plow"), in which lines run alternately from right to left and left to right. This change occurred approximately in the 6th century BC. There are, however, some early Greek inscriptions written from left to right, and after 500 BC, Greek writing invariably proceeded from left to right.

The letters for *b*, *g*, *d*, *z*, *k*, *l*, *m*, *n*, *p*, *r*, and *t*, which are sounds common to the Semitic and Greek languages, were taken over without change. The principal Greek changes were the introduction of vowel representation, the re-arrangement of the sibilant or fricative sounds (of which the Semitic alphabets had a considerable variety), the adoption of some Semitic letters for slightly different Greek sounds, and the invention of symbols for Greek sounds not expressed by any of the Semitic letters. The different ways in which these adaptations were carried out allow the two main branches of the early Greek alphabet—the eastern and the western—to be distinguished. These again subdivide, each into secondary branches. Within this general grouping there were many local peculiarities, but the differences between all of these local alphabets involved variations in detail rather than essential structure.

Ionic and other eastern Greek alphabets

The eastern and western subdivisions were the two principal branches of the early Greek alphabet. The Ionic alphabet was the most important of the eastern variety, which also included the Greek alphabets of Asia Minor and the adjacent islands, of the Cyclades and Attica, of Sicyon and Argos, of Megara, Corinth, and the Ionian colonies of Magna Graecia. A secondary branch of the eastern subdivision was made up of the alphabets used on the Dorian islands of Thera, Melos, and Crete. The alphabets of Euboea (Chalcidian), Boeotia, Phocis, Locri, Thessaly, the Peloponnesus (except its northeastern part), and of the non-Ionian colonies of Magna Graecia belonged to the western subdivision. It is a controversial point whether the eastern or the western branch was the earlier in time, whether there was any derivative link between one and the other, or whether they represent two quite independent adaptations of the Semitic alphabet (the latter alternative seems rather improbable).

Gradually, the Greek local alphabets became more and more similar. In 403 BC, the Ionic alphabet of Miletus was officially adopted in Athens and later also in the other states. By the middle of the 4th century BC, almost all the local alphabets had been replaced by the Ionic, which became the common, classical Greek alphabet of 24 letters.

After this time, the development of the Greek alphabet was almost wholly external, in the direction of greater utility, convenience, and, above all, beauty. The classical style was retained as a monumental script at the same time that more cursive forms grew up for writing on such surfaces as parchment, papyrus, and wax. The classical letters were also retained as the capital letters in the modern print (though some of the capitals in modern Greek handwriting are borrowed from the Latin alphabet). On the other hand, the classical Greek alphabet also evolved into the Greek uncials, the cursive, and the minuscule script. (Uncial letters were somewhat rounded and separated versions of capital letters or cursive forms; minuscule letters developed from cursive writing and have simplified, small forms.) Up until about 800 AD, the uncials were used as a book hand; later, the minuscule

script was employed for the same purpose. The cursive scripts evolved into the modern Greek minuscule.

Accent marks in Greek

In the middle of the 3rd century BC, the Greek scholar Aristophanes of Byzantium introduced the three accents —acute, grave, and circumflex—that were thereafter used to assist students, particularly foreigners, in the correct pronunciation of Greek words; these continue to be used in most Greek texts printed today. Originally, these marks indicated tone or pitch, not stress.

Countless inscriptions have been discovered all over the Hellenic and Hellenistic world and beyond. They include official decrees, annals, codes of law, lists of citizens, civic rolls, temple accounts, votive offerings, ostraca (fragments of pottery), sepulchral inscriptions, coins, lettering on vases, and so forth. These, along with many thousands of Greek manuscripts, both ancient and medieval, serve as sources for the studies known as Greek epigraphy and Greek paleography and are of untold importance for all branches of ancient history, philology, philosophy, and other disciplines.

The most direct offshoots from the Greek alphabet were those adapted to the languages of the non-Hellenic peoples of western Asia Minor in the 1st millenium BC: the scripts of the Lycians, Phrygians, Pamphylians, Lydians, and Carians. The first three of these were derived directly from the Greek; the Lydian and Carian were strongly influenced by it. The Coptic alphabet was the other non-European offshoot from the Greek and the only one used in Africa. Twenty-four of its 31 letters were borrowed from the Greek uncial writing, and seven were taken over from a particularly cursive variety of the Egyptian demotic writing; the demotic letters were used to express Coptic sounds not existing in the Greek language.

More significant, however, were the European offshoots. In Italy, two alphabets derived directly from the Greek: the Etruscan and the Messapian. The Messapii were an ancient tribe who inhabited the present Apulia (in southern Italy) in pre-Roman times; their language is presumed to belong to the Illyrian group. Over 200 Messapic (or Messapian) inscriptions have been discovered. In southeastern Europe, there were three offshoots from the Greek alphabet—the Gothic, the Cyrillic, and the Glagolitic alphabets. The Gothic alphabet, not to be confused with the so-called Gothic script (a variety of the Latin alphabet), was a script created by the Gothic bishop Ulfilas (or Wulfila), who died in AD 381 or 383. The script consisted of 27 letters, of which some 19 or 20 were taken over from the Greek uncial script. Ulfilas translated the Bible into Gothic; of this translation, some fragments are extant in manuscripts of the 5th and 6th centuries. The most important manuscript is the *Codex Argenteus*, preserved in Uppsala, Sweden.

The Gothic alphabet

Cyrillic and Glagolitic alphabets. The two early Slavic alphabets, the Cyrillic and the Glagolitic, were invented by St. Cyril, or Constantine (*c.* 826–869), and St. Methodius (*c.* 815–885). These men were Greeks from Thessalonica, who became apostles to the southern Slavs, whom they converted to Christianity. An early tradition, in attributing the invention of an early Slavic writing to Cyril, does not indicate whether his contribution was the Cyrillic or the Glagolitic. It is just possible that both alphabets were invented by him. The earliest dated Old Slavic documents belong to the late 10th and the 11th centuries. The Cyrillic and the Glagolitic alphabets differed widely in the form of their letters, in the history of their development, and partly also in the number of the letters, but they were alike in representing adequately the many sounds of Slavic.

The Cyrillic alphabet was based on the Greek uncial writing of the 9th century. It originally had a total of 43 letters; the two Hebrew letters tzade and shin were transformed into the Cyrillic letters for the sounds *ch*, *sh*, and *shch*. The modern forms of this alphabet have fewer letters. Glagolitic writing consisted of 40 letters, externally very unlike either the Greek or Cyrillic scripts.

Cyrillic became, with slight modification in each case, the national script of the Bulgarians, Serbs, Russians, Belorussians, and Ukrainians. (The other Slavic peoples —the Slovenes, Croats, Czechs, Slovaks, Wends, Lusatians, and Poles—use the Latin alphabet.) As mentioned above, in Yugoslavia a single language is written in Cyrillic by the Greek Orthodox Serbs and in the Latin alphabet by the Roman Catholic Croats. For a time, Cyrillic was also adapted to the Romanian language, and in recent times, through the medium of Russian script, it became the writing of a number of Finno-Ugric languages (Komi, Votyak, Mordvinian, Vogul, Ostyak, etc.), Turco-Tatar languages (Chuvash, Turkmenian, Azerbaijanian, etc.), Iranian languages (Ossetic, Kurdish, Tajiki), and Caucasian languages (Abkhaz, Circassian, Avar, etc.).

The history of the Glagolitic alphabet is particularly connected with the religious history of the Slavic peoples of southwest central Europe and the western Balkan Peninsula. In the second half of the 9th century, it was introduced, together with the Slavonic liturgy, into the Moravian kingdom, but with the banning of this liturgy by the pope, it disappeared from Moravia. It was, however, accepted (also with the Slavonic liturgy) in Bulgaria and Croatia and spread along the Dalmatian coast southward into Montenegro and westward into Istria. Although the Glagolitic script soon disappeared among the Greek Orthodox Slavic peoples because of the victory of the Cyrillic, it continued, notwithstanding the opposition of the higher Catholic authorities, to be employed among the Catholics of the western Balkan Peninsula together with the Slavonic liturgy and finally succeeded in obtaining the special license of the pope. It is still employed in the Slavonic liturgy in some Dalmatian and Montenegrin communities; the inhabitants of these places are the only Roman Catholics to use the Slavonic liturgy. The earliest preserved Glagolitic secular document dates from 1309. Glagolitic had a short flourishing period in the 16th and 17th centuries.

Etruscan alphabet. The Etruscans, a highly civilized people who were the ancestors of the modern Tuscans and the predecessors of the Romans, inhabited what is now modern Tuscany in central Italy; their language, still mainly undeciphered, has come down in over 11,000 inscriptions, the earliest being the 8th-century-BC Marsiliana Tablet, preserved in the Museo Archeologico in Florence. This is also the earliest preserved record of a Western alphabet. The early Etruscan alphabet, unlike any early Greek alphabet found in the Greek inscriptions, contains the original—the prototype—Greek alphabet, consisting of the 22 North Semitic letters, with the phonetic values given to them by the Greeks, and the four additional Greek letters at the end of the alphabet. The Etruscans introduced various changes in their script, and several features in the modern alphabets can be attributed to the influence of the ancient Etruscans. An example is the phonetic value of "k" for the letters *c*, *k*, and *q*. Like the Semitic and the early Greek alphabets, Etruscan writing nearly always reads from right to left, though a few inscriptions are in boustrophedon style. The probable date of the origin of the Etruscan alphabet is the late 9th or early 8th century BC.

About 400 BC, the "classical" Etruscan alphabet took its final form of 20 letters—four vowels and 16 consonants. Because the voiced and voiceless sounds *b* and *p*, *d* and *t*, *g* and *k* were not differentiated in the Etruscan language, letters *b* and *d* never appear in pure Etruscan inscriptions, and after the disappearance of *k* and *q*, the letter *C* was employed for *g* and *k*.

The Etruscan alphabet had many varieties and several offshoots. Among the offshoots, apart from the Latin, were many alphabets used by Italic populations of pre-Roman Italy and by non-Italic tribes (*e.g.*, the Piceni).

Latin alphabet. The adaptation of the Etruscan alphabet to the Latin language probably took place at some time in the 7th century BC. From this century, there is a gold brooch known as the Praeneste Fibula (preserved in the Museo Preistorico Etnografico Luigi Pigorini in Rome). The inscription, written in an early form of Latin, runs from right to left and reads clearly: *manios: med:fhefhaked:numasioi*, which in classical Latin is *Manius me fecit Numerio* "Manius made me for Numerius."

Oldest known record of Latin alphabet

Dating from the end of the 7th or the beginning of the 6th century BC is a famous cippus (small pillar) from the Roman Forum; it is inscribed vertically on its four faces, in boustrophedon style. Another inscription, probably of the 6th century BC, is known as that of the Duenos Vase and was found in Rome, near the Quirinal Hill. It is also written from right to left. Some Sabine inscriptions belong to the 5th or the 4th century BC. There are also a few inscriptions belonging to the 3rd and 2nd centuries BC.

The Roman capital letters, a form of writing that was used under the empire with unparalleled effectiveness for monumental purposes, became a byword for precision and grandeur, despite a very unprepossessing beginning. Indeed, for the first six centuries of its existence, Roman writing was relatively unimpressive. Only with the advent of the 1st century BC were there signs of magnificence to come.

An opinion that used to be commonly held, and still is held by many, is that the Latin alphabet was derived directly from the Greek in a form used by Greek colonists in Italy. The theory rested on an assertion that the Latin alphabet corresponds to the Chalcidian variety of the western group of Greek scripts employed at Cumae in Campania, southern Italy. This theory is unlikely; indeed, as already mentioned, the Etruscan alphabet was the link between the Greek and the Latin. For instance, the most interesting feature in the inscription of the Praeneste Fibula is the device of combining the letters *f* and *h* to represent the Latin sound of *f*. This was one of the Etruscan ways of representing the same sound. Also, most of the Latin letter names, such as a, be, ce, de for the Greek alpha, beta, gamma, delta, and so on, were taken over from the Etruscans.

Etruscan as link between Greek and Latin

Runic and ogham alphabets. Runes, in all their varieties, may be regarded as the "national" script of the ancient North Germanic tribes. The origin of the name rune (or "runic") is probably related to the fact that the ancient Germanic tribes, like all primitive peoples, attributed magic powers to the mysterious symbols scratched on armour, jewels, tombstones, and so forth. This is given credence by two related Germanic forms that mean "mystery, secret, secrecy": the Old Germanic root *ru-* and the Gothic *runa*. The most interesting runic inscriptions are those that were cut for magical purposes and those that appeal to heathen deities.

The origin of the runes offers many difficult problems, and has been hotly argued by scholars and others. The theory of the *Urrunen* (forerunners of the runes), a supposed prehistoric north Germanic alphabetic script, holds that it is the parent not only of the runes but also of all the Mediterranean alphabets, including the Phoenician. This belief, based on racial and political grounds, need not be seriously considered. Some scholars propounded the 6th century BC Greek alphabet as the prototype of the runes; others have suggested the Greek cursive alphabet of the last centuries BC. Several eminent scholars have proposed the Latin alphabet as the source of the runes. The most probable theory, supported recently by many scholars, is that the runic script derived from a North Etruscan, Alpine alphabet. In that case, it is very probable that it originated about the 2nd century BC or a little later.

Origin of the runes

It is still unknown whether the runes were originally employed mainly for magical purposes, as suggested by the name *runa*, or as a usual means of communication. The earliest extant runic inscriptions, numbering over 50, come from Denmark and Schleswig and date from the 3rd to the 6th century AD. About 60 inscriptions from Norway date from the 5th to the 8th century, slightly later than the continental ones. There are also about 50 Anglo-Saxon runic inscriptions extant, including the Franks Casket (about AD 650–700); the right side of the casket is in the Museo Nazionale del Bargello, in Florence, Italy, and the rest is in the British Museum. The largest number of inscriptions, about 2,500, come from Sweden; most of these date from the 11th and 12th centuries AD.

There is no certain evidence of wide literary use of runes in early times, but some scholars hold that the runic writing was widely employed for all kinds of secular documents, such as legal provisions, contracts, genealogies, and poems. The known manuscripts are, however, rare and relatively late. The gradual displacement of the runes coincided with the increasing influence of the Church of Rome. The runic scripts lingered on for a long time after the introduction of Christianity, however; indeed, the use of runes for charms and memorial inscriptions lasted into the 16th or even the 17th century.

From D. Diringer, *The Alphabet: A Key to the History of Mankind*

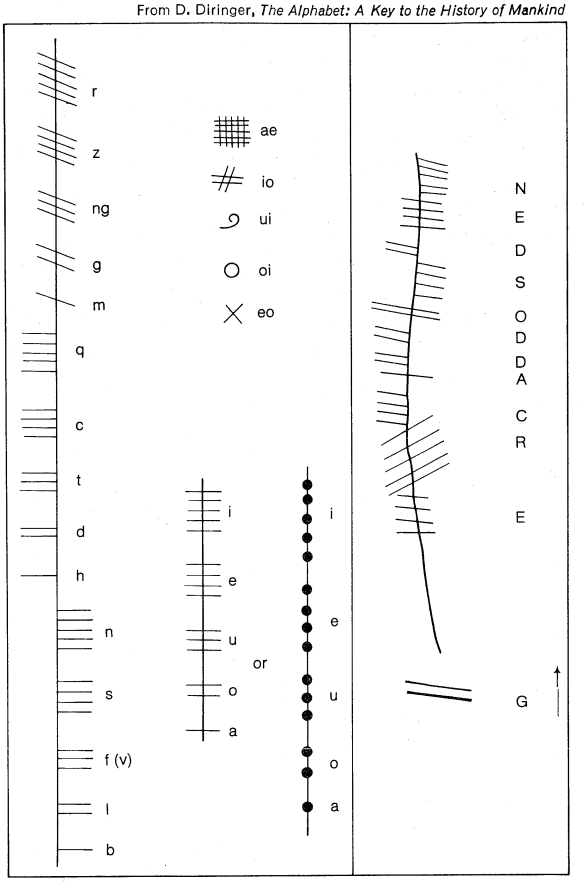


Figure 4: (Left) Ogham alphabet. (Right) Pictish ogham.

The ogham alphabet was restricted to the Celtic population of the British Isles. There are over 375 known inscriptions: 316 of them have been discovered in Ireland, chiefly in the southern counties, with only 55 from the northern counties; 40 inscriptions have been discovered in Wales; two come from Devon; and one from Cornwall. One inscription was discovered at Silchester in southern England. About ten come from the Isle of Man, and a few from Scotland. The Welsh inscriptions are usually bilingual, Latin-Celtic. With one exception, the Irish records are in ogham alone. Most peculiar is the runic-oghamic inscription from the Isle of Man (the runes being a kind of "secret" writing and the oghams being a cryptic script). The distribution of the ogham inscriptions, combined with their language and grammatical forms, point to South Wales or south Ireland as their place of origin, and to the 4th century AD as the date of their origin.

The ogham character was used for writing messages and letters (generally on wooden staves), but sometimes it was also written on shields or other hard material and was employed for carving on tombstones. The oghams formed a cryptic script, and there were several varieties, such as wheel oghams, bird oghams, tree oghams, hill oghams, church oghams, colour oghams, and others. The main ogham alphabet consisted of 20 letters represented by straight or diagonal strokes, varying in number from one to five and drawn or cut below, above, or right through horizontal lines, or else drawn or cut to the left,

right, or directly through vertical lines. The ogham alphabet was divided into four groups (*aicme*), each containing five letters. Oghams were employed during the Middle Ages; the 14th-century *Book of Ballymote* reproduces the earliest keys for translation. In many cases, the ogham inscriptions run upward.

Pictish oghams

Several ogham inscriptions known as the Pictish oghams were found in western Scotland, on the small island of Gigha off the western coast, in Argyll in northeastern Scotland, and on the northern isles, such as the Shetland Islands. They either belong to the same type as the Irish and Welsh oghams, or are written in another ogham variety.

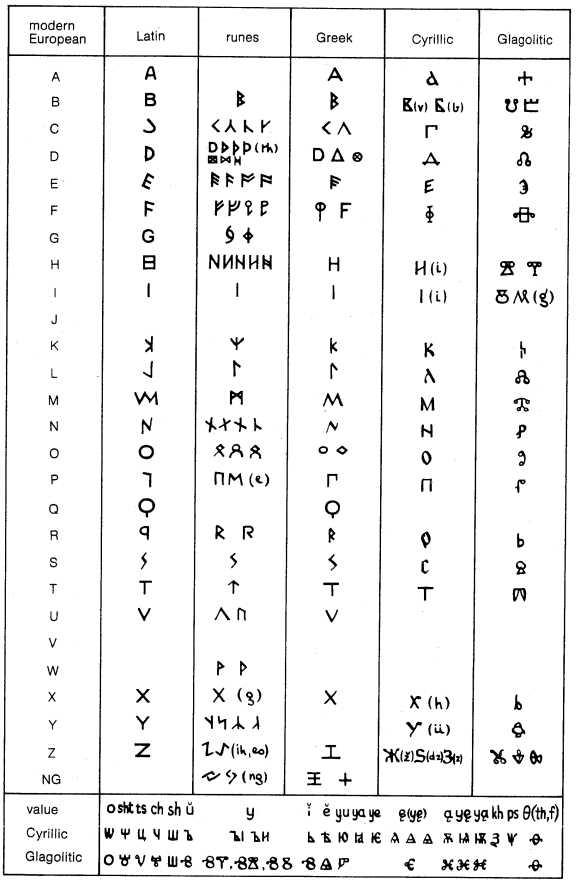


Figure 5: Comparison of Latin, runic, Greek, Cyrillic, and Glagolitic alphabets.

Later development of the Latin alphabet. As already mentioned, the original Etruscan alphabet consisted of 26 letters, of which the Romans adopted only 21. They did not retain the three Greek aspirate letters, theta, phi, and chi, in the alphabet because there were no corresponding Latin sounds but did employ them to represent the numbers 100, 1,000, and 50. Of the three Etruscan *s* sounds, the Romans kept what had been the Greek sigma. The symbol that represented the aspirate later received the shape *H* as it did in Etruscan. *I* was the sign both of the vowel *i* and the consonant *j*. *X* was added later to represent the sound *x* and was placed at the end of the alphabet. At a later stage, after 250 BC, the seventh letter, the Greek zeta, was dropped because Latin did not require it, and a new letter, *G*, made by adding a bar to the lower end of *C*, was placed in its position.

After the conquest of Greece in the 1st century BC, a large number of Greek words were borrowed by the Latin language. At that time the symbols *Y* and *Z* were adopted from the contemporary Greek alphabet, but only to transliterate Greek words; hence, they do not appear in normal Latin inscriptions. They were placed at the end of the alphabet, and the Latin script thus became one of 23 symbols.

A few permanent additions, or, rather, differentiations from existing letters, occurred during the Middle Ages, when the signs for *u* and *v*, and *i* and *j*, previously written interchangeably for either the vowel or the consonant sound, became conventionalized as *u* and *i* for vowels and *v* and *j* for consonants. *W* was introduced by Norman scribes to represent the English sound *w* (a semivowel) and to differentiate it from the *v* sound.

The connection of the capital letters of modern writing with the ancient Semitic-Greek-Etruscan-Latin letters is evident even to a layman. The connection of the minuscules (*i.e.*, the small letters) with the ancient Latin letters is not as evident, but in fact both the majuscules and the minuscules descended from the same ancient Latin alphabet. The different shapes of the small letters are the result of a transformation of the ancient letters by the elimination of a part of the letter—as, for instance, *h* from *H* or *b* from *B*—or by lengthening a part of it—for instance, *d* from *D*. Moreover, the change of the Latin writing into the modern script was caused by technical bearings of the tool, primarily the pen, and the material of writing, mainly papyrus and parchment, and, from the 14th century onward, also paper. It was the pen, with its preference for curves, that eliminated the angular forms; it was the papyrus, and still more the parchment or vellum, and, in modern times, paper, that made these curves possible.

Derivation of majuscules and minuscules from Latin alphabet

In ancient times, the minuscule did not exist, but there were several varieties of the capital and the cursive scripts. There were three varieties of the capitals: the lapidary capitals (used mainly on stone monuments); the elegant book capitals, somewhat rounded in shape; and the rustic capitals, which were less carefully elaborated than the lapidary script and not as round as the book capitals, but more easily and quickly written. In everyday life, the cursive script—*i.e.*, the current hand—was developed with continuous modifications for greater speed. There were several varieties of it, such as those of Pompeii and Alburnus Major (a town in ancient Dacia, modern Roşia Montană, Romania). Between the monumental and the cursive scripts, there was a whole series of types that had some of the peculiarities of each group. There were lapidary mixed scripts and book semicursive scripts, and there was the early uncial or, rather semiuncial script of the 3rd century AD, which seems to have developed into the beautiful uncial script.

When the various European countries had shaken off the political authority of Rome and the learned communities had been dissolved and their members scattered, a marked change took place in the development of the Latin literary, or book hand. Several national hands, styles of the Latin cursive, assumed different features. There thus developed on the European continent and in the British Isles the five basic national hands, each giving rise to several varieties: Italian, Merovingian in France, Visigothic in Spain, Germanic, and Insular or Anglo-Irish hands. At the end of the 8th century, the Carolingian (Caroline) hand developed and, after becoming the official script and literary hand of the Frankish Empire, developed as the main book hand of western Europe in the following two centuries. The combination of the majuscules, or capital letters, and minuscules, or small letters, can be attributed mainly to the Carolingian script.

National cursive hands in Europe

In the course of the next centuries, various book hands or chart hands and other cursive scripts developed from the Carolingian style. In the late 12th century, and during the next two centuries, the letters gradually became angular in shape; this resulted from the pen being held in a position that made a slanting stroke. The new hand, termed black letter, or Gothic, was employed mainly in northwestern Europe, including England, until the 16th century. It is still used, though rarely, in Germany, where it is called Fraktur script.

In Italy, the black letter was also used, but the Italians preferred a rounder type, called *littera antiqua* "old letter." During the 15th century, the round, neat, humanistic or Renaissance hand was introduced in Florence and was employed for literary productions, while the needs of everyday life were met by an equally beautiful, though not as clearly legible, cursive hand. The two styles devel-

oped into two main varieties: (1) the Venetian minuscule, nowadays known as italics, traditionally (though wrongly) considered to be an imitation of Petrarch's handwriting; and (2) the Roman type, preferred in northern Italy, chiefly in Venice, where it was used in the printing presses at about the end of the 15th and the beginning of the 16th centuries; from Italy it spread to Holland, England (about 1518), Germany, France, and Spain. The classical Roman character was adopted for the majuscules. This majuscule writing, along with the Roman type minuscule and the italics, spread all over the world. In England, they were adopted from Italy in the 16th century.

The survival of the black letter (Gothic) in Germany is attributed to the fact that it was the current style at the time of the invention of printing in Germany—it was employed by Gutenberg; in Italy the *littera antiqua* was used by the German printers Konrad Sweynheym and Arnold Pannartz, as well as by Nicolas Jenson, the great Venetian printer who perfected the Roman type.

The modern national alphabets of the western European nations are, strictly speaking, adaptations of the Latin alphabet to Germanic (English, German, Swedish, Dutch, Danish, etc.), Romance (Italian, French, Spanish, Portuguese, etc.), Slavic (Polish, Czech, Slovak, etc.), Baltic (Lithuanian, Latvian), Finno-Ugric (Finnish, Hungarian, etc.), and other languages. The adaptation of a script to a language is not easy, especially when the language contains sounds that do not occur in the speech from which the script has been borrowed. There arises, therefore, the difficulty of representing the new sounds. This difficulty was met quite differently in various alphabets. For instance, the sound *shch* as in English "Ashchurch," which in Russian is represented by one sign (щ), is represented in Czech by two signs (*šč*), in Polish by four (*szcz*), in English likewise by four, though different ones, and in German by as many as seven (*schtsch*). Thus, in these instances, combinations of two or more letters were introduced to represent the new sounds.

Invention of new signs and diacritical marks

In other cases, new signs were invented; *e.g.*, in the early Greek alphabet, and in the Anglo-Saxon adoption of the Latin alphabet. In more recent times, the most common way of representing sounds that cannot be represented by letters of the borrowed alphabet has been to add diacritical marks, either above or under the letters, to their right or left or inside. To this group belong the German vowels *ü, ä, ö;* the Portuguese and French cedilla in *ç;* the tilde on Spanish and Portuguese *ñ*; the Italian *à, é, è, ì, ù*, etc.; the great number of marks in the Latin-Slavic alphabets (Polish, Czech, Croatian, etc.)—*ą, ę, č, ć, š, ś, ž, ż, ź*, and so on. The Latin-Turkish alphabet, introduced in 1928, became general throughout Turkey in 1930. It contains 29 letters, of which two vowels (*ö* and *ü*) and three consonants (*ç, ĝ*, and *ş*) are distinguished by diacritical marks; in one instance there is a distinction in reverse—the dot from *i* is eliminated (*ı*) to represent a new sound.

Attempts to make an ideal alphabet. In a perfect alphabet each sound would be represented by a single symbol, and no more than one sound would be represented by any symbol. But there are no completely perfect alphabets. All alphabets omit symbols for some sounds, and all contain redundant letters. Indeed, living speech hardly conforms to the written word. There are, however, languages, such as Italian, Spanish, Portuguese, German, and even Bantu, that are relatively accurately represented in graphic form. In English, the spelling in many words is almost an arbitrary symbolism, which gives rise to interest in phonetic spelling and spelling reform. The I.T.A., Initial Teaching Alphabet, invented by Sir James Pitman, is an attempt to regularize English orthography.

The International Phonetic Alphabet

The International Phonetic Alphabet (IPA) is a system of writing designed to overcome the inconsistencies and redundancies in regular alphabets. Developed at the end of the 19th century by members of the International Phonetic Association, this alphabet has been used by linguists and others who must record with a greater degree of phonetic accuracy than can be achieved with ordinary alphabets. There are many additional types of phonetic alphabets. (See also PHONETICS.)

With all of its deficiencies, the alphabet has been the only, or at least the best, means of international communication until the present day. Its privileged position can be attributed to a natural development lasting many centuries and accompanied by many other elements. The external development of the individual letters, strongly influenced by calligraphers and great graphic experts, is mainly the result of aesthetic considerations.

BIBLIOGRAPHY. The following is a highly selective bibliography of works on alphabets.

Eighteenth and nineteenth centuries: T. ASTLE, *The Origin and Progress of Writing* (1784; 2nd ed., 1803; another ed., 1876); K. FAULMANN, *Illustrirte Geschichte der Schrift* (1880); J.B. SILVESTRE and F. MADDEN, *Paléographie universelle*, 4 pt. (1841; Eng. trans., 2 vol., 1850); I. TAYLOR, *The Alphabet: An Account of the Origin and Development of Letters*, 2 vol. (1883; 2nd ed., 1899), an outstanding work; H. WUTTKE, *Geschichte der Schrift und des Schrifttums* (1872).

Twentieth century: Almost all of these works contain excellent, copious, and classified bibliographies. A.B. ALLEN, *The Romance of the Alphabet* (1937); M. COHEN, *La Grande invention de l'écriture et son évolution*, 2 vol. (1958), also outstanding; D. DIRINGER, *The Alphabet: A Key to the History of Mankind* (1948; 3rd rev. ed. in 2 vol., 1968); *Writing* (1962), also translations into Russian, Danish, Norwegian, and other languages; E. DOBLHOFER, *Zeichen und Wunder* (1957; Eng. trans., *Voices in Stone: The Decipherment of Ancient Scripts and Writings*, 1961); W.C. DURFEE, *Alphabetics As a Science* (1956); J.G. FEVRIER, *Histoire de l'écriture*, 2nd ed. (1959); J. FRIEDRICH, *Geschichte der Schrift* (1966); I.J. GELB, *A Study of Writing*, rev. ed. (1963); C. HIGOUNET, *L'Écriture* (1955); K.G. IRWIN, *The Romance of Writing, from Egyptian Hieroglyphics to Modern Letters, Numbers, and Signs* (1956); V.A. ISTRIN, *L'Écriture, sa classification, sa terminologie* (1957); H. JENSEN, *Die Schrift in Vergangenheit und Gegenwart* (1969; Eng. trans., *Sign, Symbol and Script*, 3rd rev. ed., 1970); W.H. LANGE, *Schriftbibel*, 3rd ed. (1951); S.A.B. MERCER, *The Origin of Writing and Our Alphabet* (1959); A.C. MOORHOUSE, *The Triumph of the Alphabet: A History of Writing* (1953); O. OGG, *The 26 Letters*, rev. ed. (1971); A. PETRAU, *Schrift und Schriften im Leben der Völker* (1939); P. RUDLAND, *From Scribble to Script* (1955); A. SCHMITT, *Untersuchungen zur Geschichte der Schrift*, 2 vol. (1940); K. SETHE, *Vom Bilde zum Buchstaben* (1939); M. SPRENGLING, *The Alphabet: Its Rise and Development from the Sinai Inscriptions* (1931); H. STURM, *Einführung in die Schriftkunde* (1955); S.W. THOMPSON, *The A B C of Our Alphabet* (1942); J. TSCHICHOLD, *Geschichte der Schrift in Bildern* (1946; Eng. trans., *An Illustrated History of Writing and Lettering*, 1947).

(D.D.)

Alpine Folk Cultures

Distinctive Alpine folk cultures extend over the whole area of the Alps, which cover parts of Switzerland, France, Italy, Germany, Austria, and Yugoslavia. Only within the Alpine crescent, where the cultures are similar to those of most other European plains, is there a non-Alpine culture. Alpine cultures are rural and are distinguished from other rural cultures by the presence of extensive pasture grounds above the timberline, called *Alpen* or *Almen* in German, *alpages* in French. One-quarter of Switzerland and one-fifth of Austria consist of such high grasslands. Their exploitation involves seasonal movement of herds and herdsmen between permanent settlements and *alpages*.

Location of the Alpine cultures

Cool mountain valleys with intensive precipitation, which are unsuited for cultivation, will support cattle breeding and dairy farming. Larger, sunnier, and drier valleys permit the growing of produce. Thus either cattle or produce is at the centre of Alpine folk life.

There are historical, ethnic, religious, and linguistic differences in addition to those of climate and ground profile. The main languages are German (33 percent of the Alpine population), French (26 percent), Italian (22 percent), and Slavonian (11 percent). The ancient language of romanized Illyrian-Celtic groups, however, still survives in three minor descendant languages—namely, Rhaeto-Romanic in the Grisons (Swiss canton of Graubünden), Ladin in some Italian Dolomite valleys, and Friulian in the Italian part of the Carniche (Carnic) Alps. Most inhabitants of the Alps are Roman Catholics. The

Swiss cantons of Vaud, Bern, Glarus, and parts of the Grisons and Appenzell, however, are Protestant, and in the French and Piedmontese Alps there are some Protestant groups deriving from the medieval religious movement known as the Waldenses. In addition, there are a few other Protestant groups in the Austrian Alps.

Economic basis of Alpine folk cultures

Alpine folk cultures are characterized more by the type of their economy than by ethnic, political, lingual, or religious features. The biggest contrast is between the northern and eastern Alpine regions, where cattle breeding and dairy farming are exclusive or predominant, and the inner and southern Alpine regions, where cultivation and cattle farming are combined in a self-sustaining subsistence economy. A more exceptional condition exists in the regions where sheep are driven to the high Alpine pastures in summer and to the Mediterranean lowlands in winter.

The regions of the Alps can be classified according to economy and occupation as follows: regions of exclusive cattle breeding and dairy farming—northern Savoie (Savoy) in France; Gruyère, Bernese Oberland, the Four Forest Cantons (Uri, Schwyz, Unterwalden, Luzern), Glarus, and Appenzell in Switzerland; regions of predominant cattle breeding and dairy farming—Allgäu, Upper Bavaria, and Berchtesgaden in Germany; Vorarlberg, Tirol, Salzburg, Salzkammergut, Kärnten (Carinthia), and Steiermark (Styria) in Austria; upper Kranj (Carniola) in Yugoslavia; regions with a balanced agricultural and pastoral subsistence economy—Dauphiné and southern Savoie in France; Valais, Ticino, and the Grisons in Switzerland; parts of Piedmont, Alto Adige (South Tirol), and Friuli in Italy; regions of transhumance—southern French Alps from the Mediterranean to the Durance River in France; parts of Piedmont and the Orobie (Bergamasque) Alps in Italy.

TRADITIONAL CULTURAL PATTERNS

Because the Alps have sporadic steep slopes, a short growing season, and a poverty of exploitable soil, the people have been forced to seek subsistence over large areas and at different altitudes by means of seasonal movements of cattle and the storage of hay for the long winters when the pasture grounds are covered with snow. The tyranny of the steep slopes, the conquest of which takes much human and animal effort, puts the Alpine peasants and herdsmen at a disadvantage, as compared to people of the lowlands. The isolation of the mountains from the lowlands, however, is only a relative one. At all times, the life of European mountain dwellers has been connected to the economies of the lowlands. There dairy products, cattle, and wood are sold, and population surpluses are absorbed temporarily or permanently.

There are many relics of archaic culture in the Alpine countries today. Wooden tally sticks, as well as vessels, lamps, and stoves of potstone, are often still in use. In some places children can be seen playing with toy cows made out of branches and twigs. Sledges for use over grass and snow, wooden plows, carts with wooden wheels that have no metal tires can still be found more often than one would expect. Alpine dialects are characterized by many archaic traits. Customary law is still widespread and controls such things as community life, free-grazing systems, dowry, and inheritance.

Because the Alpine environment provides but a scanty basis for making a living, there is often rivalry among members of the community, but there is also, by necessity, cooperation between families when there are tasks to be accomplished that cannot be done by one family alone. The relationship between family individualism and collectivism is best expressed by a saying that mountain dwellers cooperate only when they have to; yet, they often have to.

Individualism, stoicism, and conservatism

Since Roman days Alpine peoples have been known for their struggle for freedom and independence. Feudalism was never as authoritarian in the Alps as in the lowlands and faded away sooner than elsewhere. In wartime the Alps were often a refuge of liberty. The popular association between freedom and Alpine regions reflects an actual relationship. While European lowlands often gave rise to huge centralistically organized states, Alpine communities united in more or less self-governing confederations (Tirol, the Grisons, the Four Swiss Forest Cantons, Valais, Savoie, Valle d'Aosta, Briançonnais, Vallée de l'Ubaye). In Switzerland and Austria, Alpine culture is considered an important part of national culture, though only small portions of the respective populations belong to the groups of Alpine herdsmen and peasants.

Among the Alpine dwellers, physical strength, endurance, and toughness are much appreciated, as is indifference to pains and diseases and composure in the face of death and heavy blows. The difficult conditions of the Alpine economy require great caution. Well-tried solutions, though not perfect, are preferred to uncertain experiments, since the vicissitudes of life in the Alps do not permit failure. Initiative and innovation are thus not encouraged, while tradition is strongly respected. Alpine populations are conservative, especially in economic matters. A strong religious feeling of dependency arises from being at the mercy of the overwhelming forces of nature. This, however, does not lead to fatalism but, instead, creates a tenacious will for survival.

Social organization. *Local and territorial organization.* Although kinship no longer has any formal political or economic importance nowadays, there is still a strong feeling for kin in the whole Alpine region. This has its roots in the centuries' old residency of families within the same communities or even homesteads. When looking at gravestones or crosses in an Alpine churchyard, one notes only a few different family names, for most people are interrelated by blood or marriage. Often, though, there is feuding within the community among the most powerful kindreds in recollection of pristine, deadly feuds. Only upon threat or attack from the outside do such communities unite, as they have in the past, to prevent total subjugation.

Kinship and community

Besides kinship, there is a territorial social unit that has gained more and more significance. The communal territory is usually an economic unit bounded by natural limits. In the past, this furnished most of the subsistence for the members of the community. Whereas in the Alpine forelands economic individualism since the 18th century has led to the splitting up of community-owned land, in the Alps themselves, pastures and woods are still usually owned in common. There, to a large extent, cattle farming depends on pasturing on common high grasslands and on greens near permanent settlements. This calls for rigid regulation of privileges and duties, especially those involving obligatory communal services. Community endogamy is widespread, and the sale of land to persons from the outside is frowned upon.

Within the community, however, groups of neighbours may form small economic units owning pastures and woods, baking houses, washhouses and, in places, even a chapel.

Most communities are also parishes, and the parish church is the spiritual centre of the community. In addition to regular religious services, there are annual festivities commemorating the patron saint and the dedication of the church, which include processions and the offering of prayers for the common concerns of the community. Baptisms, weddings, and funeral services take place in the community church, and former community members from the outside are often buried in the local churchyard.

In the past, Alpine communities in many places united to form larger political bodies, in order, for example, to monopolize transportation of goods across the Alps. In the Middle Ages, many such autonomous and semi-autonomous confederations developed, of which the Swiss proved to outlast them all.

Family and kinship patterns. Within the Alpine region the male-centred nuclear family, consisting of the parents and their unmarried children, is the rule. Forming an economic unity, its members strongly depend on each other and subjugate themselves to the common goal. Children generally leave their paternal house upon marriage and start a home of their own. Parents help them by lending some land, cattle, or money to be charged upon their hereditary claims. Wherever partible inheritance is

Patterns of inheritance

in use, parents generally remain proprietors of their holdings until death. In regions in which one of the offspring —usually the eldest son but, at times, the youngest—inherits the whole holding, the parents often retire with the understanding that their subsistence is guaranteed until life's end; other children are paid in coin or promissory notes. The holding is ceded to a single heir in German and Austrian Alpine provinces (with the exception of Vorarlberg and the western part of Tirol) and in the north of Switzerland, as well as in the Dauphiné. There is a high rate of illegitimate children (20 to 30 percent) in the Austrian regions with impartible inheritance. This is a consequence of the late retirement of the parents. Women do not like to marry before their grooms have taken over the parental holding, since they are afraid of being servants of their parents-in-law.

In Valais, Ticino, and the Grisons, in Vorarlberg and the western part of Tirol, and in the Italian Alps, the property is split up equally among the children upon the parents' death. This means great parcelling of landholdings and even of dwellings and barns. Through marital alliances, however, there are attempts to join inheritances in parcelled land so that new single holdings are formed. This is often supplemented by buying additional ground. Marriage between close kin is not infrequent since marriage with persons from the outside is frowned upon, thus allowing only a limited choice of partners. In endogamous Alpine regions the group of relatives with whom one deals and who can be asked to help in times of need is rather extensive and can include even third cousins. Spiritual kinship of sponsors leads to a permanent and close relationship between godfather and godmother and the family of the godchild.

Stratification. The lack of large estates in the Alps precluded the development of a rural proletariat. Most holdings are run by one family. Farmhands who are not blood related are rare and mostly found in the Austrian Alpine provinces and in upper Bavaria. Wherever such dependencies exist, they have a rather paternalistic character; there is hardly any difference between the way of life of owners and servants.

Status within a community depends largely on the size of one's landed property and herds. Peasants owning little or no land who can only keep goats instead of cows and who depend almost entirely on the common have a low status. Those newly settled in the village have either no or restricted privileges in the utilization of community-owned land and are considered second-rate burghers. When there are miners living next to peasants, as in the mining regions of Kranj and Steiermark, they form distinct groups, not mixing with the peasants. The same is true of Austrian woodsmen. Decision-making power in community affairs is primarily in the hands of males of the landowning families of long-standing, even if they are a minority in the communal population.

Role of "outsiders" and non-peasants

Socialization and education. Children are required to help with chores as soon as they are old enough to be able to do so. One of their first tasks is to bring meals to the places of work in the fields. Later, they watch the herds; this gives them a chance to play out-of-door games. Indoor playing is rare. Children's games are not encouraged by adults. Although formal education does not seem to be essential for mountain farming, school attendance, especially in winter, is good. Illiteracy, which still persists markedly on the Mediterranean, is practically unknown in the Alps. On reaching 12 years of age, boys and girls are considered full-scale farmhands. In earlier days boys would join the political community council and the military at age 16, or even 14.

Clubs or societies of young unmarried males (*Knabenschaften*) used to be widespread all over the Bavarian, Austrian, and Swiss Alps and had an important function for adolescents in molding them in accordance with local customs and in maintaining the customs. Nightly visits of unmarried young men to single girls are governed by custom in Austria, Bavaria, and Switzerland; these can be either individual—for serious suitors—or collective, when several young men call on a girl. Since almost every family is interrelated, baptisms, weddings, and burials are a matter of almost everybody's concern in isolated mountain communities.

Economic systems. *Settlement patterns and housing.* In the 11th, 12th, and 13th centuries, Alemanic and Bavarian tribes formed dispersed homestead settlements, which are still typical of the northern Swiss, Bavarian, Austrian, and South Tirolese Alpine landscapes. In the French and Italian Alps, life is centred on the villages, which often have a townlike character; in the Alps de Provence the villages are characteristically perched on cliffs in natural defensive positions. The hamlet is common nearly everywhere in the Alps, since it combines the advantages of the compact settlement with those of the single homestead and permits the farmer to live nearer his land.

Alpine homesteads and villages

Because the Alpine peasant gets his subsistence by covering a large area and at different heights, he is often compelled to maintain several buildings. He may own numerous agricultural buildings, such as barns for hay and grain and so forth, combined with stables, at more distant meadows. The multiplicity of buildings allows for only very simple equipment in each of them. A sort of temporary station between the main settlement and the high pasturelands is called *Maiensäss* in German or *mayen* in French; there cattle are grazed on private lands before being driven to the *alpages.*

The most common building material in the Alps is wood. Houses and huts on the northern slopes of the central Alps and in Austria are made of horizontally laid logs of spruce or fir. These log cabins are comparatively small in size, limited to the length of the logs, and rectangular in shape. Several such log cabins can be united to make a bigger house. In the Alpine areas from Valais, via the St. Gotthard Massif, to the eastern limits of the Grisons, living rooms and bedrooms are built in log-cabin style; the kitchen, however, is made of stone. A distinguishing feature of the southern slopes of the Alps and the southern French Alps is often a towerlike building in which stable or cellar, kitchen and living room, bedroom and storeroom are positioned one on top of the other. The typical house of the Engadine (in the Grisons) is a log cabin revetted by masonry. In the uppermost parts of some valleys of Savoie, Dauphiné, and Piedmont, a more primitive type of house can be found in which the inhabitants live with the animals during the winter in the bottommost room, separated from them by only a fence-like partition.

The Alpine homestead can combine living area, barn, and stable under one roof, as in Bernese Oberland, Tirol, and Salzburg; or, as in other areas, it can consist of a group of buildings that may be scattered or united in three- or four-sided farmsteads. Roofs are comparatively low angled, so that the roof shingles or slabs of stone covering the roofs cannot slip off and so that the snow can remain indefinitely to insulate the house from storms and cold.

Seasonal cycle of pastoralism and agriculture

Patterns of production and technology. In winter cattle are stall fed in permanent settlements; in summer they are grazed at high altitudes, the *alpages.* While the cattle are in the care of cowherds on the *alpages,* the owner's family cuts, dries, and collects the hay necessary for winter feeding from its meadows and scattered grassy spots. In mixed farming areas during this period, the grain is harvested, the fields plowed, and the vines, if any, tended.

In summer most of the milk is made into cheese; in winter it is sold to a cooperative village dairy, which either makes it into dairy products or sells it.

In the inner and southern Alpine valleys, where mixed farming prevails, only patches of land that fulfill specific conditions regarding available sunshine, soil quality, moisture retention, and slope gradient are suitable for cultivation. Such patches are very often scattered over a large surface. Fields, meadows, and vineyards are sometimes hours apart from one another and can be reached only by mule or on foot. Hay, grain, wood, and grapes have to be carried by pack animals or, often enough, on the human back or shoulder. The cultivation of slopes tends to cause the soil to slip downward; periodically, therefore, it has to be carried up on to the patches again.

There is little mechanization in Alpine farming. Only hoes can be used on the very steep slopes, and grain is frequently cut with the sickle and threshed with flails or primitive threshing sticks. It does not ripen properly in the higher areas and therefore has to be dried on special drying racks, which may be separate constructions, as in the Austrian Alps, Kranj, and parts of the Grisons and of Ticino, or built on to other farm buildings in the form of balconies.

Patterns of transhumance

The southern French Alps have a special place in the Alpine economy because of their arid pastures that make them less suitable for cattle than for sheep and goats. Besides the sheep that are raised locally, there are large flocks that are brought in to graze during the summer and that belong to farmers who do not live within the Alps. In winter all sheep are grazed in the snow-free areas of the Camargue and La Crau and are never stabled. The ewes are milked, and the milk is either made into cheese locally or sold to the commercial producers of such cheeses as the famous Roquefort. A similar form of transhumance is practiced by shepherds of the Orobie Alps, who, in winter, pasture their sheep in the lowlands between Bergamo and Milan and drive them up to the high Alps in summer.

In thickly wooded areas, particularly in the Austrian Alps, forestry is of some importance. Carpenters of the Salzkammergut and of the Tirolean Inn Valley have developed a high standard of wooden-house and bridge construction. Mining is also important in Kärnten and Steiermark.

Carters and muleteers played an important part nearly everywhere in the Alps before the advent of the railway and motorized traffic.

Up until the 20th century, home handicrafts were made in many mountain valleys in winter. Nearly everywhere, women spun and wove their own cloth and made carpets and embroideries either for their own use or sometimes for the outside market; the men in some places carved wooden vessels and figures.

Traditional seasonal migration

Migration played an important part in winter in the French Alps and in many valleys of the southern Alps. The migrants worked mainly as door-to-door peddlers or occasional workers, such as chimney sweeps or tinkers in larger towns. Many were also attracted to jobs in transport, such as coachmen or, later, taxicab drivers. The migrant workers of Tirol left in spring and returned in autumn. Even children had to leave the overpopulated mountain areas of Tirol and the Grisons to mind the cows in Swabia during the summers.

Because Alpine communities encouraged little individual initiative, Alpine migrants preferred occupations in which they were directed, such as the occupations of policemen or, formerly, of mercenary soldiers. Temporary migration often became permanent, but it was customary for the emigrants to visit their native village during the summer season for generations. Many emigrants retire in their original Alpine homes in old age.

Property and exchange systems. Ownership of land and cattle is the chief measure of status; tenant farming is rare. In the inner Alpine valleys of Switzerland (Valais, the Grisons, and Ticino), an individual has his property widely scattered, often owning more than 100 patches of land, not to mention the various buildings. There is less scattering of property in the northern and eastern Alps, where cattle breeding is predominant. Originally, anyone could use the Alpine pastures, but, because of the growing population, grazing rights had to be limited: the participant had to live in the community that owned the pastures, and he could only graze as many cattle there as he could feed in winter. Many pastures belong to cooperatives, which give each individual certain grazing rights. In the Swiss Alps the grazing rights of every individual member are, in a few places even today, set out on tally sticks. On producing the stick, the owner may drive the indicated number of cows to the cooperative pastures. There are often regulations forbidding the sale of grazing rights to outsiders. It should be noted that some highland pastures are owned wholly by individual families.

Grazing rights

People of Alpine communities are often contemptuous of laws and state regulations running counter to local traditions. Poachers lose no honour upon prosecution, since the common opinion is that game belong to the people, not to the state. Similarly, state forestry laws are considered as an arbitrary intrusion into established local rights, dating from ancient days.

Milk, milk products, and cattle—wherever these are the mainstays—bring in cash, which otherwise is rare in these largely self-sufficient economies. Where there is little cash, there is little paid labour. Therefore, it is all the more common for neighbours to help each other with tasks that are too much for the individual family. By and large, Alpine regions are among the poorest in Europe in capital and income.

Traditional religion

Beliefs and aesthetic systems. Alpine countries are a stronghold of Roman Catholicism. Even the fleeting visitor cannot fail to notice the comparatively large churches, the countless chapels and roadside shrines and crucifixes. Priests bless both pastures and cattle after the spring migration to the *alpages*. In the Four Swiss Forest Cantons and in the mountains of St. Gall and Liechtenstein, the herdsman chants his prayer every evening out over the pastures and the cattle in his care. His prayer is believed to be effective as far as it can be heard. The cult of saints is widespread. There are pilgrimages to sanctuaries, some at very high peaks, such as the Rocciamelone in Piedmont at an altitude of 11,600 feet (3,500 metres) above sea level. In Swiss Protestant mountain regions many inhabitants, especially those of isolated homesteads, belong to sects.

Religion in Alpine areas, however, is a traditional rather than an individual creed. Persons often abandon it when they move into towns where there is no immediate control over their religious comportment.

The harshness of life gives Alpine people a strong rational nature. Because of the relative scarcity of professional medical attention, many resort to folk remedies using Alpine herbs and animal fats, and there is a widespread belief that phases of the moon influence plant growth, weather, animals, and human beings; but it cannot be said that the Alpiners are much inclined to belief in magic or superstition.

Festivities and carnivals

The more important cattle breeding is, the more festive is the driving of the herd up to the high pastures at the beginning of summer and its return some three months later. Led by herdsmen in colourful costume, the animals are decorated with flowers and wear large bells around their necks. Northern Switzerland, which is particularly famous for these festivities, also has Alpine games, usually while the cattle graze on the high pastures—games such as *Schwingen* (a form of wrestling) and *Steinstossen* (a form of putting the stone). In many Alpine valleys of Austria and Switzerland, there are carnivals having a kind of archaic character, with the wearing of wood masks and the worship of nature deities and ancestors.

Folk art in Alpine countries is expressed in music, poetry (usually song), dance, wood carving, and embroidery. In the cattle-breeding northern areas, there are many primitive forms of song and music involving, for instance, yodelling, a way of singing marked by rapid switching of the voice to falsetto. There are also trumpet-like instruments made of bark, the perfect form of which is the alphorn. Folk music in mainly pastoral areas, such as northern Switzerland, Bavaria, the Austrian Alps, and upper Kranj, has wide-ranging, floating melodies, whereas in the crop-growing regions of the inner and southern Alps, more songlike melodies of limited range are common. The most frequent poetical themes are love and longing for the homeland, as well as historical, patriotic, pastoral, and hunting themes. The vitality of the Alpine folk culture can also be seen in very expressive dances, such as the *Schuhplattler* (in Upper Bavaria, Tirol, and Salzburg), with its rapid sequence of jumping and hopping movements.

Wood-carving consists partly of chip carving for the decoration of everyday objects, such as milking stools, neckbands for bells, wooden spoons, and distaffs, and partly of figure carving, especially of nativity figures (in Upper Bavaria, Berchtesgaden, and the Ladin valleys of

South Tirol). Decoration of house facades with sayings usually of a religious nature is widespread in Protestant Alpine areas (in Bernese Oberland and parts of the Grisons), but it can also be found to a lesser extent in Catholic regions, such as German-speaking upper Valais.

Embroidery, as a form of folk art, can be seen especially at prominent points of women's clothing, such as cuffs, stomachers, hats, scarves. It has become a home industry in parts of eastern Switzerland.

EVOLUTION OF THE CULTURES TODAY

Until the 19th century there were very few alternative ways of making a living in the Alps. Both the pastoral and crop-growing economies had reached their definite form at an early stage. With the comparatively primitive standard of technology, everyone within the village community could achieve more or less the same socio-economic level. They worked and lived as they had learned from their fathers, and the strict social control of the village community ensured the continuity of tradition. There was a balance between the number of people and animals, on the one hand, and the available agricultural land, on the other—a balance that could be maintained by the well-tried measures of extensive celibacy, late marriage, and temporary or permanent emigration. In this way Alpine communities were able to withstand almost unchanged even such drastic historical influences as the Reformation and the French Revolution. All their customs and their whole ethos served to maintain the status quo, which was seldom questioned. A basic feature of these self-sufficient communities was a strict conformity in all aspects of life that allowed for little individual initiative.

Effects of the Industrial Revolution

The outbreak of the Industrial Revolution found the Alpine population unprepared. When the Alps were opened up by roads, rails, and cableways and were exploited for lumbering, tourism, and hydroelectricity, mountain dwellers began to recognize the differences in standard of living that had cut them off from the people of the lowlands. Furthermore, a new type of worker-peasant evolved when factories were set up at the mouths of the French Alpine valleys, in Valais, and in northern Italy; he would commute daily or weekly between home village and factory and so combine traditional agricultural occupations with factory work.

More money came to the Alpine valleys from casual earnings in forestry and in the construction and maintenance of roadworks, avalanche protective works, hydroelectrical installations, and pipelines; to these earnings were added state subsidies to the mountain economy, children's allowances, and old-age pensions. All these financial aids made possible a certain degree of mechanization, with tractors and motorized mowers and plows in places in which the quality and position of the ground permitted it. Even electric fences and pneumatic hay blowers were put to use.

Widespread abandonment of farming

An economy based on money and rationalization soon made less indispensable the services that the village community had performed for individual families for centuries. It became obvious that it would never be possible to make mountain farming as productive as lowland farming. Increasing emigration led to the abandonment of parts of the cultivated land. Thus, today, especially in the inner and southern Alps, one can find deserted homesteads, hamlets, and even whole villages. Farming is carried on in many places only by the old people who are by and large resistant to innovations. At present the worker-peasants still help to slow down the depopulation of mountain villages, but it is improbable that the children will continue the split existence of their fathers. Nonagricultural part-time earnings are often only a transitional phase in the complete abandonment of farming.

Even in the once wealthier pastoral northern Alpine regions, the situation is becoming critical. More and more pastures are being abandoned because of a lack of hands. Cowherds will accept the isolation of their job only for very high wages, which jeopardize the profitability of the enterprise. There are marketing difficulties for milk products made on the *alpages*, because they can be produced more efficiently elsewhere in factory-like operations.

Social and economic progress seem to have united against the exploitation of the Alpine pastures. Not only cattle in the Alps but also sheep in the traditional regions of transhumance are decreasing in number because of the increasing difficulty of finding grazing areas for wintering close to the Mediterranean coast, the land there being intensively cultivated throughout the whole year now.

Tourism

Because the introduction of industry to mountain villages has usually not been a success, tourist development, especially involving winter sports, is, at present, the most promising way of warding off the depopulation of the Alps. Not all Alpine areas, however, are equally attractive for tourists. For the people of the Alpine communities, brought up strictly in the traditional way, there is little of the initiative necessary to assume a successful role in tourist development. Besides, there is locally neither the necessary capital nor the professional training. Some parts of the Alps, thus, are currently the scene of spectacular investments by outside developers who are turning mountain areas into technocratized, highly elaborate recreational resorts. Landownership, once the pride of the Alpine peasant, is now increasingly reverting to urban developers and to city dwellers who can afford a holiday house in the mountains.

The traditional costumes of Alpine cowherds and peasants and their wives, Alpine music and dance, and colourful local customs have become part of the tourist-entertainment industry, and via the mass media they are now integrated into the international folklore.

These changes have not occurred in all regions to the same degree and in the same way. There are still strongholds of tradition quite close to areas in the full process of urbanization. But, everywhere in the Alps, traditional folk culture is involved in a struggle for which it is not sufficiently prepared and in which external forces of change are stronger than internal forces of tradition.

BIBLIOGRAPHY. A systematic, comparative study of Alpine folk cultures taking in the Alps as a whole, does not yet exist. A good preliminary view on the Alpine mixed-farming areas may be found in ROBERT K. BURNS, "The Circum-Alpine Culture Area," *Anthrop. Q.*, 36:130–155 (1963). There are innumerable local or regional monographs written by ethnographers, folklorists, and geographers; of the more comprehensive ones that merit mention none has been translated into English. JOHN FRODIN, *Zentraleuropas Alpwirtschaft*, 2 vol. (1940–41); and RICHARD WEISS, *Das Alpwesen Graubündens* (1941), both deal with Alpine economy, material culture, and customary law. Alpine folk life is described in PHILIPPE ARBOS, *La Vie pastorale dans les Alpes françaises* (1922); ARNOLD VAN GENNEP, *Le Folklore du Dauphiné: Isère*, 2 vol. (1932–33), and *Le Folklore des Hautes-Alpes*, vol. 1 (1946); RICHARD WEISS, *Volkskunde der Schweiz, Grundriss* (1946); MICHAEL HABERLANDT (ed.), *Oesterreich, sein Land und Volk und seine Kultur* (1927); PAUL SCHEUERMEIER, *Bauernwerk in Italien, der italienischen und rätoromanischen Schweiz*, 2 vol. (1943–56); VILKO NOVAK, "Übersicht über Viehhaltungsformen und Alpwesen in Slowenien," in LASZLO FOLDES, *Viehzucht und Hirtenleben in Ostmitteleuropa* (1961); and RUDOLF KRISS, *Sitte und Brauch im Berchtesgadener Land* (1963). Useful detailed information on cultural items may be found in the national ethnological atlases and in their comprehensive commentaries: *Atlas der schweizerischen Volkskunde* (1950–); *Österreichischer Volkskundeatlas* (1959–); and *Atlas der deutschen Volkskunde* (1937–39 and 1958–). Alpine settlement and housing is handled best in RICHARD WEISS, *Häuser und Landschaften der Schweiz* (1959); and MAX GSCHWEND, *Schweizer Bauernhäuser* (1971). Of special interest for the understanding of today's development are JOHN W. COLE, "Economic Alternatives in the Upper Nonsberg," *Anthrop. Q.*, 42:186–213 (1969); and PLACIDE RAMBAUD and MONIQUE VINCIENNE, *Les Transformations d'une société rurale, la Maurienne (1561–1962)* (1964), the latter dealing with the Maurienne valley in Savoy. URS JAEGGI, *Berggemeinden im Wandel* (1965), is a study about social and cultural change in the Bernese Oberland; GERALD BERTHOUD, *Changements économiques et sociaux de la montagne* (1967), describes recent economic, social, and cultural change in a small mountain community of Valais.

(A.Ni.)

Alpine Lakes

The lakes of the Alps, set in some of the most magnificent and rugged of European landscapes, have often served as

a focus for settlement and an attraction for tourists; as a result, many have been affected by various forms of pollution. All are of considerable scientific interest. Most of the Alpine lakes lie in valleys that were formed by mountain-building processes during the uplift of the mountain chain of the Alps. During the Ice Age of the geologically recent Pleistocene Epoch, less than 2,000,000 years ago, glaciers flowed through these valleys, deepening and excavating the ground, and leaving moraines (deposits of waste material) when they shrank at the end of the glacial period. Water filled up the excavations or was dammed up by the moraines. The lakes that originated in valleys of the mountain region are of a long and narrow shape and are generally very deep (*e.g.*, Como, with a depth of 345 feet [410 metres]). In some cases the glaciers advanced from the Alps into the adjacent plains where they began to diverge fanwise. In such cases the terminal part of the associated lakes broadens (as in Lago di Garda) or bifurcates (Lake of Constance). The lakes are divided into a northern and a southern group by the Alpine watershed running from west to east. The southern group is made up of Lake Geneva and the Insubrian lakes (Maggiore, Lugano, Como, and Garda). While this group lies in an Alpine environment, parts of the northern lakes are situated in the foothill zone of the Alps or even some distance beyond it (*e.g.*, Lakes Neuchâtel, Zürich, Constance, Chiemsee, Attersee). The Wörther See lies in an inner Alpine depression of the Eastern Alps.

Scientific study. The scientific study of the Alpine lakes started in Switzerland. The stationary oscillations of the water level (seiches) caused by the wind were studied by F.A. Forel, who also made classic observations on the interrelationship between physical and biological processes in lakes. In his work *Le Léman* (1892–1904) he created the term limnology to characterize the comprehensive study of lakes. Later on, the increasing pollution of Lake Zürich caused attention to be paid to chemical and biological changes. In the lakes of the Eastern Alps the phenomenon of the thermocline (zone of rapid decrease of lake temperature below the warm surface stratum in summer) was first studied in the Wörther See (1891). In the same lake, in 1931, the discovery was made that there was an absence of total water circulation during the winter in lakes with wind-sheltered sites. These lakes were henceforth characterized as being of meromictic type. In Lake of Constance, Auerbach, Maerker, and Schmalz investigated the currents caused by the Rhine, which flows through the lake (1926). By the late 20th century, a number of institutes were engaged in the study of the pollution of the Alpine lakes.

Composition. The water composition of the Alpine lakes is quite uniform. The chief constituent in solution (up to 96 percent) is bicarbonate associated with calcium, to a lesser degree with magnesium. Brown humic (derived from organic decay) substances may occur in different quantities. Small amounts bring about a shift from the blue colour of pure water (*e.g.*, Lago di Garda) toward green (Lago Maggiore), larger quantities cause an olive or brown-green coloration. The content of phosphate, although very small, is of importance for the development of the microscopical algae known as phytoplankton, which float in the upper strata of the lakes. About 100 years ago almost all Alpine lakes were poor in plant nutrients, especially in phosphates. During the 20th century, many lakes were manured by the water wastes from households and hotels, and thus the phosphorus content increased and the algae multiplied, a process called eutrophication and a major factor in the pollution of the Alpine lakes. The extreme growth of phytoplankton under these conditions makes the water turbid and less suitable for bathing. It also intensifies oxygen consumption in the deep layers of the lake as a result of the increased decomposition of dead algae. In extreme cases the spawn of some fish species developing near the bottom can be endangered. Two methods are used as a corrective to eutrophication. In Switzerland attempts have been made to remove the organic substances by mechanical and biological purification and also to eliminate the phosphate by additional treatment. In Germany pipelines around the lake borders are used to collect the waste water from the catchment areas of the lakes.

Pollution

Lake Geneva (Lac Léman). Lake Geneva is the largest of the Alpine lakes. The main source of its inflowing water is the Rhône (with a water flow of 6,350 cubic feet per second [180 cubic metres per second]). Each year the Rhône brings 105,000,000 cubic feet (3,000,000 cubic metres) of sediment into the lake. The northern shore is Swiss and rises to charming and fertile slopes covered by gardens and vineyards. The southern is French and less cheerful, but the steep ranges in the background are imposing. In the west the lake becomes narrow, and this region is known as the Petit Lac. The climate of the lake area is semicontinental and humid. The mean yearly temperature is 48.9° F (9.4° C), and the precipitation amounts to 39 inches (1,000 millimetres) a year. As a result of the mild climate the lake is never frozen over, and its surface layers warm up to 72°–75° F (22°–24° C) in summer. The water is blue and transparent, but symptoms of eutrophication are nevertheless to be seen in the environs of larger towns. The northern shore is densely populated, with Geneva (Genève) and Lausanne as the largest settlements. The region around Lausanne represents the largest zone of viticulture in Switzerland. Along the shore there are many winter and health resorts, notably Montreux and Vevey. To the south of the lake, the famous French spas Évian-les-Bains and Thonon are of importance. The lake is not as rich in fish as other Alpine lakes. The most valuable fish is the "fera," a kind of whitefish.

Lake of Constance (Bodensee). After Lake Geneva the Lake of Constance is the largest Alpine lake (see the Table). It fills the former valley of the Rhine between Bregenz and Stein. Because of structural depressions and the spreading out of the former Rhine Glacier, the main basin (Obersee) bifurcates at its western end into the Überlingen and the Untersee. Politically, the northern shore belongs to Germany, the southern is Swiss, and the eastern end is Austrian. The chief influent, or water source, is the Alpenrhein, with a flow of 8,010 cubic feet per second (227 cubic metres per second). It brings about 116,450,000 cubic feet (3,300,000 cubic metres) of rubble into the lake basin every year—sufficient to fill the lake up within 12,000 years. The lake discharges through the Hochrhein, which leads away 423,000,000 cubic feet (12,000,000 cubic metres) of water a year to feed the famous cataracts of Schaffhausen. The mean temperature of the region is about 47.5° F (8.6° C). Precipitation increases with proximity to the surrounding mountains: from Constance (Konstanz), 31 inches (790 millimetres), it increases to 59 inches (1,500 millimetres) at Bregenz. The temperature of the lake water does not exceed 72° F (22° C) in summer. In winter the lake has been frozen over only once or twice per century. As the result of pollution the quality of the lake water has decreased rapidly, especially since the mid-20th century. The mass of phytoplankton has grown tenfold, the nutrient conditions for fish have improved, and their growth has been intensified. On the other hand, the eutrophication of the lake has endangered the water supply of various parts of southern Germany. The Stuttgart region takes drinking

Alpine Lakes

lake	area (km^2)	length (km)	depth (m)	altitude (m)	territory
Geneva	581	72	310	372	Swiss–French
Constance	541	66	252	396	German–Swiss
Garda	370	54	346	65	Italian
Neuchâtel	218	38	153	429	Swiss
Maggiore	212	54	372	194	Italian–Swiss
Como	146	50	410	198	Italian
Luzern	114	38	214	434	Swiss
Zürich	88	29	143	406	Swiss
Chiemsee	80	15	74	518	German
Attersee	46	20	171	467	Austrian
Wörther See	19	17	85	439	Austrian

water from a pipeline from the Überlingen. There are many small villages around the shore of the lake, but the total number of inhabitants is under 250,000. In the western part of the region, which is favoured by a milder climate, the land supports vegetable gardening, fruit farming, and viticulture; in the east, around Friedrichshafen, there is some industry. A chain of summer resorts and watering places is evidence of an important tourist traffic.

Lago di Garda. This is the largest and bluest of the Italian lakes. The lake extends from north to south and is made up of two quite different parts. The upper (northern) part fills a fjordlike valley of structural origin; the lower end extends into the Lombardian Plain and is raised by dams formed of morainic (glacial transported) material. The chief influent is the Sarca, which flows from the north. The lake is drained by the Mincio, which runs into the Po. The most conspicuous attribute of the lake is its opal-like blue colour due to the calcareous environment. The region is very favoured climatically because of its site along the south slopes of the Alps (January mean temperature 36°–37° F (2°–3° C). The southern part of the region surrounding the lake is densely populated, with a large number of watering places and health resorts. Palm, olive, and lemon trees winter in the open air. Lake navigation is now of little importance, since a most beautiful scenic route (the "Gardesana") was built around the lake from 1932–34.

BIBLIOGRAPHY. The classic work on this subject is EDUARD RICHTER, *Die Temperaturverhältnisse der Alpenseen* (1891). For current information, see B.H. DUSSART, "Les grands lacs d'Europe occidental," *Année Biol.*, vol. 2 (1963), a general description of the lakes; and G.E. HUTCHINSON, *A Treatise on Limnology*, 2 vol. (1967), a basic work on modern limnology.

(I.F.)

Alps Mountain Ranges

From the Gulf of Genoa in the Mediterranean Sea, the mountain range popularly known as the Alps reaches north and northeast as far as Vienna to form a giant spine in south central Europe. These are, however, only a small segment of a discontinuous mountain chain stretching a quarter of the way around the globe, from the Rif Atlas Mountains of Africa across southern Europe and Asia to beyond the Himalayas. From the peaks and uplands of their 750-mile (1,200-kilometre) length and 75- to 125-mile (120- to 200-kilometre) breadth flow many of the Continent's major rivers: the Rhône, the Rhine, most of the Danube, the Adige, and the Po. Main communications thoroughfares wind tortuously through the valleys and high passes, and tunnels have been cut through the mountains. In European history, the Alps have served as a barrier, as a unifier, and as a determining influence on the life within the range. The Alps form part of six nations—France, Switzerland, West Germany, Austria, Yugoslavia, and Italy—but none is wholly an Alpine state.

Influences of the Alps on European history

Though they are less formidable and extensive than such other mountain systems as the Andes, the Himalayas, and the Urals, the Alps are responsible for major geographical changes in the heart of Europe. The climates of the Mediterranean and the Atlantic Ocean and those of the land to the north and east are contained by this barrier, which permits only a small part of each to pass through. On mountains rising in some places above 13,000 feet huge glaciers create the impressions of a polar landscape in the midst of a temperate zone. Men have lived from prehistoric times in even the most remote valleys of these mountains, which have given them a great deal in common in terms of economics and way of life despite differences in their political, linguistic, religious, and other allegiances.

Continually, the cultures of the Alps and of the surrounding plains have interpenetrated each other, the vertical world of the former never being completely independent of the latter. In the course of their long, uninterrupted human settlement by men who learned to exploit their many resources, the Alps have become perhaps the most "humanized" of the world's mountains. Though natives once fled from Alpine winters and from the economic and social primitiveness of much of the territory, the Alps are now one of the most sought-after regions of Europe: its resorts for health, mountaineering, skiing, and general tourism are among the most extensive in the world, and its hydroelectric-power resources afford rich possibilities for industrial development. (For information on related topics, see the article EUROPE and those on the six Alpine nations.)

Study and exploration. Records of ascents of various peaks in the Alps date from at least as early as the 14th century, and, in the late 18th and the 19th centuries, the interest in this activity created a vogue for serious mountaineering throughout the world. Horace Bénédict de Saussure, a professor at the University of Geneva, made ascents of the peaks and encouraged others to do so in the 1780s, when he made the earliest scientific observations of the mountains. With the systematization of geology as a science in the second half of the 19th century, Alpine investigation made great advances. The first geological maps indicated the main structural outlines of the mountain system, which revealed unusually complicated geological phenomena that seemed to be the result of something more than a massive upheaval within the earth. One explanation was the theory of nappes—enormous masses of rock apparently squeezed and thrust across long distances by internal pressures in the earth, eventually overlying strata of more recent age and different origin. This theory has been reinforced by the hypothesis of continental drift, which posited the realignment of entire land and water areas during vast geological eras. Today, however, the tendency is to suppose a vertical uplift of the crystalline rock and a sliding down (under the effects of both gravity and the relatively plastic nature of the layers of schists and marls) of the sedimentary cover, and a somewhat more complicated origin for the nappes.

Theories on origins

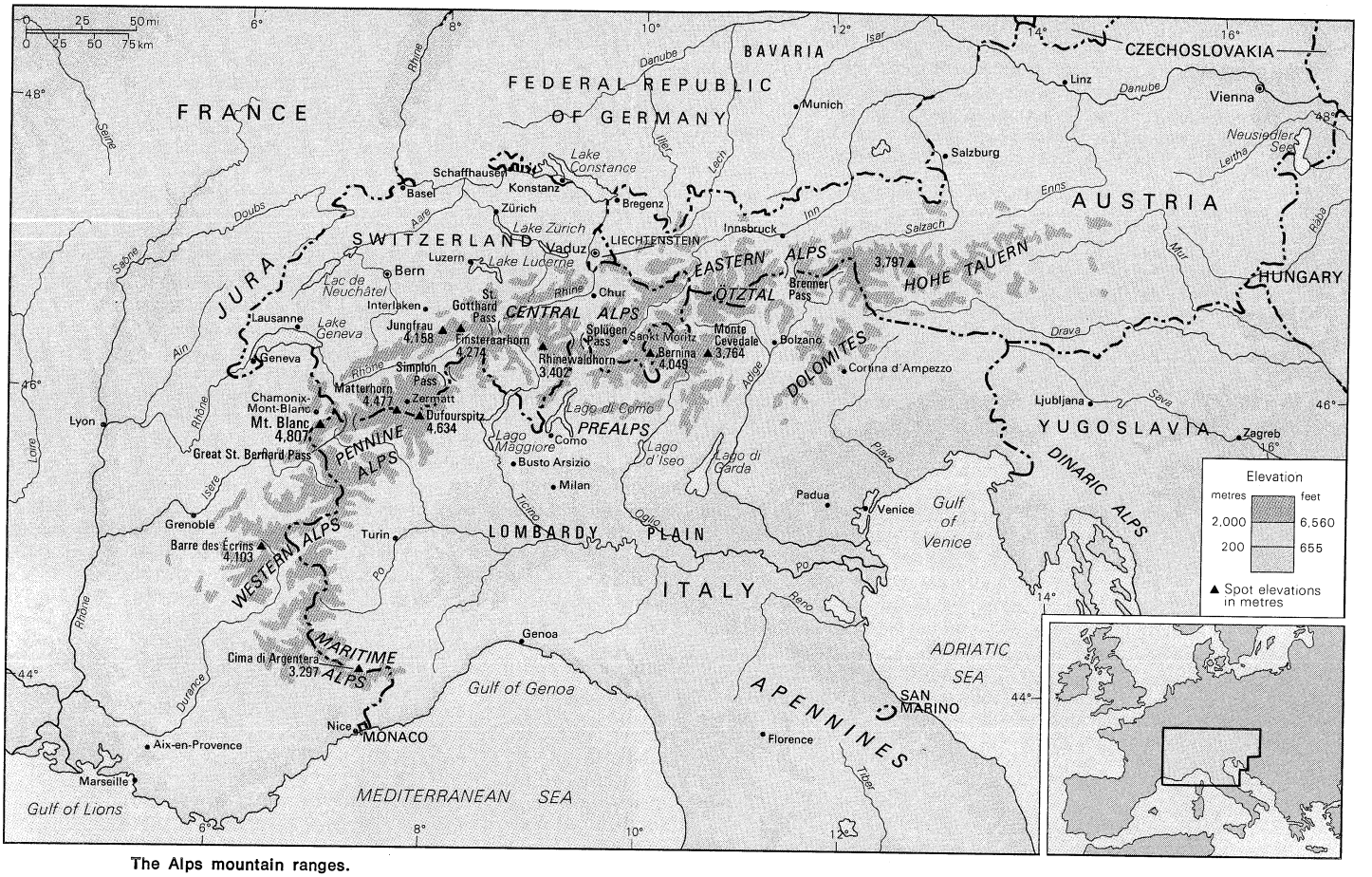
Simultaneous with this conjecture on geological regions, the subjects of Alpine relief, climatology, and vegetable, animal, and human geography also came in for observation and analysis. The Alps may well be the most thoroughly studied mountain system on earth, particularly at the Universities of Grenoble, France, and Innsbruck, Austria; yet they continue to present hosts of complex, puzzling scientific problems.

The natural environment. From the Mediterranean to Vienna, the Alps are divided into western, central, and eastern segments, each of which comprises several separate ranges. The western Alps run from the coast through southeastern France and northwestern Italy, with the Rhône to the west and north, to near the Great St. Bernard Pass on the Swiss–Italian border. Their peaks include Mont Blanc, at 15,771 feet (4,807 metres) the highest in the Alps. The central Alps, however, extending eastward to the area of the Splügen Pass (also on the Swiss–Italian border) and Lago di Como in north central Italy, are generally higher. The eastern Alps spread north into West Germany and south into Yugoslavia and cover most of Austria.

Divisions of the Alps

Geological character. The outer portion of the Alps, a series of ranges known as the Pre-Alps, lie in the extreme west of the system—in France—and in the northeast, in northern Austria in the region of Salzburg and Vienna. The Pre-Alps are, as their name implies, in the nature of lower foothills, and geologically their sedimentary structure is in striking contrast to the harder crystalline rock that forms the double- and sometimes triple-tiered inner Alps. The lack of such pre-Alpine formations in northern and northwestern Italy and the existence of several sedimentary zones between the crystalline chains are relatively minor contradictions to the theories that seem best to account for the formation of the Alps.

The essence of this process, best seen in the western Alps, was the upward and outward thrust of the crystalline material from deep in the earth and the resultant downward slide of the sedimentary crust. When this thrust and sliding occurred unchecked over great distances, nappes were formed, whereas the presence of mere folds suggests the obstruction of the sliding by oth-



The Alps mountain ranges.

er forces. The western Pre-Alps are thus taken to be the result of the force created by the rise of the stronger crystalline massifs to the east, including Mont Blanc. The spread of softer materials was checked by the deeply imbedded land formations of the Continent against which it was being pressed. Surrounding Lake Geneva, however, are two mountain masses with hard cores, amid the Pre-Alps, that were apparently driven great distances over the Mont Blanc region before that uplift occurred.

The greater complexity of the central Alps brings the assumptions underlying the theory of nappe formation into question. It seems more likely that the younger materials reached their lower lying position through breaks in the crystalline structure that was being thrust upward, rather than that they were overlain by those older formations. Numerous geological contradictions also tend to refute the "nappist" position in the eastern Alps, but a new synthesis of scientific opinion is yet to be made.

Topography. The crystalline mountains tower the highest. In addition to Mont Blanc, they include the Pennine Alps, which reach 15,203 feet (4,634 metres) in the Dufourspitze; the Aar Massif, 14,022 feet (4,272 metres); the Bernina, 13,281 feet (4,049 metres); the Ötztal, 12,379 feet (3,774 metres); and the Hohe Tauern of Austria, 12,454 feet (3,797 metres). The Pre-Alps, comparatively low in France, become more massive in Switzerland, Austria, and northeastern Italy.

The shapes of the relief have been formed more by frost and glaciers than by running water. High peaks, the sharp pinnacles known as aiguilles, and vertical walls characterize much of the crystalline regions, notably in the Matterhorn, which at 14,691 feet (4,478 metres) is the most spectacular isolated mountain. Huge cliffs and canyons dominate the landscape in limestone regions, especially in the Dolomites, while broad valleys are common in the areas of softer rock, particularly in France and Austria. Glaciers lowered valley floors, creating great differences in height between summits and adjacent valleys, and depressed the bottoms of such Alpine lakes as Maggiore, Como, and Garda below sea level.

The network of the valleys, an essential part of Alpine relief, is largely transverse, descending outward from the summits like two slopes of a roof. The valleys that run longitudinally, in the direction of the chain's trend, are fewer but are longer and wider. Erosion, mostly glacial, was cut into low points of the ridges to form the numerous passes that have allowed comparatively easy movement through the Alps since prehistory.

Climate. Temperatures generally decrease by 2° F (1° C) with each 650-foot increase in altitude, and winters and periods of frost become longer. At 6,500 feet the mean annual temperature is about 32° F (0° C). The southern slopes are warmer than the northern. Shadows cast by the relief may deprive valleys of sunlight for months during the winter.

The Alps are humid, but the distribution of precipitation is unequal. The outer mountains, which first cause rain clouds to rise and jettison their moisture, receive as much as 79 inches (2 metres) of precipitation a year. The inner mountains, though higher, are less humid, and some areas receive only 20 inches (500 millimetres) of moisture and require irrigation for crops.

Glaciers

Mountains above 10,000 feet may accumulate more snow than can be thawed during the summer months. The effects of pressure and of alternate melting and freezing turn it slowly into the pre-ice known as névé, and finally into ice. From this process the multiform glaciers develop, as domes, amphitheatres, and walls, in gullies and valleys. Glaciers cover about 1,500 square miles. The largest in the Alps is the Aletsch Glacier in southwestern Switzerland, covering an area of over 50 square miles. Glaciers in the Chamonix Valley of France are smaller but more accessible.

Rivers. The abundant precipitation and moderate evaporation are largely responsible for the vast river systems flowing from the Alps. Except in the winter, when levels are low, they are excellent producers of hydroelectric power. The construction of dams has created reservoirs that equalize the flow throughout the year and prevent disastrous floods during thaws.

Vegetation and animal life. Several vegetative subzones exist within the basic forest zone of the lower reaches and the virtually treeless higher pastures of the Alps. Beeches predominate in the lower forest areas, spruce and fir in the higher; larch and pine are found on the interior mountains. The Alpine turf, bearing grasses, shrubs, and flowers, varies in thickness with the soil, and with altitude becomes less frequent.

A few species of animals have adapted well to the higher mountains. Bears have vanished, but the ibex, which like the chamois is endowed with extraordinary nimbleness, was saved by Italian royal game preserves. Marmots hibernate in underground galleries. The mountain hare and the ptarmigan—a grouse—assume a white coat for winter. Several national parks amid the ranges ensure preservation of the native fauna.

Human imprint on the Alps. From prehistoric times, the Alps have never ceased to support a large population. Throughout the centuries, they were heavily Romanized in the west and Germanicized in the east.

Traditional ways of life. A self-sufficient economic system based on agriculture lasted until the middle of the 19th century. The plains produced cereal grains, while rye, vegetables, hemp, and flax grew higher up. The high pastures were ideal for grazing herds of cattle, sheep, and goats, which provided both food and clothing. Complex pastoral usages regulated the up-and-down movements of the herds and flocks—up to the open land in the spring, down for more abundant food supplies during the winter.

Agriculture and cottage industry

Two complementary activities were necessary for economic survival: industry of a handicraft nature and seasonal migration of natives to find work elsewhere. The industry exploited local resources: charcoal and other articles were made from wood; iron and silver ores were processed and fashioned: and flax, hemp, and wool were spun and woven into clothing. The Italian chimney sweeps came to symbolize the migration at its most miserable, but other persons repaired watches or became itinerant vendors of fabrics, haberdashery, and other items. Most returned home for the summer tasks, but some left forever.

Modern intrusions. Toward the middle of the 19th century, the Alps became more heavily populated, largely because of the introduction of the potato, which supported a larger population than previous crops could. The economic revolutions in the lowlands undermined the agriculture, the cattle raising, and the cottage industries of the mountains. Railroads penetrated up the valleys and through tunnels, bringing goods from the plains and often returning from the mountains with large contingents of permanent émigrés. They also made possible tourism and other new kinds of industry. Hydroelectric power was harnessed as early as 1869 at Lancey, near Grenoble, France, to supply electrometallurgical and electrochemical industries.

The tourists were attracted both by the beauties of the Alpine landscape and by the vogue of mountaineering. Chamonix-Mont-Blanc, Zermatt, Interlaken, Sankt Moritz, and Cortina d'Ampezzo became established as vacation centres for the upper and middle classes. Chamonix-Mont-Blanc and Switzerland were frequented by the English, the eastern Alps by the Germans. From among the mountain farmers arose the first generation of guides.

The contemporary Alps. Until the middle of the 20th century, only a few centres were animated by industry and tourism. Agriculture, technically backward and deprived of its complementary craft industries, could no longer sustain a large population. The French and Italian Alps experienced a serious depopulation, though the Germans in Switzerland, Austria, and Bavaria held more tenaciously to their lands.

Roads built after World War II made accessible places that could not be reached by train. Tourism increased prodigiously, not only for the cool summer months but also for winter sports, above all skiing, which experienced a great resurgence. Although there was skiing in the 1930s, it was slow in developing because of the extensive facilities and equipment it requires. Only in the 1950s were the Alps able to offer the properly prepared slopes, the chair lifts, and the centrally heated hotels in sufficient quantity.

Tourism and skiing

Skiing has become the major attraction of the Alps, which have become accessible by air travel to enthusiasts from as far as the Americas and Japan. Summertime skiing on glaciers has created a year-round industry. Though many thousands of acres of farmland have been sacrificed to the sport, villagers in Switzerland and Austria participate in the boom by renting out their rooms and by building small hotels.

Prospects. The major result of the skiing and tourist boom may be the increasing stability of the population of the Alps regions of Europe. The manufacturing industries of Austria, Italy, and Switzerland that depend on hydroelectricity appear to be able to continue on their own, but the nationalization of the electrical industry in France appears to many persons to threaten that sector of the economy in the French Alps. Towns throughout the mountains have continued to grow, however, and their nontourist industries have continued to diversify. Grenoble, France, with more than 300,000 inhabitants, is the major Alpine city, and Innsbruck, Austria, and Bolzano, Italy, both have more than 100,000. Numerous middle-sized cities lie at the entrances to the Alps or within the range, and smaller towns and villages are numerous. In spite of their separation from major centres, they have become more or less stable economically and socially. Most important, their pure air and water, their serenity, their promise of outdoors excitement on the slopes will probably assure that they will continue to flourish amid the clamours of the 20th century.

BIBLIOGRAPHY. The literature on the Alps is immense, but it is mainly in French and German. The numerous universities in or near the Alps have accumulated an enormous quantity of documents, many of them unpublished.

A general study of the Alps is found in P. and G. VEYRET, *Au Coeur de l'Europe: Les Alpes* (1967); P. VEYRET, *Les Alpes* (1971); P. GABERT and P. GUICHONNET, *Les Alpes et les États alpins* (1965); G. DAINELLI, *Le Alpi* (1963); and P. GEORGE and J. TRICART, *L'Europe centrale*, 2 vol. (1954).

The physical geography is covered in numerous works. The theory of the nappes is set forth in L.W. COLLET, *The Structure of the Alps* (1927); a geological synthesis was written by E. KRAUS, *Die Baugeschichte der Alpen*, 2 vol. (1951); the history of the glaciations was retraced by A. PENCK and E. BRUCKNER, *Die Alpen im Eiszeitalter*, 3 vol. (1901–09); the nature of the high mountains is well-analyzed in R. GODEFROY, *La Nature alpine* (1957).

There are fewer works on the human geography: J. FRODIN, *Zentraleuropas Alpwirtschaft*, 2 vol. (1940–41); M. CEPEDE *et al.*, *La Vie rurale dans l'arc alpin* (1960; Eng. trans., *Rural Problems in the Alpine Region*, 1961); P. BRUNNER, *Les Chemins de fer aux prises avec la nature alpestre* (1935); and M.E. DELLENBACH, *La Conquête du massif alpin et de ses abords par les populations préhistoriques* (1936).

(P.V.)

Altaic Languages

The term Altaic includes the languages of the Turkic, Mongolian, and Manchu-Tungus language families. Named after the Altai Mountains, the languages are spread over an area that extends from the northeastern region of the Asian continent through the northern and northwestern provinces of China, Mongolia, Central Asia, southern Siberia, the Volga region, and Turkey, down to the Near East and the Balkan Peninsula. In the early 1970s there were more than 70,000,000 speakers of these languages.

On the basis of correspondences in vocabulary and structural similarities, several scholars have concluded that the Altaic languages are genetically related and thus use the term Altaic as the name of a language family. Others consider the correspondences and similarities to be only traces of ancient contacts or areal (regional) convergences and reject the genetic relationship of these languages. For them, the term Altaic refers to a language group that displays a certain historical and typological unity. Still others admit the possibility that the languages are related but do not consider the relation provable on the basis of present knowledge.

Question of genetic relationship

Attempts have been made to determine further genetic relations of the Altaic languages. The examination of structural and etymological correspondences gave rise to the hypothesis that the Uralic (including Finnish and Hungarian) and Altaic languages are related; this hypothetical group is called the Ural-Altaic language family. But the Ural-Altaic theory has an ever-decreasing number of supporters. Some adherents of the Altaic relationship are seriously considering the possibility of genetic ties to certain ancient strata of the Korean language.

As a result of the historically very active role of the Altaic peoples, their languages are found spread over a large geographical area. Altaic peoples overpowered their present territories in Asia and Europe in succeeding waves. They once played a dominant role in the history of China, Iran, Byzantium, the Arab caliphate, and India, and their migrations had effects on the history of eastern Europe as well. Their present political formations in Inner and Central Asia and in the Near East are of great importance.

Languages of the Altaic group. The following tables give information about the literary languages and the major dialects of the three language families, their geographical division, and the number of speakers. Historical connections of the individual languages are also indicated. The names of subdivisions of the language families are of geographical or historical origin, in conformity with scientific traditions.

Table 1: Turkic Languages*

	number of speakers
Common Turkic, or z-Turkic, languages	
Southeast (Uighur or Chagatai) group	
Uzbek (U.S.S.R. 6,015,000, Afghanistan 1,200,000†, China 15,000)	7,230,000†
Uighur (China 4,200,000, U.S.S.R. 95,000); the literary language, is also used by the Salars (China 35,000) and Yellow Uighurs (China 4,000)	4,295,000
Southwest (Oguz or Turkmen) group	39,000
Turkish (Turkey 30,569,000†, Balkan countries 1,009,000, Cyprus 98,000, U.S.S.R. 35,000, Arab countries 19,000, other countries 22,000)	31,752,000†
Azerbaijani (Iran 3,200,000, U.S.S.R. 2,940,000, Afghanistan 30,000†); other Oguz dialects spoken in Iran (near Azerbaijani)	6,170,000† 970,000
Turkmen (U.S.S.R. 1,002,000, Iran 330,000, Afghanistan 270,000†, Arab countries 122,000, Turkey 92,000†)	1,816,000†
Gagauz (U.S.S.R. 124,000, Bulgaria 5,000)	129,000
Northwest (Kipchak) group	
Karaim (U.S.S.R.)	6,000
Kumyk (U.S.S.R.)	135,000
Karachay-Balkar (U.S.S.R., Karachai 81,000, Balkar 42,000)	123,000
Tatar (U.S.S.R. 4,968,000, Romania 22,000, Turkey 13,000†, China 8,000, Bulgaria 6,000, other countries 1,000)	5,018,000†
Bashkir (U.S.S.R.)	989,000
Kazakh (U.S.S.R. 3,621,000, China 580,000, Mongolian People's Republic 40,000, Afghanistan 3,000†)	4,244,000†
Kara-Kalpak (U.S.S.R. 173,000, Afghanistan 2,000†)	175,000†
Kirgiz (U.S.S.R. 969,000, China 80,000 Afghanistan 25,000†)	1,074,000†
Nogay (U.S.S.R.)	41,000
Northeast (Siberian or Altai) group	
Khakass (U.S.S.R.); the literary language used also by the Shors (U.S.S.R.)	57,000 15,000
Altai (formerly Oirat; U.S.S.R.)	45,000
Tuvinian (U.S.S.R. 100,000, Mongolian People's Republic 18,000)	118,000
Yakut (U.S.S.R.)	237,000
Khalaj (Iran)	20,000†
r-Turkic language	
Chuvash (U.S.S.R.)	1,470,000
Total	66,168,000

*The geographical location of the language area and data specifying the distribution and number of speakers are given in parentheses. †Estimated.

Because the name Turkish usually referred to a single language—Ottoman Turkish, the idiom spoken in Turkey, the Balkan countries, Cyprus, and elsewhere—the term Turkic is preferred in scientific literature for the whole language family.

On the evidence of certain archaic features in which Chuvash corresponds to the Mongolian languages, adherents of the theory of an Altaic relationship tend to look upon Chuvash as a separate unit among the Turkic and Mongolian languages.

Mutual intelligibility is possible within each of the Turkic language groups, but there are major difficulties in comprehension among the groups, and understanding is impossible among the speakers of the Yakut, Khalaj, and Chuvash tongues and those of the so-called common Turkic languages.

Table 2: Mongolian Languages*

	number of speakers
Western group	
Kalmyk (U.S.S.R.)	106,000
Oirat (Mongolian People's Republic 71,000, China 60,000†)	131,000†
Mogol (Afghanistan)	50,000†
Eastern group	
Mongol (Mongolian People's Republic 711,000, China 1,620,000)	2,331,000
Buryat (U.S.S.R. 253,000, Mongolian People's Republic 28,000)	281,000
Daghur (China)	50,000
Monguor (China)	60,000
Santa (China)	180,000
Total	3,189,000

*The geographical location of the language area and data specifying the distribution and number of speakers are given in parentheses. †Estimated.

The name Mongolian is used for the members of the language family taken collectively, as well as for certain idioms spoken in the Mongolian People's Republic (Outer Mongolia) and China (Inner Mongolia). To make an exact distinction between the family and these languages, the latter are sometimes designated by the name Mongol. The reverse use of these terms, however, is also known.

Kalmyk, Mongol, and Buryat (Buriat) are contemporary literary Mongolian languages. There are no great differences among these languages, and mutual intelligibility thus exists, though the possibility of mutual comprehension is decreasing among speakers of isolated and non-written dialects.

Table 3: Manchu-Tungus Languages*

	number of speakers
Southern (Manchu) group	
Manchu (China)	2,800,000
Nanai (U.S.S.R. 8,000, China 1,000)	9,000
Other dialects (U.S.S.R. 4,000, China 23,000)	27,000
Northern (Tungus) group	
Evenki (U.S.S.R. 25,000, China 6,000, other countries 1,000)	32,000
Lamut, or Even (U.S.S.R.)	9,000
Total	2,877,000

*The geographical location of the language area and data specifying the distribution and number of speakers are given in parentheses.

The census datum for Manchu has an ethnic rather than a linguistic value. It is generally assumed that Manchu must be considered as the language of a minor group, and it is, in fact, almost an extinct language.

There are considerable differences between the northern and southern groups of this family and among the individual languages and dialects of the southern group. Despite their immense geographical area, languages and dialects belonging to the northern group are less differentiated.

LINGUISTIC CHARACTERISTICS

Phonology. Simplicity is the characteristic feature of the phonological (sound) systems of the Altaic languages. Richness in vowels is compensated for by the scantiness

Sound harmony

of consonants. In addition, clusters of consonants are rare. Sound harmony is one of the most characteristic features of the Altaic languages. Its essence is that in an individual word only certain combinations of vowels may occur, and, in the case of certain consonants, specific variants must be selected according to the type of vowels in the word. The best known forms of vowel harmony are palatal and labial harmony; *i.e.*, only back (velar) or front (palatal) vowels and rounded (labial) or unrounded (nonlabial) vowels may occur in a word. For example, in Turkish, back vowels include *a, ı, o, u;* front vowels include *e, i, ö, ü;* rounded vowels are represented by *ö, ü, o, u* and unrounded vowels by *a, ı, e, i.* This harmony extends to the suffixes, too, producing variants when added to words belonging to different vowel classes. The change in the vowels of the last two syllables (both suffixes) in the following Turkish words is notable: *at-lar-da* "on the horses," *it-ler-de* "on the dogs," *gör-ür-üm* "I see," *getir-ir-im* "I bring."

Sound harmony in the Altaic languages is the product of a complicated historical development. The form of vowel harmony varies from one language to another. Sound harmony is less developed in the Manchu-Tungus group, in which it may occasionally reflect an ancient state or special aberration. Adherents of the Altaic relationship maintain that the ancient state of these languages naturally also exerted a considerable influence on musical tone on the end syllable was the result of a later development. The ancient state seems to be reflected in the contemporary Mongolian tongues and a group of Manchu-Tungus languages; in Turkic and in another group of the Manchu-Tungus languages, stress shifted to the end syllable.

Grammar. From the point of view of grammatical features, there is a great difference between the Altaic languages and the Indo-European languages (which include English, French, and German). Altaic languages are agglutinative; that is, word formation consists of adding suffixes to the word roots. The root may take a large number of suffixes (*e.g.*, Turkish *gel-me-dik-ler-i-nden* "because of their not coming"), though the order of the combination of suffixes is governed by strict rules. While certain suffixes in the Indo-European languages may perform several functions (*e.g.*, Latin *casarum* "of the houses," in which the suffix *-arum* indicates a feminine, plural, genitive form), in the Altaic languages these functions are distributed among separate suffixes (*e.g.*, Turkish *ev-ler-in* "of the houses," which consists of a stem plus a plural suffix plus a genitive suffix).

Altaic languages lack definite articles (such as English "the") and grammatical gender (*e.g.*, German masculine, feminine, and neuter words), and use the singular form of nouns with numerals (*e.g.*, "two girl"). In the Manchu-Tungus languages, a distinction is made in the 1st person plural of the personal pronoun between exclusive ("we"—without you) and inclusive ("we"—with you). The same peculiarity is found in Mongolian and, from the Mongolian influence, in the Turkic languages of Siberia. There are no prepositions in the Altaic languages, but postpositions are used very frequently. For example, an Altaic speaker does not say the equivalent of "until night" or "for John" but rather "night until" and "John for." In the Tungus languages, nominal (noun) inflection has approximately 20 cases, while the Manchu languages show a simpler state, with about five to eight cases. Turko-Mongolian languages also show a simple case system: they have about six to ten cases. Because the plural is expressed by independent suffixes, the case suffixes are identical in singular and plural.

Parts of speech in the Altaic languages

Parts of speech in the Altaic languages are less differentiated than they are in the Indo-European languages; essentially, only nominals (words used as nouns), verbs, and particles (words that are neither nouns nor verbs) may be considered as separate parts of speech. In contrast with the Tungus languages, the Turkic and Mongolian languages do not possess any morphological means, such as specific suffixes, to distinguish between nouns and adjectives. Altaic languages are rich in verbal nominals (nouns, adjectives, and adverbs derived from verbs), as well as in nouns of action, participles, and gerunds; these perform important grammatical functions in Altaic sentences. A conspicuous ancient Altaic peculiarity is that verbal predicative expressions (finite verbs) are of verbal-nominal origin; that is, these verbal formations are in most cases verbal nouns in predicate positions and express their personal relations with added personal or possessive pronouns: *e.g.*, Old Turkic *bil-t-im* "I knew" is literally "having known my," and Old Turkic *bil-ir-biz* "we know" is literally "knowing we." In Altaic languages, there is no word that corresponds to the English "to have." To express this concept, a periphrastic construction (circumlocution) is used; *e.g.*, an Altaic speaker might say "The book is with me" rather than "I *have* the book."

Despite major differences in syntax, there are also significant common Altaic features. The modifier (attribute, appositive), for example, always precedes the modified word (*e.g.*, a noun), and the noun in the genitive case (the possessive) precedes the possessor (*e.g.*, "captain of the ship" is "of the ship, captain"). In the Tungus languages the attribute agrees in number and case with the modified noun. This feature is completely lacking in the Turkic languages, though agreement in number existed in the early periods of Mongolian.

In Altaic sentence structure, subordinate clauses that begin with conjunctions or relative pronouns ("The man *who was bald*") are alien; wherever they appear, they indicate foreign influence. Instead of subordination, coordination by verbal nominals (nouns) serves as a basis for sentence construction. The following Turkish sentences illustrate this feature. *Ahmedin bize gelmesini isterim*, translated as "I want Ahmed to come to us," is literally "I want Ahmed's coming to us"; *Şehre giden otobüs gecikti* "The bus, which goes to the town, was late" is literally "The bus going to the town was late"; *Eve varınca hemen telefon etti* "As soon as he arrived home he telephoned" is literally "Arriving home, he immediately telephoned." As a result of a more varied system of morphology, the Manchu-Tungus languages are characterized by a higher degree of syntactic freedom, while the Turkic and Mongolian languages have a higher degree of syntactic constraints (requirements for the positions of the elements of a sentence) and a simpler system of morphology.

Altaic languages have changed at a slow rate. The archaic character is particularly conspicuous in the Manchu-Tungus and Mongolian languages. Because of this difference in historical character, one language may provide useful information for the history of another in regard to the development of earlier, supposedly common features.

Vocabulary. During an immense span of space and time, the vocabulary of the Altaic languages has shown a high degree of receptivity. Opponents of the genetic Altaic relationship look upon the ancient common elements of the vocabulary of these languages as loanwords resulting from once-frequent contacts between Turkic and Mongolian and between Turko-Mongolian and Manchu-Tungus. These correspondences in vocabulary show closer ties between the Turkic and Mongolian languages than between the Turko-Mongolian and Manchu-Tungus languages. The oldest loanwords of the vocabulary are connected with the Semitic, Indo-European, and Uralic languages. In the eastern part of the Altaic language area, many loanwords came from the Chinese to the Old Turkic and Middle Mongolian languages, as well as to contemporary Turkic and Mongolian, and also to the Manchu-Tungus languages, particularly to Manchu. Through the medium of religious texts translated from Sanskrit; Tocharian, and Sogdian, many loanwords found their way to Uighur and from there to Mongolian. Mongolian also had contacts with the Tibetan language through the Buddhist literature and through certain dialects. A strong linguistic interaction may be observed among the local Turkic and Mongolian languages in Mongolia and southern Siberia and among the Turkic and Iranian languages in Central Asia.

Oldest loanwords from Semitic, Indo-European, Uralic

A number of Mongolian words were adopted into the

Yakut language, which in turn had contact with the Tungus languages. The spread of Islām paved the way for Arabic and Persian loanwords to enter the Turkic languages. Because of relations among the Ottoman Empire, Turkey, and the Western civilizations, a considerable number of French, Italian, German, and English words were borrowed into Turkish. Russian provided the major part of the vocabulary of modern civilization for the Turkic languages of the U.S.S.R., although there was also considerable influence from other spheres. The Balkan, Anatolian, Caucasian, and certain Uralic languages had only minor influence on the Turkic languages.

In certain Turkic languages, linguistic contacts go far beyond the simple borrowing of words. Although in the East the development of the Buddhist translation literature was followed by a comparatively modest linguistic influence by the source languages, the Arabic and Persian influence that followed the adoption of Islāmic culture proved very strong. An avalanche of Arabic and Persian words flooded the vocabulary of the literary languages of the settled Islāmic-Turkic peoples and also filtered into the dialects. This layer of loanwords also led to changes in the structure (*i.e.*, the phonological system and sound harmony) of certain languages. Iranian influence gave rise to similar profound changes in Uzbek. Currently, Russian influence has been greatly intensified by the continuous contact taking place within an identical political and cultural framework and also by the increasingly widespread bilingualism of the Turkic peoples. This process is reflected in the growing number of calques (loan translation words), as well as in syntactic (sentence) structure.

In the course of linguistic contacts, the Altaic languages naturally also exerted a considerable influence on other languages; *e.g.*, Persian, Russian, Finno-Ugric, the Balkan languages.

Earliest Turkic records

Writing systems. The Altaic peoples used a variety of writing systems. The first-known writing of the Turkic peoples is the runic script (named after the Germanic runes), which is most probably of Semitic origin, particularly Aramaic. The first dated texts originate from Inner Asia in the 8th century; they are called the Orkhon inscriptions. Archaeological investigations indicate that this early script was spread over an extremely large area. The Turkic peoples (mainly Uighurs) who settled in the Tarim Basin used Brāhmī and Manichaean scripts in the 8th to 10th century, as well as the Uighur writing, as it developed from the Sogdian cursive writing, from the 8th to 17th century. Nestorian (Christian) Turkic peoples in Central Asia used the Estrangelo script of Syriac origin in the 13th and 14th centuries.

After the adoption of Islām in the 11th century, the Arabic script became the most widespread writing system among the Turkic peoples. In Anatolia, certain linguistic groups or minorities used the Greek alphabet (16th–20th centuries) and the Armenian script (17th–20th centuries). The latter was also employed by the Turkified Armenians in the Ukraine in the 16th and 17th centuries. Between the 16th and 19th centuries, Karaites used the Hebrew script; Latin and Cyrillic writing was introduced to them in the 19th century. In 1929 the Arabic script was replaced in Turkey by Latin writing. Among the Turkic peoples of the U.S.S.R., the Latin script was introduced in the 1920s; in the 1930s it was replaced by the Cyrillic script. Turks living in China, Iran, and the Arab countries still use the Arabic script.

Mongolian script developed from Uighur writing in the 12th century. After some Latin antecedents, it was replaced by the Cyrillic script among the Buryats in 1938 and among the Mongols in 1944. Mongols living in China still use the Mongolian script. The 'Phags-pa or Pa-sse-pa alphabet, introduced on the state's initiative in 1269–72, was based on the Tibetan script and did not take root. It became extinct in a few decades. From the 17th century, the Oirats used a reformed version of the Mongolian writing, called Oirat script. After 1917 this writing system was replaced among the Kalmyks by the Cyrillic alphabet; the Latin alphabet was introduced in 1931, and in 1937 the Cyrillic script was reintroduced. Oirats living in China still use the Oirat script.

The Manchu script, a reformed or adapted version of Mongolian writing, came into existence in the 17th century; it is probably still used in private life in China. For Manchu-Tungus people living in the U.S.S.R., the Cyrillic script was introduced in 1937–38 after some experiment with the Latin alphabet.

The writings of the Mongolian-speaking Khitans, who ruled in China from the 10th to the 12th century, and those of the Juchens, who ruled in China in the 12th and 13th centuries and who spoke a Manchu-Tungus language, have not yet been definitely deciphered.

HISTORY OF THE ALTAIC LANGUAGES

Original habitat of the Altaic peoples

The original habitat of the Altaic peoples is supposed to have been on the steppe area bordering on the Altai and Ch'ing-hai mountains between Tibet and China; in the north, it probably stretched to the Siberian taiga region adjacent to the steppe. It is assumed that Tungus peoples lived in the northern and northeastern parts of this area and the Mongols in the central and southeastern parts. Turkic peoples probably inhabited the northwestern and western parts, while the southern and southwestern region was occupied by Hunnic groups. Although the Hun language is not known, Huns are supposed to have spoken an Altaic idiom. In historical times, the Uralic, Indo-Iranian, Sino-Tibetan, and Korean languages have bordered on this Altaic language area.

Advocates of a genetic relationship between the Altaic languages believe that it was probably in the area delineated above that the Proto-Altaic language (*i.e.*, the parent language of the modern Altaic tongues) began to divide into the currently known main groups. Those who oppose the Altaic theory maintain that this area was the scene of contacts among the individual language groups. From this original habitat, the Altaic peoples moved to their present territories. In the steppe migrations, which were natural consequences of a nomadic way of life and which dramatically changed the ethnic and linguistic map of central Eurasia, the Turkic and Mongolian peoples played the leading role. The Manchu-Tungus peoples' separate migration toward the northeastern Asian taiga and tundra might have taken place in an early period. The south- and southwest-bound migrations of the southern Manchu-Tungus peoples were restricted to the East Asian regions.

The Turkic languages. *Division into two linguistic groups.* It was probably at a very early date that the Turkic languages split into the *z* and *š* group (commonly called *z*-Turkic) and *r* and *l* group (commonly called *r*-Turkic); *e.g.*, *z*-Turkic *yüz*, "hundred," contrasts with *r*-Turkic *śĕr* "hundred"; *š*-Turkic *yaš*, "age," contrasts with *l*-Turkic *śul* "age." The first traces of the *r* and *l* division appeared on the south Russian steppes. It is assumed that the Huns also were speakers of an *r*- and *l*-type Turkic language and that their migration was responsible for the appearance of this language in the West. The *r*- and *l*-type language is now documented only by Chuvash, a language considered as a descendant of a Volga-Bulgarian language. The rest of the Turkic languages are of the *z*- and *š*-type.

The split of the *z*- and *š*-type Turkic languages, which appeared first in Inner Asia and later developed in the presently known language groups and languages, evolved over several centuries.

Literary languages and linguistic records. As was stated above, the first records of the Turkic languages are runic inscriptions of the Turks in Mongolia. The Uighurs overthrew the Turkish Empire there, and, after the disintegration of their empire in 840, they settled down in the Tarim Basin, where the Manichaean and Buddhist religions gave rise to a sizable literature written in various scripts, flourishing from the 9th to the 14th century. In the 11th century, during the rule of the Qarakhanides (the first Turkic representatives of Islām), a new literature came into existence in eastern Turkistan. These literary efforts show the intensification

Arabic and Persian influence on Turkic literature

of Arabic and Persian influence, which culminated in the Khwārezmian literature (13th–14th centuries) and the Chagatai literature (15th–16th centuries), as well as in the postclassical products of Chagatai (17th–19th centuries). (Khwārezmian and Chagatai are Middle Turkic languages with significant literatures.) Despite the influence exerted by other Turkic language groups, these Inner and Central Asian literary languages show an organic continuity; they might be considered as the antecedents of the contemporary Uzbek and New Uighur literary languages.

The main movement of the Oğuz branch of Turkic peoples from western Turkistan was toward the Caucasus, Iran, Anatolia, and the Balkans. The first continuous records of this language group come from 13th-century Anatolia. Its linguistic material is linked with the contemporary literary language of Turkey (*i.e.*, Turkish). The Azerbaijani and Turkmen languages also have long literary traditions. The first records of the Azerbaijani language date from the 14th century; the Turkmen language came into use only gradually, because the Chagatai language was commonly used instead by writers.

The migration of the Kipchak Turks was toward the south Russian steppes. A 14th-century textbook known as the *Codex Cumanicus*, used by missionaries, constitutes the first record of the languages of these peoples. Linguistic records (*e.g.*, dictionaries, grammars, literary works) of the Kipchaks, who as mercenaries went to Egypt and Syria where they also came into power, date from the 13th to 17th century. Armenian Turks of the Ukraine left remarkable written material from the 16th and 17th centuries. All these Kipchak linguistic records, although they represent languages that have disappeared, are still connected indirectly with the contemporary Kipchak literary languages. Because of the cultural conditions of these peoples, the first records of most of these languages originated in the 20th century.

The remainder of the contemporary Turkic literary languages also developed in the 20th century.

The diffusion of the Turkic peoples over the centuries and the changes in their political and cultural centres were accompanied by the emergence of a number of literary languages. These literary languages are not necessarily the direct antecedents of the contemporary—and recently developed—literary languages. In most cases, only their membership in the same language group (Uighur, Kipchak, Oğuz) may be established, thus indicating an indirect historical connection.

Stages of the languages. On the basis of linguistic records, Turkic languages spoken from the 8th to the 10th century are termed Old Turkic, those of the 11th to 15th century Middle Turkic, and those of the 16th to 20th century New Turkic languages. To specify the form of the Turkic languages that resulted from language reforms of the mid-20th century, it is customary to use the term newest, or modern. The division into periods is based on external facts of these languages rather than the internal history and thus merely gives a practical chronological overview.

The Mongolian languages. Although the Mongolian script probably evolved as early as the 12th century, its first record, the inscription of Yesunke, dates from 1225. The *Secret History of the Mongols*, written in all probability in 1240, has been handed down in Chinese transcription. As regards the development of spoken Mongolian, the history of Mongolian is divided into three periods: Old Mongolian (to the 12th century), Middle Mongolian (13th–16th centuries), and New Mongolian (17th–20th centuries). Written Mongolian, originally based on a Middle Mongolian dialect, was characterized for centuries by an increasing conservativism in regard to its relation to the spoken language. Its history is divided into such periods as preclassical (13th–16th centuries), classical (17th–20th centuries), and modern (20th century). Classical Mongolian is continued by the contemporary literary Mongol, based on the Khalkha Mongol dialect, and by the written language of Mongols living in China.

Stages of Mongolian

The Oirats created a separate script and literary language in the 17th century. This, however, decreased in importance among the eastern Oirats from the second half of the 18th century but survived among the western Oirats (the Kalmyks); its course of development flowed into the contemporary literary language.

It was not until 1931 that the Buryats began to develop their own literary language. Until then, literate groups had used literary Mongolian.

The Manchu-Tungus languages. Within the Manchu-Tungus language family, only the history of the Manchu language is documented by linguistic records. As a result of the Manchu rule in China (1644–1911), Manchu rose to the rank of official language in the country. The earliest Manchu records date from the mid-17th century. Manchu acted as the lingua franca in China's communications with other countries. The downfall of the Manchu rule in China and the gradual Sinicization of the Manchu population ended the influence of the Manchu language. Other literary languages of the Manchu-Tungus language family came to be in the 1930s.

As has been shown, the Altaic languages, unlike the Indo-European languages, have very recent linguistic records—Turkic from the 8th century, Mongolian from the 13th century, and Manchu-Tungus from the 17th century. This situation puts limits on comparative historical investigations, making it extremely difficult to settle the problem of the Altaic relationship.

Altaic languages in the 20th century. The development of literary languages became a central problem in the independent political entities of the Altaic peoples that arose after World War I. The languages of the Altaic peoples in the U.S.S.R. underwent a veritable revolution in the 1920s and 1930s. Writing reforms were soon followed by the strengthening of literary languages on the basis of the individual national languages. For several illiterate peoples, new literary languages had to be created for the first time. In Turkey, written and spoken language reforms included, above all, a renewal of the vocabulary; this formed a significant objective of the Turkish president Kemal Atatürk's social program. In the ongoing process of modernization, a new literary language has come into existence. Considerable reform of writing and language also took place in Mongolia.

Creation of new literary languages

The linguistic policy pursued in the mid-20th century had an immense impact on the history of the Altaic languages. These efforts not only saved minor or nonwritten languages from extinction but strengthened the positions of major Altaic languages among the languages of modern civilizations. The linguistic consolidation has considerably contributed to enhancing the political and cultural significance of these peoples.

For information on the literature of the peoples who speak (or have spoken) Altaic languages, see CENTRAL ASIAN PEOPLES, ARTS OF; ISLAMIC PEOPLES, ARTS OF.

BIBLIOGRAPHY. N. POPPE, *Introduction to Altaic Linguistics* (1965), is the best up-to-date manual, giving a short but complete picture of the Altaic languages, their history and structure, as well as the theory of the relationship of these languages. *Philologiae Turcicae Fundamenta* (1959); *Handbuch der Orientalistik:* vol. 5, pt. 1, *Turkologie* (1963); pt. 2, *Mongolistik* (1964); and pt. 3, *Tungusologie* (1968), give systematic descriptions of the various historical and contemporary languages with much bibliographic data. K.H. MENGES, *The Turkic Languages and Peoples* (1968), is the best recent account of Turkic linguistics, supplementing the bibliographies in the works mentioned above. G.J. RAMSTEDT, *Einführung in die altaische Sprachwissenschaft*, vol. 1, *Lautlehre* (1957), and vol. 2, *Formenlehre* (1952), is the classical summary of the theory of relationship among the Altaic languages. R. LOEWENTHAL, *The Turkic Languages and Literatures of Central Asia* (1956); D. SINOR, *Introduction à l'étude de l'Eurasie Centrale* (1963); and G. HAZAI (ed.), *Sovietico-Turcica* (1960), are useful bibliographies that complement one another.

(G.Ha.)

Altai Mountains

General description

A great topographic spine of Inner Asia, the Altai straddle parts of China, Mongolia, and the Soviet Union,

stretching more than 1,000 miles from the Gobi (desert) to the West Siberian Plain. The jagged mountain ridges, whose name derives from the Turkish–Mongolian *altan*, meaning "golden," separate the waters of such great rivers as the Ob (flowing north to the Arctic Ocean) and its major tributary, the Irtysh, from the rivers draining into the vast Central Asian basin. There are three main prongs of the system—the Soviet, Mongol, and Gobi Altai; and the Soviet peak, Belukha, at 14,783 feet (4,506 metres) is the highest point. Historically something of a remote barrier, the Altai are now famous for their mining and hydroelectric potential, while the ancient ways of life of their peoples are being rapidly transformed.

Topography. The Soviet Altai lies in eastern Kazakhstan and the Gorno-Altai sector of the Russian Soviet Federated Socialist Republic. A belt of northern foothills separates it from the rolling West Siberian Plain, while in the northeast it borders the Zapadny (Western) Sayan uplands and the mountains of the Lebed River Basin. The Mongol Altai (rising to Hüyten Orgil, over 14,000 feet) thrusts away to the southeast, then eastward. The Gobi Altai begins some 300 miles southwest of Ulaanbaatar, the Mongolian capital, and occupies the country's southern portions, towering over the Gobi expanses.

Geological evolution. The Altai were originally formed in the great geological upthrusts of the Hercynian orogeny, about 300,000,000 years ago, and were worn down, over the long periods of geological time, into a peneplain, a plateau with generally accordant summit heights. At the beginning of the Quaternary Period, about 2,500,000 years ago, new upheavals thrust up magnificent peaks of considerable size along a zone of weakness in the Earth's surface. Quaternary glaciation scoured the mountains into rugged shapes and deepened valleys from a V- to a U-shaped cross section; river action was also intensive. As a result of these differential geological forces, the highest ridges in the contemporary Altai—notably the Katun, Severo- (North) Chu, and the Yuzhno- (South) Chu—tower more than 13,000 feet, running latitudinally in the central and eastern portions of the Soviet sector of the system. The Tabyn-Bogdo-Ola (Tavan Bogd Uul in Mongolian), the Mönh Hayrhan Uul, and other western ridges of the Mongol Altai are somewhat lower. The highest peaks are much steeper and rockier than their Alpine equivalents, but the ranges and massifs of the middle Altai, to the north and west, have ridges of about 8,200 feet, whose softer outlines betray their origins in ancient, smoothed surfaces. Valleys are nevertheless jagged and gorgelike. The ridges are separated by steppelike structural hollows (notably the Chu, Kuray, Uymen, and Kansk), which are filled with crumbly deposits up to 50,000,000 years old. Heights range from 1,600 to 6,600 feet above sea level.

The extreme dislocations suffered by the Altai over the course of geological time have occasioned a variety of rock types, many of them altered by volcanic activity when great former geosynclines, sediment-filled structural downwarps, were thrust upward in mountain-building epochs. There are considerable accumulations in the various structural depressions. The layers of nonferrous and rare metal-bearing rocks contain exploitable deposits of iron, mercury, gold, manganese, and tungsten that have a very high commercial value.

Climate. The regional climate is severely continental: because of the influence of the great Asiatic anticyclone, or high-pressure complex, the winter is long and bitterly cold. January temperatures range from 7° F (−14° C) in the foothills to −26° F (−32° C) in the sheltered hollows of the east, while in the Chu steppes temperatures can plunge to a bitter −76° F (−60° C). There are occasional tracts of the permanently frozen soil (or permafrost) that coats great stretches of the Soviet Far North. July temperatures are warm (often to 75° F [24° C], sometimes up to 104° F [40° C]) on the lower slopes, but summers are short and cool in the higher portions. In the exposed west, particularly at heights of between 5,000 and 6,500 feet, precipitation is high: as much as about 80 inches may fall quite evenly throughout the year. The amount decreases by a third farther east, and some areas have no snow at all. Glaciers coat the flanks of the highest peaks: some 1,200 in number, they cover an area approaching 600 square miles.

Temperature ranges

Rivers and lakes. The Soviet and Mongol portions of the Altai are criss-crossed by a network of turbulent, youthful rivers, fed mainly by melted snow and summer rains, occasioning spring and summer floods. The Katun, Bukhtarma, and Biya are among the biggest. Rivers of the Gobi Altai are shorter, shallower, and often frozen in winter and dry in summer. There are more than 3,500 lakes, most of structural or glacial origin. Those of the Gobi Altai are often bitterly salty.

Vegetation. There are four fairly distinct vegetation zones: mountain subdesert, mountain steppe, mountain forest, and the Alpine regions. The first, found on lower slopes and in hollows of the Mongol and Gobi Altai, reflects the high summer temperatures and low rainfall: the sparse life includes wormwood and halophytic plants (those growing in salty soils). The mountain steppe zone rises to about 2,000 feet in the north and to 6,600 feet in the south and east. Meadows and mixed grass steppes characterize the former, and grassy herbs of the wormwood-fescue type the latter: only in the higher areas are mountain vines and bush steppes found. The mountain forest zone is most characteristic of the Soviet Altai, where it covers about 70 percent of the territory, mostly in the low and medium mountain regions. Forests reach up to 6,500 feet but climb to 8,500 feet on less exposed slopes of the central and eastern Altai. Trees include coniferous growths, birches, aspens, and larches, and there are many huge clearings. A forest belt is practically nonexistent in the Mongol and Gobi Altai, but isolated clumps of Siberian trees grow in sheltered spots. Alpine vegetation—sub-Alpine shrubs giving way to meadows widely used for summer pasture and then mosses and bare rock and ice—is found only on the highest ridges.

Animal life. Animal life follows vegetation patterns. Various small rodents populate the mountainous semideserts and steppes, while bird life includes eagles, hawks, and kestrels. Most birds and animals are of Mongolian origin: marmot, jerboa (a jumping rodent), and antelope. Siberian animals (bear, lynx, musk deer) and birds (grouse, woodpecker, and crossbill) frequent the moist coniferous forests. Alpine animal life includes the mountain goat, snow leopard, and mountain ram.

The human imprint. The Soviet Altai is peopled by groups of Altaic, Russian, and Kazakh origin. Only in the 1960s did the formerly tribal Altaic peoples—distinguished on the basis of dialect—cohere into a unified group. Their principal occupation is the breeding of cattle or sheep. The Russians are mainly agriculturalists or, significantly, miners. The Kazakhs mostly raise sheep, cattle, and wild horses, utilizing the summer Alpine pastures (and erecting temporary dwellings), while keeping their stock under shelter in winter. The Mongol Altai harbours groups of Kazakhs, Mongolikhalks, and the nomadic Oyrat. Sheep, goats, and camels are raised in the drier south, with large horned cattle, as well, in the north. Horse breeding is widespread. The shortages of water in the south necessitate migratory movements of up to 120 miles on the part of the nomads, who erect temporary villages. The development of agriculture, however, is producing more settled ways of life.

Peoples and their ways of life

Research. Scientific research into the Altai dates only from the 18th century but was stimulated when gold was discovered in 1828. Russian geologists and engineers pioneered the collection of data. Soviet expeditions from the U.S.S.R. Ministry of Geology and Tomsk V.V. Kuibyshev State University and expeditions from the Academy of Sciences of the Mongolian People's Republic have uncovered a wealth of geological and topographical data, while many of the hitherto inaccessible peaks have been scaled.

The outlook. The Altai have considerable potential for continued resource and hydroelectric development. The hydroelectric stations at Ust-Kamenogorsk and Ust-Bukhtarma in the Irtysh Basin are already in operation, while agriculture is spreading into the mountains. Pas-

tures continue to support livestock, but mineral springs are attracting an increasing number of health resorts and sanatoriums. The Telets Lake area is but one of the regions attracting tourists and mountain climbers.

(N.I.M.)

Aluminum Products and Production

Aluminum, or aluminium, is the most modern of the common metals used industrially in large quantities. Its name is derived from *alumen*, the Latin name of a naturally occurring aluminous sulfate. Although crude salts of aluminum were used as early as the 5th century BC, the metal was not isolated until 1825 and was not produced commercially on a large scale until 1888.

In less than a century of commercial use, aluminum by any measure has surpassed all other metals except iron and steel. The industry's growth has been rapid in Europe and spectacular in the U.S., where its average annual rate of growth since 1893 has been 14.8 percent, and where apparent annual consumption is approximately 49.3 pounds per person. Other developed countries also manifest a large demand for aluminum—and that of the underdeveloped countries is growing rapidly.

HISTORY

Before 5000 BC, people in Iraq were making fine pottery from a clay that consisted largely of a compound containing aluminum, and almost 4,000 years ago, Egyptians and Babylonians used aluminum compounds in various chemicals and medicines. Pliny refers to *alumen*, known now as alum, a compound of aluminum widely used in the ancient and medieval world to fix dyes in textiles. By the 18th century, the earthy base "alumina" was recognized as the potential source of a metal.

The English chemist Sir Humphry Davy in 1807 attempted to extract the metal. Though unsuccessful, he satisfied himself that alumina had a metallic base which he named "alumium" and later changed to "aluminum." The name has been retained in the U.S. but modified to "aluminium" in many other countries.

First laboratory production of the metal

A Danish physicist and chemist, Hans Christian Ørsted in 1825 finally produced aluminum. "It forms," Ørsted reported, "a lump of metal which in color and luster somewhat resembles tin."

A few years later (1845) Friedrich Wöhler, a German physicist working at the University of Göttingen, made metallic aluminum in particles as large as pinheads, and became the first man to determine the following of aluminum's properties: specific gravity, ductility, colour, stability in air, etc.

Deville process. Nine more years passed before a French scientist, Henri Sainte-Claire Deville, announced a momentous improvement in Wöhler's method, which permitted Wöhler's "pinheads" to coalesce into lumps the size of marbles. Deville's process became the foundation of the aluminum industry. Bars of aluminum, made at Javel Chemical Works and exhibited in 1855 at the Paris Exposition, introduced the new metal to the public.

Although enough was by then known about the properties of aluminum to indicate a promising future, the cost of the chemical process for producing the metal was too high to permit widespread use. But important improvements presently brought breakthroughs on two fronts: first, the Deville process was improved; and second, the development of the dynamo made available a large power source for electrolysis (the production of chemical change by electric current), which proved highly successful in separating the metal from the compound sodium aluminum chloride.

The work of Hall and Héroult. The modern electrolytic method of producing aluminum was discovered almost simultaneously, and completely independently, by Charles M. Hall of the United States and Paul-Louis-Toussaint Héroult of France in 1886. (By an odd coincidence, both men were born in 1863 and both died in 1914). The essentials of the Hall-Héroult processes were identical and remain the basis for today's aluminum industry. Purified alumina is dissolved in molten cryolite and electrolyzed (broken down) with direct current. Under the influence of the current, the oxygen of the alumina is deposited on the anode and is released as carbon dioxide, while free molten aluminum—which is heavier than the electrolyte—is deposited on the carbon lining at the bottom of the cell.

Hall immediately recognized the value of his discovery. He applied July 9, 1886, for a U.S. patent and worked energetically at developing the process. Héroult, on the other hand, although he applied for French patents, apparently failed to grasp the significance of the process. He continued work on a second successful process that used copper as a cathode and produced an aluminum–bronze alloy. Conveniently, in 1888, a German chemist, Karl Joseph Bayer, discovered an improved method for making pure alumina from low-silica bauxite ores.

Hall and a group of businessmen established the Pittsburgh Reduction Company in 1888 in Pittsburgh. The first ingot produced by electrolytic smelting was poured in the fall of that year. Demand for aluminum grew and a larger reduction plant was built at New Kensington, Pennsylvania, using steam-generated electricity to produce one ton of aluminum per day by 1894. The need for cheap, plentiful hydroelectric power led the young company to Niagara Falls, where, in 1895, it became the first customer for the new Niagara Falls power development.

In a short time, the demand for aluminum exceeded Hall's most optimistic expectations. In 1907 the company changed its name to Aluminum Company of America (Alcoa). Until World War II it remained the sole American producer of primary aluminum.

European industry. Neuhausen, Switzerland, is the "nursery" of the European aluminum industry. There, to take advantage of waterpower available from the falls of the Rhine, Héroult built his first aluminum–bronze production facility, which later became the Aluminium-Industrie-Aktien-Gesellschaft. The British Aluminium Company Limited, organized in 1894, soon recognized the wealth of cheap electric power available in Norway and became instrumental in building aluminum works at Stongfjorden in 1907 and later at Vigeland. In France the Société Électrométallurgique Française, also based on Héroult's patent, was started near Grenoble around 1888. Little aluminum was produced in Germany before 1914, but the war brought an urgent demand, and several smelters went into production employing electricity generated by steam power. Later, the Soviet Union began producing substantial amounts of aluminum in the Ural industrial complex. By 1940, Japan, Sweden, and Yugoslavia had become producers.

ORES AND MINING

Aluminum is the third most abundant element on the earth's surface. Only oxygen and silicon are more common. The earth's crust to a depth of ten miles contains an estimated 8.05 percent of aluminum. Aluminum has a strong tendency to combine with other common elements and so never occurs in nature in the metallic form. Its compounds, however, are an important constituent of virtually all common rocks. It is found in clay, shale, slate, schist, granite, syenite, and anorthosite.

Aluminum is found as hydrated aluminum oxide—a combination of alumina and water—in combination with such impurities as iron oxide, titania, and silica.

Bauxite. The most important aluminum ore, an iron-containing rock consisting of about 52 percent aluminum oxide, was discovered in 1821 near Les Baux in southern France. The material was later named bauxite. Though now applied loosely to any aluminum-bearing substance, bauxite is best defined as an aluminum ore of varying degrees of purity in which aluminum in the form of hydrated oxide is the largest single constituent.

Bauxite varies greatly in physical appearance, depending on its composition and impurities. It ranges in colour from yellowish white to gray, or pink, to dark red or brown, if high in iron oxides. It may be earthy, or range in form from clay to rock. Bauxite has been found in all the world's continents except Antarctica. The richest deposits generally lie in areas that, during formation, were

in tropical and subtropical climates, providing optimum conditions of heavy rainfall, constant warm temperatures, and good drainage.

Principal deposits of bauxite

Large deposits are found in the United States, the Caribbean islands, northern South America, Hawaii, Australia, India, Indonesia, Malaysia, China, the Soviet Union, Ghana, Greece, Yugoslavia, Hungary, Italy, and France.

Not all bauxites are economical for aluminum production. Only earth with an aluminum oxide content of 45 percent or more is considered practical for making metal. Two minerals, gibbsite and boehmite, which contain 40–60 percent alumina, are presently considered economical enough to be processed. Gibbsite is found largely in tropical areas on either side of the Equator, while boehmite is found largely north of the subtropical belt in southern Europe, the Soviet Union, Turkey, and Mainland China.

Known deposits of bauxite can supply the world with aluminum for hundreds of years at present production levels. When high-grade bauxite deposits are depleted, substantial reserves of secondary ores will remain to be exploited: laterite deposits in the northwestern United States, Hawaii and Australia; anorthosite in the western United States; apatite and diaspore in Europe; kaolinite in the southeastern United States. Other nonbauxite sources of alumina are also available: alumina clays, dawsonite, alunite, aluminous shales, igneous rocks, and saprolite and sillimanite minerals. Eastern Europe, led by the Soviet Union and Poland, already is refining alumina from nonbauxitic ores, namely nepheline syenite and high-alumina clays. Vast new bauxite developments in Australia, Guinea, and Indonesia have tended to postpone interest in secondary ore technology elsewhere.

Mining. By far the greatest quantity of commercially exploited bauxite lies at or near the earth's surface. Consequently, it is mined in open pits requiring only a minimum removal of layers of soil and rock covering the ore. Bauxite beds are blasted loose, dug up with power shovel or dragline, and the ore transported by truck or rail to a processing plant, where it is crushed for easier handling, washed of clay and dirt, kiln-dried, pulverized, and shipped to a defining plant. Refining plants are located near mine sites, it possible, since transportation is the major item in bauxite costs.

Approximately 90 percent of all bauxite produced is refined into alumina, which is ultimately smelted into aluminum. The remaining 10 percent is used in other applications, such as abrasives and refractories, and in the manufacture of petroleum catalysts and paper. Bauxite yields one ton of aluminum from four to six tons of ore. World production of bauxite was more than 50,000,000 tons per year at the end of 1969. Jamaica is the largest single producer, followed by Australia, Surinam, the Soviet Union, Guyana, France, Guinea, Yugoslavia, Greece, Hungary, and the United States.

ORE PROCESSING

The production of aluminum from bauxite is a two-step process: refining bauxite to obtain alumina and smelting alumina to produce aluminum. Bauxite contains a number of impurities, including iron oxide, silica, titania, and water. If these impurities are not removed during refining, they will alloy with and contaminate the metal during the smelting process. The ore of aluminum, therefore, must be treated to eliminate these impurities. Purified alumina suitable for commercial production usually contains less than 0.1 percent of other oxides. Bayer's process, with some modification, is still the most widely used and commercially successful method for the production of alumina, and virtually all aluminum is smelted from alumina using the Hall-Héroult electrolytic process.

Refining the ore. There are a number of alkaline, acid, and thermal methods of refining bauxite, clay, or other ores to obtain alumina. Acid and electrothermal processes generally are either too expensive or do not produce alumina of sufficient purity for commercial use. A process that involves treatment of ore with lime and soda is used in the Soviet Union. Smelting is accomplished with the aid of limestone in Norway; this process may also have been used in the Soviet Union during World War II.

The Bayer process

The Bayer process (Figure 1) involves four steps: digestion, clarification, precipitation, and calcination.

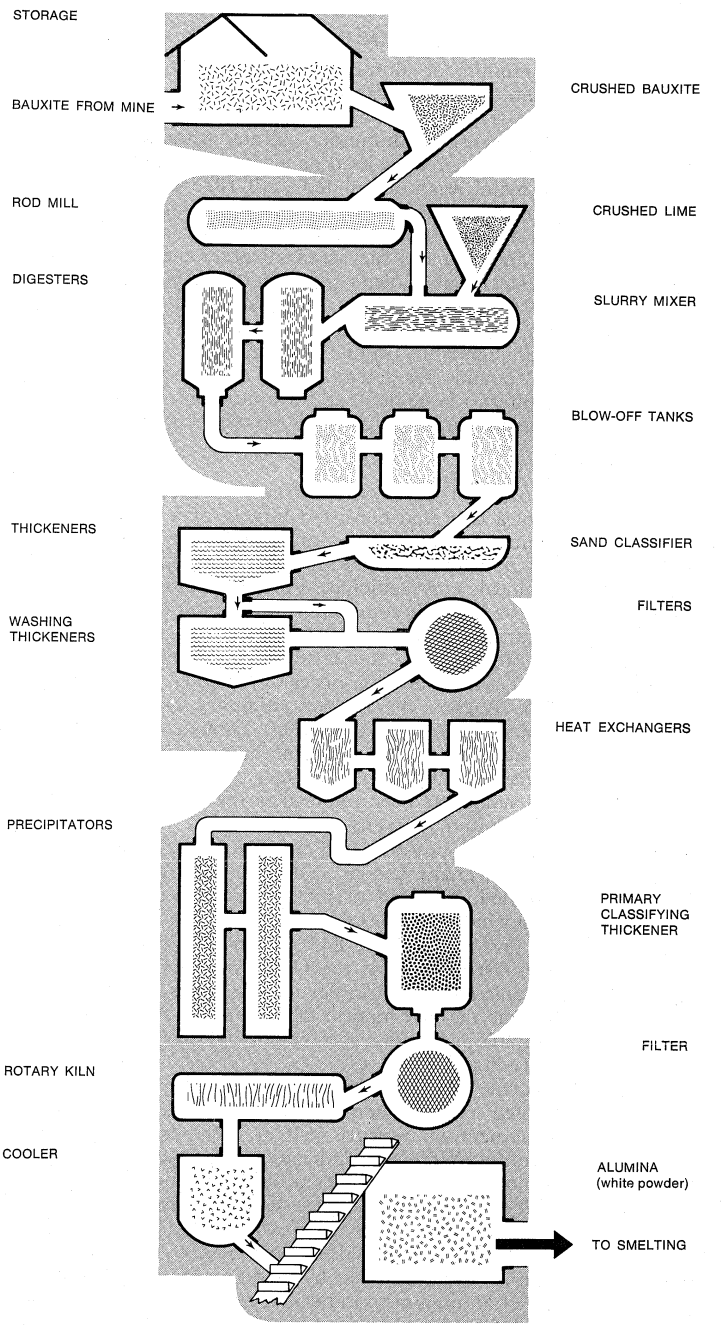


Figure 1: Principal operations in the Bayer aluminum refining process.

In the first step, bauxite is ground, mixed with a solution of caustic soda (sodium hydroxide) and pumped into large pressure tanks called digesters, where the ore is dissolved under steam heat and pressure. The sodium hydroxide reacts with the hydrated aluminum oxide of bauxite to form a saturated solution of sodium aluminate; insoluble impurities, called red mud because of their iron oxide content, settle to the bottom. During clarification, the mixture is passed through a series of pressure-reducing tanks (called blow-off tanks) and filter presses (shown in Figure 1 as the sand classifiers). Cloth filters catch the red mud, which, except at plants where lower grade ores are refined, is discarded. A resultant green liquor passes through filters into cooling towers (heat exchangers) and then into tall, silo-like precipitators. Sizable amounts of

hydrated alumina crystals are added to the solution in the precipitators as "seeding" to hasten crystal separation.

The seed crystals attract other crystals and form groups physically heavy enough to settle out of solution. This precipitate of aluminum hydrate crystals is filtered out, washed to remove impurities, and heated in long, rotary kilns at temperatures in excess of 1800° F (980° C). Free water and water that is chemically combined with the aluminum hydrate are driven off, leaving commercially pure alumina—or aluminum oxide—a superdry, fine white powder similar to sugar in appearance and consistency. It is still half aluminum and half oxygen by weight, bonded so firmly that neither chemicals nor heat alone can separate them.

Alcoa combination process

During World War II, the Alcoa combination process was developed for processing lower grade ores containing relatively high percentages of silica. Very briefly, this process reclaims the alumina that has combined with silica during the digestion process and has been filtered out with the red mud. The red mud is not discarded but is heated with limestone (calcium carbonate) and soda ash (sodium carbonate) to produce a sodium aluminate solution. This solution is returned to the digesters and proceeds through the remaining steps of the Bayer process. The waste product is a silicate material called brown mud.

Alumina produced by the Bayer process is quite pure, containing only a few hundredths of 1 percent of iron and silicon and traces of residual soda. In addition to being the primary raw material for producing metallic aluminum, alumina is an important chemical in its own right. It is used widely in the chemical, refractories, ceramic, and petroleum industries (see below *Aluminum chemicals*).

Refining four tons of bauxite yields about two tons of alumina. A typical alumina plant, using the Bayer process, can produce almost 3,000 tons of alumina per day. The cost of alumina can vary widely, depending on plant size and efficiency, labour costs, overhead, and cost of bauxite.

Smelting. Although there are several methods of producing aluminum, only three are worth attention. The Deville process, which involves direct action of metallic sodium upon aluminum chloride, was for years the basis of European aluminum production. It has been commercially abandoned in favour of the more economical electrolytic process, which involves dissolving purified alumina in melted cryolite and passing a current through the mixture to obtain metallic aluminum and carbon dioxide. A carbothermic approach, the oldest and the classical method for reducing (removing oxygen from) metallic oxides, has been for years the subject of intense research by the metal industries. This involves heating the oxide together with carbon. Carbon and oxygen combine to produce carbon monoxide or dioxide; the pure metal is left behind. The great attraction of carbothermic smelting is the possibility of bypassing the alumina refining cycle and of starting with lower-grade ores than bauxite. Despite many years of intensive research, however, no economic competitor has been found for the Bayer-Hall-Héroult approach.

The Hall-Héroult process

Although unchanged in principle, the Hall-Héroult smelting process of today differs greatly in detail from the original process. Modern technology has produced substantial improvements in equipment and materials, and has lowered final costs.

In a modern smelter (Figure 2), alumina is dissolved in smelting cells—deep, rectangular steel shells lined with carbon—that are filled with a molten electrolyte or conductor, consisting mostly of an aluminum compound called cryolite.

By means of carbon electrodes, direct current is passed through the electrolyte to the bottom of the cell, causing a crust to form over the surface of the molten bath. Alumina is added on top of this crust where it is preheated by the heat from the cell (about 950° C) and its adsorbed moisture driven off. Periodically the crust is broken, and the alumina is stirred into the bath to maintain proper concentration.

The actual process by which alumina is dissolved in the molten bath and electrolytically decomposed is a matter of dispute. The end results, however, are deposit of molten metallic aluminum on the bottom of the cell and the combination of the oxygen with the carbon, which then escapes as carbon dioxide.

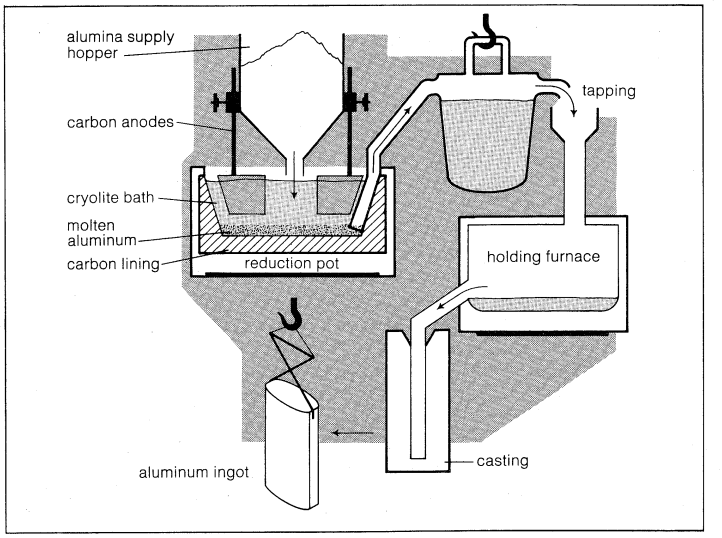


Figure 2: Simplified cross section of a modern aluminum smelter, showing the basic steps in the smelting process (see text).
By courtesy of Aluminum Company of America

The smelting process is continuous. Additional alumina is added to the bath to replace that consumed by reduction. Heat generated by the electric current maintains the bath in a molten condition so the fresh alumina is dissolved. Periodically, molten aluminum is siphoned off.

Because some aluminum fluoride from the cryolite electrolyte is lost in the course of the reaction, aluminum fluoride is added, as needed, to restore the chemical composition of the bath. An acid bath condition with an excess of aluminum fluoride provides maximum efficiency.

About 0.5 pound of carbon is consumed for every pound of aluminum produced.

Smelting cells

In actual practice, long rows of smelting cells, called potlines, are electrically connected in series. Normal voltages for cells range from four to six volts, and current loads range from 50,000 to 150,000 amperes. From 50 to 150 cells may form a single potline with a total line voltage of up to about 750 volts, thus making clear why power is one of the most costly ingredients of aluminum. Since 1900, aluminum producers have searched for sources of cheap hydroelectric power but have also had to construct many facilities that use energy from fossil fuels. Over the last four decades, technological advances have reduced the amount of electrical energy necessary to produce a pound of aluminum. In 1930, the figure was 12 kilowatt-hours. In the 1970s a single smelting cell will produce about 2,000 pounds of metal per day and use only six to eight kilowatt-hours per pound.

Finished molten aluminum is siphoned from smelting cells into large crucibles. From there, the metal may be poured directly into molds to produce foundry ingot, it may be transferred to holding furnaces for further refining or for alloying with other metals, or both, to form fabricating ingot. As it comes from the cell, primary aluminum averages 99.80 percent purity.

Automation and computer control have had a marked effect on smelter operations. The most modern reduction facilities use fully mechanized carbon plants and sophisticated potline computer control for monitoring and automating potline operations. As a result of modern technology, the man-hour requirement per ton of aluminum smelted has been reduced from about 25 in 1950 to half that in the 1970s.

ALUMINUM METAL

A ductile, silvery white metal, usually with a dull lustre due to a surface film of aluminum oxide, aluminum is

very light, weighing approximately one-third as much as an equal volume of copper or steel. It is corrosion resistant, an excellent conductor of heat and electricity, reflects both light and radiant heat, is nonmagnetic, does not readily absorb neutrons, can be safely used with foods and medicines, and can be formed by all known metal-working processes.

Aluminum can be joined by welding, brazing, soldering, adhesive bonding, rivetting, stitching, or stapling, by means of a number of mechanical assemblies, such as nuts and bolts, screws, and nails. It can be given a wide variety of mechanical finishes by grinding, polishing, buffing, abrasive blasting, and burnishing. A variety of chemical finishes can be used, such as alkaline or acid etches, bright dips (these give an extremely shiny finish to metal), chemical milling, and immersion plating. It is suited to an electrochemical process called anodizing. Or it can be electroplated with other metals, or given organic coatings such as paint, lacquer, and plastic films. Aluminum also is frequently finished by porcelain enameling or metallizing.

Strength of aluminum

High-purity aluminum (99.90 percent) is relatively soft and has a fairly low tensile strength of about 7,000 pounds per square inch in the annealed condition. (Annealing involves heating and then cooling slowly to make the metal less brittle.) By alloying and proper thermal and mechanical treatment, however, it can be made much harder and stronger, with tensile strengths as high as 100,000 pounds per square inch. Unlike some other metals, the strength and ductility of aluminum increases at supercold temperatures. Upon melting, the solid metal expands about 7 percent in volume, the solidification shrinkage being 6.6 percent of the liquid volume. Hydrogen is the only gas known to be appreciably soluble in molten aluminum; its solubility increases with temperature but becomes nearly zero when the metal freezes.

Aluminum may act as a base to form salts with acids or as a weak acid to form salts with strong alkalies. It is stable in air because of a thin, transparent oxide film that forms on exposure to air, protecting the aluminum from further oxidation and reaction. Growth of this natural oxide film is self-limiting; that is, when a thin layer is formed, further growth is halted. Molten aluminum is protected in air by a thicker oxide coating, which also deters further oxidation. Finely divided atomized or flake aluminum mixed with air and ignited will explode violently. Aluminum reacts rapidly with boiling water to liberate hydrogen and form aluminum hydroxide.

Aluminum is attacked only slowly by most dilute acids; concentrated hydrochloric acid dissolves it rapidly. Concentrated nitric acid (95–100 percent), however, has so little effect that aluminum drums and tank cars are used for shipping this chemical. Rate of attack by dilute acids is reduced as the purity of aluminum is increased. Alkalies, however, such as sodium and potassium hydroxide, attack it vigorously, regardless of its purity. Sodium carbonate solution attacks and corrodes aluminum, as does lime. Distilled water at ordinary temperatures has almost no action on ordinary aluminum, but salt water will slowly attack and corrode it. On the other hand, high-purity aluminum and certain aluminum-magnesium alloys resist corrosion indefinitely, even in seawater.

Metallic aluminum and its oxide and hydroxide are nontoxic and generally inert to living organisms. The metal is distributed so abundantly in the earth's crust that, inevitably, man breathes it in air and eats it in food, along with other minerals. Exhaustive research, however, has established no toxic effect of aluminum on the human body. On the contrary, certain properties of aluminum have led to its extensive use in medicines. Colloidal aluminum hydroxide and aluminum phosphate have been found to help in the treatment of peptic ulcers and symptomatic gastric hyperacidity. Inhalation of aluminum powders has been used as a prophylactic (preventive) therapy in the treatment of silicosis. The astringent property of some aluminum compounds has made them beneficial in the treatment of skin irritation, wounds, and burns. Water-soluble aluminum salts are used widely in cosmetics and as a component of deodorants.

ALUMINUM ALLOYS

Superpure aluminum

In its superpure condition (99.99 percent), aluminum lacks strength and hardness, but is formable, weldable, corrosion-resistant and an excellent conductor of electricity. Super-purity aluminum has many applications; in chemical equipment, reflectors, as a catalyst in making gasoline, in fine jewelry, and in electronic components. Most aluminum used today, however, is alloyed with other metals to improve its properties for special uses.

Although almost all the metallic elements readily alloy with aluminum, only a few are important major ingredients in commercial, aluminum-base alloys. These are copper, manganese, silicon, magnesium, and zinc. Smaller amounts of nickel, tin, iron, cadmium, lead, bismuth, cobalt, chromium, titanium, vanadium, boron, sodium, and zirconium may be added for particular purposes.

Aluminum–copper alloys, following heat treatment, possess mechanical properties superior to those of mild steel. These alloys have good ductility and machinability, but they do not approach aluminum–magnesium for corrosion resistance and weldability. They are well-suited for screw-machine products and structural applications requiring high strength-to-weight ratios, such as truck-trailer panels and aircraft components. Addition of about 1 percent manganese yields non-heat-treatable aluminum alloys that are stronger than commercially pure aluminum, have good corrosion resistance, and are highly workable.

Aluminum–manganese alloys are popular for cooking utensils, heat exchangers, chemical equipment, storage tanks, highway signs, roofing, siding, and many architectural applications.

Adding major amounts (about 10 percent) of silicon to commercially pure aluminum yields an alloy with a relatively low melting point. Thus, aluminum–silicon alloys are suitable for welding wire and brazing material. Because silicon imparts great fluidity to molten metal, this alloy is used also in castings. The addition of up to 5 percent magnesium yields an alloy with good tensile strength, weldability, hardness, and corrosion resistance in marine atmospheres. Current uses for aluminum–magnesium alloys include packaging (can stock), ships and boats, light standards, appliances, and cryogenic (low temperature) tanks. Adding both silicon and magnesium to aluminum produces alloys that are easily formed, machined, welded, and finished, have good resistance to corrosion, and are of medium strength. End uses include transportation equipment, architectural installations, bridge and highway structures, and machine parts.

Aluminum–zinc alloys are hard, heat-treatable, and possess very great strength. Additions of lesser amounts of magnesium, copper, and chromium—along with zinc—yield excellent alloys used widely in aircraft structures and mobile equipment.

Other metallic elements may be added to these basic aluminum alloys for special purposes. For example, bismuth, lead, and tin yield an alloy that is easier to machine; boron helps increase electrical conductivity; nickel gives improved strength at higher temperatures; and addition of titanium refines the grain of the metal, resulting in greater strength and ductility.

Casting alloys. Just as different casting methods, such as sand, die, or permanent mold, are best suited to specific applications, so are certain alloys best suited to particular casting methods. The elements found most often in aluminum casting alloys are silicon, magnesium, copper, and nickel. Silicon (up to 20 percent) improves the fluidity of molten aluminum, allowing it to flow more easily through a mold cavity. It also reduces shrinkage, decreases leaks, and makes welding easier. Copper may be added as a hardening constituent and to increase the strength of both heat-treated and non-heat-treated parts. Magnesium improves strength and corrosion resistance, and makes the alloy easier to machine. Nickel improves dimensional stability and strength at high temperatures.

In the early days of the industry, the casting alloys most used in the U.S. were of the aluminum–copper type. Britain and Europe preferred aluminum–zinc alloys. After

1920, aluminum–silicon was introduced. A 5 percent silicon alloy became popular in the U.S. for architectural and other applications in which good casting properties and high resistance to corrosion were more important than strength. After World War I, alloys containing about 10 percent copper were used throughout the world for pistons and other parts of internal-combustion engines. Most of the aluminum alloy castings made in the U.S. up to 1940 were of 7–10 percent copper, although heat-treated alloys containing 4–5 percent copper and 10 percent magnesium were employed extensively where higher strength and ductility were required.

Heat treatment required for casting

The heat treatment of castings requires a relatively long heating (6–24 hours) at temperatures of 700–1,000° F (370–535° C), the exact temperature and time depending on the alloy composition and type of casting. This is followed by quenching the casting in hot or cold water or oil, with subsequent aging at either room temperature or a temperature somewhat above the boiling point of water. The purpose of this "solution heat treatment" is to make certain that as much as possible of the alloying elements, particularly copper, magnesium, and zinc, are dissolved in the aluminum in what is termed a solid solution.

The reason for the long duration of heat treatment is inherent in the structure of the castings. When a casting is made from an aluminum-base alloy containing copper, for example, the crystals that first form during the cooling process are nearly pure aluminum. Most of the copper remains in solution in the liquid aluminum, and as the aluminum crystals grow and use up the liquid aluminum, the concentration of copper in the liquid increases. Since increasing the proportion of copper in aluminum lowers its freezing point, this process continues as the casting cools, until that temperature is reached at which the lowest melting point solution (eutectic) of aluminum and copper freezes. In the meantime, however, the crystals have grown until they almost touch one another, so that when the eutectic freezes it fills the small intergranular spaces and acts as a cement between the crystals. Aluminum can hold approximately 4.0 percent of copper in solid solution at the heat-treating temperature (about 950° F). Only a small fraction of this, however, is actually in solid solution. The rest is in the intergranular eutectic. In order to obtain a nearly saturated solid solution in the aluminum, it is necessary to hold the casting at high temperatures long enough for the copper to diffuse from the grain boundaries and distribute itself throughout the grains or crystals. The larger the grains, the longer this takes. Consequently, fine-grained castings (such as those cast in permanent molds) require a shorter heat-treating time than coarse-grained sand castings. Wrought material (in which the eutectic has been broken up and mechanically dispersed) requires a much shorter heat-treat time than permanent mold or sand castings.

Supersaturated solutions

If the casting is suddenly cooled or quenched to room temperature, a supersaturated solid solution results, since the solid solubility of copper in aluminum at room temperature is about 0.25 percent. Unless some accelerating metal, such as magnesium, is present, this solid supersaturated solution is unstable and the copper precipitates slowly during a long period of storage at room temperature. If the temperature is raised to the boiling point of water or higher, however, the mobility of the atoms of copper in the alloy is sufficiently increased so that they are able to get together in a few hours. They apparently precipitate in the form of submicroscopic particles of a copper–aluminum compound. With longer heating and at higher temperatures, these particles grow large enough to become visible under the microscope upon a polished and properly etched surface of the alloy. They are found to be arranged in quite definite patterns along certain planes in the individual crystals. Apparently these minute particles of hard constituent remain between the crystals, just as particles of fine sand, placed between sheets of paper, cause them to resist the tendency to slip. The precipitation process thus increases the tensile strength and yield strength of the alloy but decreases its ductility. In the presence of 0.5 percent magnesium, this precipitation occurs spontaneously in a few days at room temperature. Such precipitation at room temperature is spoken of as natural aging, while precipitation induced by heating at elevated temperatures is called artificial aging.

In certain aluminum–magnesium alloys, the solid solubility of magnesium at the heat-treating temperature and at lower temperatures is considerably greater than that of copper, but the processes and results of heat-treating, quenching, and aging are similar. The solubility of silicon varies little with temperature, so that heat treatment of an aluminum–silicon alloy has little effect on its properties. If magnesium is present, however, a magnesium–silicon compound is formed, which has an appreciable solid solubility at heat-treating temperature and precipitates by artificial aging after quenching. A heat-treatable casting alloy of this type may contain about 5 percent silicon, 0.5 percent magnesium, and 1.5 percent copper. Its tensile strength is lower than that of the heat-treated copper or magnesium alloys, but it has much better casting characteristics.

The mechanical properties of castings vary greatly, depending not only on composition but also on casting conditions. Thin sections generally show a somewhat higher tensile strength than do thick sections. Heat-treated alloys as a rule show higher strength and ductility than those not heat treated. Typical strengths for ordinary sand-cast alloys are in the range of 19,000–30,000 pounds per square inch, while heat-treated alloys may show 30,000–45,000 pounds per square inch. As a general rule, an alloy cast in permanent (metal) molds will show tensile strengths 5,000–10,000 pounds per square inch higher than when sand-cast.

Typical strengths of sand-cast alloys

Wrought alloys. Wrought alloys are available in worked form, such as sheet, foil, plate, extrusions, forgings, rod, bar, and wire. Working operations and heat treatments change ingot structure into a recrystallized or severely fragmented structure, characteristic for each wrought alloy, which determines strength, corrosion resistance, and other properties. Wrought alloys fall into two basic classes: non-heat-treatable alloys and heat-treatable alloys.

The first wrought-aluminum alloy was probably a relatively pure aluminum, since the pure metal's softness and high ductility permitted it to be rolled, forged, or drawn with a minimum of effort. Alloying elements usually cause the pure metal to become stronger, harder, and less ductile and, consequently, more difficult to fabricate. A high grade of pure aluminum was used early in this century for electrical conductors. Known as EC (electrical conductor) metal, this wrought material is usually 99.45 percent aluminum. Commercial "pure aluminum" sheet contains 99.5–99.79 percent aluminum. Foil is generally made of high purity metal, ranging up to 99.99 percent. For the flexible packaging market, products such as collapsible tubes are usually made of 99.8 percent aluminum. Wrought products of pure aluminum are used in many fields, with major consumption in the packaging, chemical, electrical, and architectural fields.

In the manufacture of wrought aluminum articles, maximum ductility is produced by annealing the article, usually by heating for some time at about 650° F (340° C). The hardening and strengthening effect of even small amounts of alloying metals is evident in comparing typical properties of commercial-purity annealed aluminum sheet with sheet of 99.99 percent purity. The typical tensile strength of about 7,000 pounds per square inch for the super-purity metal increases to 13,000 pounds per square inch for commercial-purity sheet, while the yield strength increases from 1,500 pounds per square inch to 5,000 pounds per square inch. (The yield strength of aluminum is defined as the load in pounds per square inch that produces a permanent stretch or elongation of 0.2 percent.) At the same time, the elongation in two inches is reduced from 50 percent to 35 percent in commercial-purity sheet.

Cold-working (rolling, forging, drawing, and so on, at ordinary temperatures) hardens the metal, increasing both tensile and yield strength. This effect increases with the amount of cold work. The magnitude of change is seen from the fact that a typical tensile strength for hard-

rolled commercial-purity aluminum sheet is 24,000 pounds per square inch with a corresponding yield strength of 22,000 pounds per square inch and an elongation of only 5 percent in two inches. Intermediate amounts of cold working produce intermediate tempers.

Aluminum–magnesium alloys, containing five to six percent magnesium, form a useful series of non-heat-treatable, solid solution, wrought alloys. Strengths after the metal has been annealed range from 16,000 pounds per square inch tensile strength and 6,000 pounds per square inch yield strength to 52,000 pounds per square inch tensile strength and 24,000 pounds per square inch yield strength. Elongations may exceed 25 percent. Aluminum–magnesium wrought alloys have good strength, forming, and welding characteristics and high resistance to corrosion. Finishing and decorative capabilities are also characteristic of the aluminum–magnesium alloy-system.

Aluminum–magnesium–silicon forms a family of alloys used widely in wrought products for architectural and transportation applications. Aluminum–zinc–magnesium alloys, when heat-treated and aged, have very great strength and are employed principally in aircraft components and mobile equipment.

Aluminum–copper wrought alloys combine wide strength range with good weldability, superior resistance to stress corrosion, and good elevated-temperature properties. Additions of lead and bismuth provide superior machinability. Aluminum–copper–magnesium alloys, introduced over 50 years ago, are outstanding structural alloys in the aircraft industry.

Miscellaneous alloys in heat-treatable and non-heat-treatable classes span a broad range of composition and applications. Increasingly popular and useful are composite wrought products, consisting of an aluminum-alloy core coated on one or both surfaces with a thin layer of corrosion-resistant pure aluminum or aluminum alloy. The combination of core and cladding alloys is selected so that the cladding provides electrochemical protection for the core. When a corrosive solution (electrolyte) is in contact with the product, current from the cladding flows through the electrolyte to the core and the cladding is dissolved, thus protecting the core. The most widely used clad aluminum products are sheet and plate.

ALUMINUM POWDER AND PARTICLES

The phrase "powder and particles" encompasses a family of forms in which 99.5 percent unalloyed aluminum ingot is formed into powders or granules. Ingot is melted in a furnace and passed through streams of air or inert gases. The molten aluminum forms a spray that solidifies into granules and is collected. These atomized granules range in size from 0.015 micron (one micron equals one 1/1,000 of a millimetre) up to 17,000 microns. Powders normally are defined as being finer than 1,000 microns, and particles range up to 17,000 microns. In shape, powders and particles may be granular, globular, irregular, spherical, elongated, ribbon, flake, or other, with varying surface area, unit weight, and density.

Applications include paints and pigments, missile fuels, chemicals, explosives, and a variety of putties, cements, and solders. Aluminum pigments are used widely in metallic paints on storage tanks, bridges, and towers, where durability and corrosion resistance are required. Particles in flake form in aluminum paint arrange themselves in parallel layers five to ten flakes deep, forming, in effect, a thin layer of metal. Atomized powders are major components of rocket propellants, flares, and fireworks. Mixed with uranium compounds, they are used in nuclear fuel elements. Aluminum powder is utilized to produce aluminum-clad steel wire for the electrical industry. In aluminothermic reactions, aluminum powders are combined with the oxides of other metals with the production of heat. Aluminum's great affinity for oxygen makes it useful in the aluminothermic smelting of steel, chromium, and vanadium.

Use in metallic paints

Aluminum powders are used in powder metallurgy processes in which the powders serve the same function as molten metal in casting operations. Powders are put in a die and subjected to pressure. The compacted piece is heated to produce aluminum parts having specific desirable properties.

ALUMINUM CHEMICALS

Aluminum oxide. Aluminum oxide exists in several different crystallographic forms, of which corundum is most common. Corundum is characterized by a high specific gravity (4.0), a high melting point (about 2,015° C or 3,660° F), great insolubility and hardness.

Aluminum oxide is the major ingredient in the commercial chemicals known as aluminas. Of the pure, inorganic chemicals, aluminas are among the largest volume produced in the world today. Rubies and sapphires are crystalline, nearly pure varieties of alumina, coloured by small amounts of impurities. Synthetic rubies and sapphires are made commercially by fusing a mixture of high purity aluminum oxide with colouring agents in an oxy-hydrogen blowpipe flame. Most are cut and drilled to form tiny "jewel" bearings in watches and various precision measuring instruments.

Activated alumina is a porous form of aluminum oxide from which much of the chemically combined water has been driven off. It is chemically inert to most gases, nontoxic, and will not soften, swell, or disintegrate in water. It has the ability to adsorb and hold moisture without change in form or properties, and has high resistance to shock and abrasion. Activated alumina is used in oil, chemical, and petrochemical industries as a dehydration agent and purifier in the manufacture of gasoline, petrochemicals, natural gas, and hydrogen peroxide.

Calcined alumina is aluminum oxide that has been heated at temperatures in excess of 2,000° F to drive off nearly all chemically combined water. In this form, alumina has great chemical purity, extreme hardness (nine on the Mohs hardness scale, on which diamond is ten), high density and a high melting point (above 3,700° F). It possesses good thermal conductivity, heat and shock resistance, and great electrical resistivity at elevated temperatures. This combination of properties makes calcined alumina useful in abrasives, glass, porcelains, spark plugs, and electrical insulators, but the greatest quantity of calcined alumina is smelted to obtain aluminum.

Tabular alumina

Tabular alumina is aluminum oxide that has been heated to temperatures above 3,000° F. Composed of tablet-like crystals, it has very high capacity and thermal conductivity, and exceptional strength and volume stability at high temperatures. For these reasons, a major usage for tabular alumina is in the production of high-quality refractories, the materials used for lining industrial furnaces. High-alumina refractories are used in the metal and glass industries in boiler installations, large furnaces and kilns for smelting metals and firing glass, pottery and porcelain, and in the manufacture of building bricks.

Most refractories are produced in the form of brick, bonded and fired in furnaces. Some castable refractories are made in the form of mortars, usually tabular alumina with calcium aluminate cement as a binder. These mortars, called grog, are sprayed under pressure to form the linings of the steel industry's basic oxygen furnaces, electric ore furnaces, steel ladles, and coke ovens, and for steam boilers, rotary kilns, and many other high temperature applications.

In addition, fused aluminas are used in special refractories for the glass industry. Fused alumina is calcined alumina that is melted in electric arc furnaces, cooled, crushed, and recast into desired shapes. In another application, industrial processes requiring hot gases use a unique heat-transfer device called a pebble heater. Gases to be heated are passed through a bed of tabular alumina balls that have been heated to extreme temperatures. In still another application, an aluminous insulating material is formed by melting alumina and silica in an electric furnace and subjecting the molten mixture to high-velocity gases to produce fine white fibres.

Aluminum hydrates and other compounds. The hydrous forms of alumina, called aluminum hydrates, may contain either one (monohydrate) or three (trihydrate) molecules of water. Each may exist in two different

crystalline forms, known as "alpha" and "beta." In both hydrates, the alpha variety is more common. The alpha trihydrate (gibbsite) and alpha monohydrate (boehmite) occur in bauxite.

Hydrated alumina is used extensively in the production of aluminum chemicals, such as alum, sodium aluminate, aluminum fluoride, and hydrous aluminum chloride. It is a constituent of petroleum catalysts, and an ingredient in the manufacture of plastic and rubber goods, paper, glass and vitreous enamel, adhesives, varnishes, and toothpastes.

Aluminum sulfate is used in water purification. Aluminum chloride in various forms is used as a catalyst in organic chemistry, and in the cosmetic industry as a deodorant. Aluminum fluoride is a popular compound used widely in the purification of aluminum.

ALUMINUM FABRICATION

Aluminum has been called the obliging metal because it can be fabricated by all the techniques known to metalworking. Similarly, it can be joined to itself or to other materials by the total range of joining procedures and finished easily and economically by many methods (see also METALLURGY).

Fabricating methods. *Casting.* In casting molten aluminum is poured—or forced under pressure—into a mold, the shape which it then takes. There are many variations. In sand-casting, expendable molds of special sands and clays are formed around a pattern into a shape. Molten aluminum at about 1,300° F is then poured into the sand cast. The cast is discarded after use. The process is employed for large shapes and small quantity runs. Plaster mold casting, a refinement of sand-casting, provides parts with smooth surface finishes and fine dimensional control. Permanent mold castings employ metal molds and cores into which molten aluminum at about 1,350° F is poured. Such a mold may be used thousands of times to produce parts with smooth surfaces and superior strength and soundness. In die-casting, molten metal at about 1,200° F is injected into water-cooled, heat-treated steel dies at high velocities under pressures up to 20,000 pounds per square inch. This is the preferred method for long production runs and for products requiring good finishes, dimensional accuracy, and internal soundness. Shell mold casting uses a thin shell of sand, which has been preformed and baked in a metal pattern. Investment casting is a two-step process in which wax is first cast in permanent molds, a new plaster mold prepared around the wax, the wax melted out, and molten aluminum poured into the cavity. Centrifugal casting uses centrifugal force to drive molten metal into a complex die.

Newly developed is premium-engineered casting that utilizes combinations of different casting processes to develop final products with excellent mechanical properties, close control of dimensions, thin sections, and smooth surfaces.

Rolling. Rolling is the process by which aluminum ingot is reduced by passing it between pairs of rolls. Rolling ingot, which may be up to 30 feet (9 metres) long, 36 inches (90 centimetres) wide and 20 inches (50 centimetres) thick, generally passes through a scalper that trims and smooths the rolling surfaces and edges of the ingot. The ingot then is heated to about 800° F and passed through a hot reversing mill that in successive passes rolls the ingot down to plate size which varies from six inches (150 millimetres) to 0.250 inches (6 millimetres). For further reduction to sheet, 0.249 inches to 0.006 inches, plate is passed through a series of cold-rolling mills. Before the final cold-rolling, it is usually annealed to attain optimum metallurgical qualities. Heat-treatable sheet is passed through a heat-treating furnace after it has been cold-rolled. Non-heat-treatable sheet goes directly from cold rolling to finishing operations. Plate, which is not generally cold-rolled, goes from hot rolling to heat treating or finishing operations. Further rolling of sheet down to less than 0.006 inches produces foil. Aluminum foil can be rolled so thin that it takes ten layers to equal the thickness of a newspaper page. Rolling also may be used to produce rod and bar stock. Specially formed blocks or bars of the metal, called blooms, are passed through successively smaller and variously shaped rolls to achieve the desired shapes and sizes.

Aluminum foil

Extruding. In extruding, a hydraulic ram forces an ingot or billet, which generally has been heated to about 800° F, through an orifice in a die. The metal takes the shape of the orifice, in cross section, and is sawed to length. Theoretically, there is no limit to the variety of shapes that can be extruded. Size is limited only by the power of the press and the size of the die block. Hydraulic presses range from about 500 tons pressure for soft alloys to 14,000 tons for hard alloys. Extruded shapes generally fall into three categories: solid, a shape other than rod and bar that does not enclose a void, as in the letter I; semihollow, a shape with part or all of its cross section partially enclosing a void, as in the letters Y and E; and hollow, a shape with part or all of its cross section completely enclosing a void, as in the letters P and O.

Forging. In aluminum forging, preheated ingots or billets are formed into desired shapes by repeated blows of a hammer, or by the application of pressure in a press. Forging hammers range from 500 to 50,000 pounds capacity. Drop forging employs a hammer blow to force the metal into a die cavity. Dies are usually produced from forged alloy steel blocks. Press forging uses a fast push, rather than a blow to force metal into the die. Die forging, hand forging, and ring forging are variations of the basic process.

Drawing. Aluminum can be drawn; that is, pressed into a hole or cavity whose shape it retains. This basic process, called deep drawing, is used to produce aluminum cups, pots, and similar containers. In cold-drawing, a pointed length of rod is pulled through a die orifice that is smaller than the overall diameter of the rod. Aluminum wire is formed by this process, being drawn through successively smaller die orifices until the desired gauge is reached. Generally, rod becomes wire when it is less than ⅜ inch in diameter. Cold-drawn tubular products are made in much the same manner, except that a bar of steel (mandrel) is employed to form the core of the tube, and is withdrawn after the tube is complete.

Other processes. Aluminum and its alloys are distinguished by the ease with which they can be machined. Castings, forgings, and other products are brought to final shape, dimension, and finish on lathes, joiners, routers, drill presses, shears, grinders, and many other high-speed metal-working machines. Other fabricating techniques used successfully with aluminum are spinning, stamping, embossing, coining, bending, rotary swaging and cold heading (see also MATERIALS PROCESSING).

Heat treatment. Besides processes that change the metal's shape or mold the metal into desired forms there are certain processes that insure the metal's final characteristics and metallurgical qualities. They include heat treating and stress relieving.

There are four general types of heat treatments applied to aluminum and its alloys: preheating, annealing, solution heat treatment, and precipitation heat treatment.

Cast ingots are usually preheated before cold- or hot-working. Preheating improves workability, helps control recrystallization and grain growth, and causes desirable diffusion of alloying elements. Most aluminum alloys are heat-treated in gas-fired or radiant-heated-air or controlled-atmosphere furnaces. Depending on size and alloy, typical preheats for aluminum forging or extrusion ingots range from four to 48 hours at temperatures of 800 to 1,200° F (425 to 650° C).

Annealing results in a metal of softened, workable condition, relieves internal stresses, and stabilizes properties and dimensions. Certain non-heat-treatable, wrought-aluminum alloys are heated rapidly to annealing temperature to insure fine grain size. Typical annealing conditions for heat-treatable wrought alloys are two to three hours at 775° F (410° C). Careful control of cooling rate is essential to obtain maximum softness and stability of properties.

Typical annealing conditions

Solution heat treatment improves mechanical properties by developing maximum concentration of hardening constituents in solid solution. This requires heating the metal

to a temperature close to the eutectic temperature, holding it there for the required time, and cooling it fast enough to retain the desired solid solution. Wrought alloys frequently are treated in molten salt baths, usually sodium nitrate. Proper quenching is vital. As the metal cools, alloying elements start to precipitate, affecting the mechanical properties of the metal. Immersion quenching in water at room temperature is used primarily for wrought alloy products. Most casting alloys are quenched in hot (150–212° F) water, which reduces internal stresses and minimizes distortion in the product. Hot oil and cold water quenching also are used for certain products.

Precipitation heat treating, often called "artificial aging," is heating at elevated temperatures to harden the alloy. This process also is used where reduction of internal stresses is required to stabilize the dimensions of a part or to prepare the metal for extensive machining.

Stress relieving reduces internal pressures in a metal body. These pressures, or stresses, are unequal plastic deformations in the metal, caused by fabricating or temperature gradients or other external forces. Such residual stresses may be relieved by annealing or by mechanical operations, such as stretching or roller levelling, or by combinations of mechanical and thermal methods.

Joining. Aluminum can be joined by most methods used for other metals. Of 40 welding processes aluminum has been commercially welded by 32. Aluminum can also be joined by brazing and soldering, that is, by welding with the addition of an alloy with a high melting point (brazing) or a low melting point (soldering; see also WELDING, BRAZING, AND SOLDERING). The light metal can be joined by a wide range of mechanical fasteners, including rivets, nuts and bolts, screws, nails, clips, and special structures. It can be adhesive bonded and, in lighter gauges, stitched or stapled.

Finishing. Aluminum lends itself to a wide variety of finishing techniques for appearance, or for corrosion, erosion, or abrasion resistance. These may be classified into mechanical, chemical, electrochemical, and electroplated finishes and applied coatings.

Mechanical finishing. Mechanical-finishing operations are carried out to remove surface irregularities, smooth the surface, and apply a texture to the surface. Grinding, the rapid removal of surface metal, is accomplished with an abrasive belt or wheel. Polishing, a form of grinding, uses finer abrasives with special lubricants. Buffing brings out high lustre by eliminating surface marks produced during polishing. In buffing, a fine abrasive powder is embedded in a grease binder and rubbed on, instead of being glued to the face of the wheel. Satin finishing imparts a soft, smooth finish with low reflectance by means of fine, irregular lines scratched on the surface of aluminum with a rotating wire brush, buffing wheel, or steel wool.

Satin finishing

In barrel burnishing, abrasive chips (in water, lubricant, detergent, or other medium) rub against the metal in a rotating or vibrating container, imparting a bright, smooth surface. This is an economical way to finish small parts. In abrasive blasting, handsome matte or satin finishes are produced by subjecting the metal to an air blast or air–water blast into which has been introduced an abrasive, such as fine silica, glass beads, or steel shot. Shot peening is similar to abrasive blasting except that large-diameter steel shot is used. Peening cleans and textures aluminum surfaces, while at the same time it work-hardens the metal, and improves its fatigue resistance. In embossing, patterns are applied to aluminum sheet by passing it through engraved rolls.

Chemical finishing. Chemical finishing treatments on aluminum function as cleaners, etchers, and depositors. Chemical cleaners prepare the surface for subsequent operations. Chemical etches roughen the natural surface layers of aluminum. Chemical depositors add metallic film or coatings to the surface.

Alkaline etches impart an inexpensive matte finish to aluminum. The most frequently used alkaline etchant is an aqueous solution of sodium hydroxide (caustic soda). An acid bath is necessary to remove smut produced by the chemical reaction. The surface is microscopically pitted, which gives the metal a characteristic nonreflective appearance. Acid etches achieve a variety of results. Mixtures of sulfuric and chromic acids are used as mild etches to remove stains made during heat treatment, or to prepare aluminum alloys for printing, anodizing, or other finishing operations. Mixtures of nitric and hydrofluoric acids produce semibright matte finishes on aluminum alloys. Concentrated solutions of hydrochloric acid are used to deep-etch aluminum. The addition of cobalt or nickel salts hastens the action, yet leaves a smooth surface.

Both alkaline and acid etchants are used widely in chemical milling, by which the metal thickness is reduced in a controlled manner. Dissolution rates of about 0.001 inch of metal per minute are typical. Chemical milling (also called deep etching) is used to obtain surface patterns through selective surface metal removal, or to remove metal from highly complex shapes, or to achieve overall reduction in dimension and weight of large aluminum parts in cases in which machining would be infeasible or too costly.

Immersion plating is a type of chemical finish in which thin films of other metals are deposited galvanically when aluminum is immersed in aqueous solutions of certain lead, tin, copper, silver, brass (zinc and copper), cadmium, iron, or zinc compounds. These have different purposes: zinc or tin may serve as a surface preparation for electroplating; tin, lead, or brass surfaces reduce friction; silver promotes surface electrical conductivity.

Electrochemical finishing. Anodizing is an electrochemical finishing process unique to aluminum; an electric current creates an aluminum oxide coating on the surface that is integral with the aluminum. Billions of pores in the film permit the electrolyte to reach the metal surface and the current to flow. Different alloying constituents in the base aluminum can produce a variety of attractive colours in the product. Also, high porosity permits the oxide coating to adsorb applied dyes, pigments, and lubricants. The oxide film can then be effectively sealed by immersion in pure boiling water. Thus, the colour or dye or additive becomes integral with the metal. Anodic coatings on aluminum are very popular for architectural applications.

Electroplating. Electroplated coatings are applied to aluminum alloys to obtain a specific appearance, or to impart a special property, such as increased wear resistance, electrical conductivity or improved solderability.

Applied coatings. Aluminum's inherent resistance to corrosion makes it an ideal receiver for applied coatings such as paints and lacquers, plastic films, and porcelain enamels. Organic coatings, paints, lacquers, and enamels, can be applied by any of several conventional techniques, such as spraying, roller coating, dipping, and flow coating. Plastic films can be laminated to aluminum sheet, using proper adhesives. Plastic powders can also be applied by electrostatic spraying; particles, negatively charged as they leave the spray gun, are attracted to the object to be painted, which is grounded. The term porcelain enamel refers to a family of finishes with widely varying composition and properties. An opaque porcelain coating can be applied to aluminum by fusing a finely milled and pigmented glass frit (a partly fused mixture) to a prepared surface by high-temperature firing. Such coatings give aluminum the characteristics of glass: hardness, resistance to chemical attack, and colour fastness. Finally, certain metallic coatings can be applied to aluminum by vacuum metallizing, hot dipping, and spraying to add the properties of the applied metal in specific applications.

Porcelain enamel

USES OF ALUMINUM

Aluminum's high thermal and electrical conductivity, light weight, and corrosion resistance were the basis of most of the early large uses of the metal, including the manufacture of cooking utensils and electrical conductors. With the development of stronger, improved cast and wrought alloys between World Wars I and II, aluminum became increasingly important as a structural material, particularly in transportation and architecture. During World War I, Germany pioneered the use of

wrought alloys of the duralumin (aluminum–copper–magnesium–manganese) type. These were used in the framework of Zeppelin airships and in the structural parts of airplanes, replacing wood. Aluminum castings were used in airplane engines in World War I, and subsequently the light metal was used in aircraft in the form of sheet, extrusions, and tubing. At the same time, aluminum's high strength-to-weight ratio and economy led to wider use in bodies, crankcases, and other parts for passenger automobiles, buses, and trucks. Aluminum in tank trucks and highway trailers became popular, as it did in high-speed trains and passenger automobiles.

Aluminum's attractive properties have led to its increasing popularity for architectural uses after about 1925. Aluminum windows, doors, roofing, and miscellaneous hardware gained wide use in the period prior to World War II. Its nontoxicity and resistance to many corrosive materials led to applications in the chemical, processing, and brewing industries. At the same time, the widespread use of aluminum paint developed in the U.S. Aluminum foil gained popularity as a wrapping material and was used in the electrical capacitors of radios. Foil and thin aluminum sheet went into many types of caps and closures for bottles. During World War II, aluminum production was devoted almost entirely to aircraft applications; what little remained went into other military and naval equipment.

In the 1950s and '60s aluminum applications spread through the construction, transportation, electrical, packaging, machinery, and many other industries. In building construction the use of aluminum for doors, windows, screening, roofing, siding, gutters, railings, and many other applications has grown rapidly, replacing a variety of once traditional building materials. The 1,350-foot towers of the World Trade Center in New York have roughly 4,500 tons of aluminum in their exterior curtain walls. Automobiles use aluminum in transmissions, engines, trim, electrical systems, air conditioning, brakes, and paint. The largest commercial aircraft have become virtually all aluminum. Urban rapid-transit systems are another major new user of aluminum. In the electrical industry, aluminum goes into machinery and equipment of all kinds, including cable and wire for power transmission. In packaging, the largest growth has been in aluminum cans for food and beverages.

Since World War II, aluminum production in the United States, the leading producing nation, has outdistanced the growth of the gross national product by two to one, with substantial exports to Europe and Japan. In West Germany the demand for processed aluminum has increased at about 20 percent annually, while overall market growth in Japan is about 30 percent. France, the United Kingdom, and other western European nations are also large users. The three principal fields of application outside the U.S. are, generally speaking, electrical systems, transportation, and construction.

Growth of the industry

The worldwide aluminum industry has grown at a steadily increasing pace. In 1928 only ten countries in the world were aluminum producers; only the United States, the Soviet Union, and Canada produced substantial amounts. By the 1970s, some 15 countries, including China, Japan, and Taiwan, had primary production capacities exceeding 100,000 tons per year. Estimated world production for 1969 was 10,019,000 tons. Of this total, North America accounted for 49 percent, Europe 37 percent, and Asia 9.0 percent. The eight leading producers were the United States, the Soviet Union, Canada, Japan, Norway, France, West Germany, and Italy. Yet, in view of the fact that virtually every country in the world is an aluminum consumer, the restriction of production to a few countries represents a situation that most certainly will change; emerging nations will develop their own natural resources and build up their industries; and major importing nations will expand domestic capacities.

BIBLIOGRAPHY. General references include: K.R. VAN HORN (ed.), *Aluminum*, 3 vol. (1967), a comprehensive treatment; J.D. EDWARDS, F.C. FRARY, and Z. JEFFRIES, *The Aluminum Industry*, 2 vol. (1930), a general survey; P.V. FARAGHER, *Fundamentals of Aluminum* (1959, reprinted 1967); AMERICAN METAL MARKET, *Aluminum Supplement* (annual); ALUMINUM ASSOCIATION, *Aluminum Statistical Review* (annual); ALUMINUM COMPANY OF AMERICA, *Aluminum Handbook Series* (1967). Other suggested readings are: S.L. BRUBAKER, *Trends in the World Aluminum Industry* (1967); A.C. HUANG, *Past and Prospective Trends in the World Aluminum Industry* (1968); S. SPECTOR, *Aluminum Industry Review 1968–69* (1969); U.S. DEPARTMENT OF COMMERCE, BUSINESS AND DEFENSE SERVICES ADMINISTRATION, *Aluminum Factbook* (1963). Articles on this subject appear regularly in the periodical *Metals Week*.

(K.R.V.H.)

Amazonas

Amazonas, the largest state of Brazil, and situated in the northwestern part of the country, is bounded to the northwest by Colombia; to the north by Venezuela and the Brazilian territory of Roraima; to the east and southeast by the Brazilian states of Pará and Mato Grosso; to the south by the Brazilian territory of Rondônia; to the southwest by the Brazilian state of Acre; and to the west by Peru. It has an area of 604,036 square miles (1,564,445 square kilometres). Despite its size, it is the most thinly populated Brazilian state, with only 961,000 inhabitants.

Amazonas occupies the greater part of the Hiléia (Hylea), the forest zone of the Amazon Basin, and has its northernmost tracts north of the Equator. The capital, Manaus, stands in the eastern part of the state at the confluence of the Rio Negro with the mainstream of the Amazon. The state takes its name from the Amazon River—Rio Amazonas in Portuguese. Most of the territory is tropical forest, penetrable only by means of the rivers intersecting it (for associated physical features, see AMAZON RAIN FOREST; AMAZON RIVER).

Exploration by Francisco de Orellana

History. The Spanish explorer Francisco de Orellana passed through this region in 1541–42, during a voyage down the Amazon from the Coca, one of its Andean headwaters, to its Atlantic estuary. Near the Nhamundá confluence (on the easternmost border of the modern state), Orellana encountered what he took to be a tribe of female warriors; by associating them with the Amazons of classical legend he gave the river its name. In 1669 a Portuguese captain, Francisco da Mota Falcão, founded the fort of São José do Rio Negro, on the site of the future Manaus; and in 1755 the Capitania de São José do Rio Negro was established, with its seat of administration at Mariuá, the future Barcelos, some way further up the Rio Negro. The rubber trees of the region had already been discovered by the Portuguese in the early 18th century, when a missionary, Frei Manuel da Esperança observed the Cambeda Indians producing rubber.

After Brazilian independence, Rio Negro remained dependent on the state of Pará until 1850, when it gained autonomy, becoming the province of Amazonas in 1852. After the overthrow of the imperial regime in 1889 (see BRAZIL, HISTORY OF), the province became a federal state, adopting a constitution in 1891.

In 1866 a law of the imperial regime had opened all ports on the Amazon, as far as the Peruvian frontier, to foreign shipping. From 1880 to 1910 the rubber trade brought great prosperity to Amazonas, for which a modern port was constructed at Manaus in about 1900. The state's rubber trade later suffered from the competition of new rubber plantations in The Netherlands East Indies and in Malaysia; but production rose again in World War II, when the Allies lost their Far Eastern plantations to the Japanese.

In 1946 the Brazilian government launched a plan for the economic development of Amazônia, and a new federal constitution stipulated that not less than 3 percent of the union's revenue from taxation should be devoted to the Hiléia for at least the next 20 years. After the expiration of that minimum period, the plan was continued into the 1970s.

Physical geography. Geologically the land consists mainly of sedimentary deposits of recent formation. Except on the northern borders, where Neblina Peak (Pico

da Neblina) reaches 9,870 feet (3,108 metres) (the highest point in Brazil), the mean altitude is not more than 300 feet above sea level.

Rivers

The mainstream of the great Amazon River (known as the Rio Solimões from the Peruvian frontier to the Rio Negro confluence) traverses the state from west to east, its major tributaries being the Içá, the Japurá, and the Negro from the north and the Javari, the Juruá, the Purus, and the Madeira from the south.

With an annual average temperature of 79° F (26° C) and an annual rainfall of 79 inches (2,000 millimetres), the climate is warm and superhumid. Apart from small areas of savanna (grassy parkland) on the northern borders, the equatorial rainforest of the Hiléia covers the whole country.

Among the plant life of Amazonas, the rubber tree has already been mentioned. The Brazil nut tree, which is extremely tall, is characteristic of the highlands. The rich soil of the *várzeas* or forested lowlands, a few miles wide along the banks of the main river, has a vegetation distinct from that of the hinterland *várzeas* and is subject to periodical flooding.

Animal life includes many endemic genera (families) and species—the species being unusually numerous and varied. Mammals are best represented by monkeys, bats, and rodents; birds by ant thrushes, parrots, toucans, and various marsh birds; reptiles by caymans (crocodilians, related to alligators), turtles, boas, anacondas, and iguanas; fish by characins (a species of tropical freshwater fishes) and catfish; insects—amongst many other sorts of invertebrate—by beetles, butterflies, and numerous Orthoptera (grasshoppers, crickets), Hymenoptera (bees, wasps, ants, and other social insects) and Diptera (flies, mosquitoes, gnats). The mosquitoes of the marshland are a menace to man and beast alike.

Population. Most of the people of the areas remote from the Amazon mainstream live in settlements on the banks of the other rivers. Nearly all of the rural population consists of caboclos—persons of mixed European and Indian descent. There is also a large group descended from immigrants from northeastern Brazil who arrived during the rubber boom. In 1970 there were about 42,000 Indians, representing 60 percent of the Indian population in Brazil. The Indian groups or tribes, of which about 30 can still be distinguished, have been progressively reduced in numbers by imported disease and by economic misery. Small Portuguese, Japanese, and Syro-Lebanese business communities also exist.

Population density

The overall density of population is less than 1.59 per square mile (four per square kilometre) and large areas of the state's territory are uninhabited. As many as 25 percent of the population are concentrated in Manaus. Careiro Island, a little way downstream from Manaus, is also quite thickly populated in comparison with the rest of the country. Manaus, still profiting from its privileges as a free port and the sole great city of the state, transmits something of its prosperity to adjacent areas.

The language of Amazonas is Portuguese, but the local vocabulary also incorporates many words from the Indian languages.

Roman Catholicism is the religion of the majority, but other Christian persuasions are also represented. The Indians preserve something of their original religions.

Administration and social conditions. The governor of the state, according to the constitution, is elected for a four-year term and is assisted by a vice-governor. The Legislative Assembly seats 30 deputies, also elected for four years. The judiciary comprises a court of justice, state judges, municipal judges, jury courts, and a system of military justice.

A constant struggle against disease is carried on in some 20 hospitals. Yellow fever, malaria, leprosy, and other tropical diseases are sporadic; and the average lifespan is only 38 years.

In the early 1970s, about 50,000 students attended elementary public schools, and 15,000 elementary private schools; there were about 15,000 pupils in public or private high schools. The Universidade do Amazonas, at Manaus, was founded in 1965.

Economy. The natural resources of the state consist almost wholly of the vegetation, the products of which form the basis of the economy: rubber, Brazil nuts, timber, guarana (a climbing shrub containing tannin and caffeine), vegetable oils, and fibres. These products are usually extracted by elementary methods, and very little processing is done in Amazonas, which sends much of the raw material to Belém in Pará state, to be processed for export. Known mineral deposits are few, and there is only one small iron foundry in operation.

For agriculture, short-cycle crops such as manioc, jute, sugarcane, bananas, and sweet potatoes are grown on the belts of land fertilized annually by the rivers. The yearly flooding of the rivers makes it difficult to raise livestock, but cattle are gradually being introduced to the higher lands. The plentiful fish and game available add protein to the people's diet.

Exports

Brazilian industry absorbs most of the raw materials produced by Amazonas, but rubber, timber, jute, vegetable oils, nuts, resins, aquarium fishes, and skins are exported to foreign countries.

Transportation. Transport is mainly by water, the rivers being navigable by large ships and canoes alike. There is, however, a land highway from Manaus to Itacoatiara; and the Transamazônica road was being built from east to west across the state in the 1970s, to link the Hiléia with northeastern and central Brazil.

BIBLIOGRAPHY. GASTAO CRULS, *A Amazônia que eu vi* (1930), observations made during a voyage on the Amazon, *Hiléia Amazônica* (1944), an illustrated work on flora, fauna, archaeology, and ethnography; JOSE FRANCISCO ARAUJO LIMA, *Amazônia, a terra e o homem* (1945), a study of the environment, people, history, and education; CHARLES WAGLEY, *Amazon Town: A Study of Man in the Tropics* (1953), a sociological study of the town of Gurupá, in Pará, where conditions are comparable to those in Amazonas; *Anuário Estatístico do Brasil*, vol. 30 (1969), statistical data; *Geografia do Brasil*, vol. 1, *Grande Região Norte* by ANTONIO TEIXEIRA GUERRA (1959), illustrations and maps on the climate, flora, fauna, geology, hydrography, industries, animal husbandry, agriculture, social organization, and transports.

(J.C.M.C.)

Amazon Rain Forest

The Amazon region (Amazonia) covers a territory of approximately 2,700,000 square miles (7,000,000 square kilometres), largely in northern Brazil between latitudes 2° N and 16° S. It is the greatest river basin of the tropical world, although the length of the Amazon is exceeded by that of the Nile. Its immense tropical rain forest contains mankind's biggest reserve of natural resources. The river dominates the ecological system of Amazonia, which is based on the circulation of water evaporated from the sea, through the air to the forest and its soil, and back through the river to the sea.

The mainstream rises in glacier-fed lakes of the Peruvian Andes, although it poses a geographical puzzle to say where it really begins. The first European to descend from Peru to the Atlantic Ocean was Francisco de Orellana, in 1541. Attacked by savages, he believed that the women of the tribe fought with the men against him, so he called the unknown river the Amazonas, after the female warriors of Greek mythology. The legend is still alive. At campfires, one can still hear tales about cities and cultures in which women warriors dominate, cities said to be hidden in the dense jungle. Today, the river is called Marañón (Spanish) or Maranhão (Portuguese) from its source to Iquitos, Peru; and Solimões from Iquitos to Manaus, Brazil. There it meets the Rio Negro and becomes the Amazon proper.

The vast Amazon Valley is characterized by lack of mountains. Alluviate soils form a moderately undulated plain. To the north lies the Guiana Highlands; to the west, the Andes; and to the south, the Brazilian central plateau. The entire area belongs politically to the Guianas (French Guiana, Guyana, and Surinam), Venezuela, Colombia, Peru, Ecuador, Bolivia, and Brazil. Half of it lies within Brazil and constitutes 40 percent of Brazil's total area, forming the states of Amazonas, Pará, Acre, and the federal territories of Roraima and Rondônia.

The forest. The forest stretches from the swampy mangroves and floating meadows (floating islands composed of intertwined land and water plants) in the east near the Atlantic, to the tree line on the Andes. High in the Andean region, there are large and small meadows (paramos; Spanish *páramos* or *paramillos*) of grass, sometimes with gigantic cacti. The flora in dry areas is distinctly different from that in swampy locations. On lower levels, swamps with arborescent calla-like Araceae (Montrichardia) are common. The Andean forest gradually turns into tropical rain forest, which can be divided into three main types: forest, the ground level of which is not inundated during the rainy season; areas that are regularly inundated; and areas that are occasionally inundated. One can frequently paddle a canoe between the trunks of tree species adapted to such nearly amphibious living. On the Brazilian central plateau, the forest proper is limited by a parklike vegetation of grasses and small to medium trees and bushes, characterized by twisted trunks, thick bark, and leathery leaves. Many of them have big subterranean woody trunks, and roots that reach deep into the groundwater. Where the trees leave sufficient space for a jeep to pass, the vegetation is called *cerrado*. Nearer to the streams, the trees grow taller and denser until they transcend into the forest proper. Pampas (open plain) sometimes suddenly occur.

The three types of forest area

Relief and drainage. The Amazon Valley is like an immense canyon opening into the Atlantic Ocean with a mouth over 250 miles wide. The tributaries, especially, those from the Brazilian Highlands (Planalto Central), meander with relatively slow current and few rapids for thousands of miles before descending great rapids to the level of the valley. They drain four-tenths of South America and twice as much as any other drainage area. The Amazon carries an enormous amount of water and is of extraordinary depth. Near Óbidos a depth of 300 feet has been recorded and, near Manaus, 330 feet. In spite of the low average gradient of about 0.6 inch per mile, the current speed is relatively high. The average discharge into the Atlantic has been determined at 6,350,000 cubic feet per second—four times the discharge of the Congo and 10 times that of the Mississippi (See also AMAZON RIVER.)

Soil types. The appearance of the forest vegetation is extremely exotic, leading to the erroneous conclusion that the underlying soil must be very fertile. The luxurious growth, however, stems only from the fragile ecological balance between soil and plants, influenced by light, temperature, and water. Generally, the soil free from flooding is well drained, porous, and of variable structure. Often it is sandy and of low to medium natural fertility because of its lack of phosphate, nitrogen, and potash and relatively high acidity. Small areas are underlain with basaltic and diabasic rocks, which are of high natural fertility. In Brazil these fertile soils are called *terra roxa* (reddish soil) and *terra preta dos Indios* ("black earth of the Indians"). The potentiality of the semiflooded areas is great; their soils do not lack nutrients, but use for agricultural purposes is limited by the periodic inundations.

Types of vegetation. The floral province of Amazonia is characterized by the occurrence of rubber-bearing *Hevea* species and *Gnetum* but exhibits many local variations. Four general floral subprovinces may be discerned: (1) the upper Rio Branco, up to the Guianas, with predominantly grasslands and patches of forest; (2) the Jari-Trombetas river area, with varied vegetation of semideciduous, dry forest and *cerrados*; (3) the Rio Negro area, which, despite its heavy rainfall, contains much desert-like sclerophilious vegetation (plants with thick leaves, resistant to water loss). This subprovince comprises the Rio Negro, the upper part of the Orinoco, the lower and middle Japurá, and the western part of the Trombetas Basin; (4) the rain forest of the Amazon Basin, extending from the base of the Andes to the Atlantic Ocean and up to some parts of the northern boundary of the Amazon Valley.

The variety of trees is enormous; 117 species have been counted in an area of half a square mile. The most frequent varieties are myrtles, laurels, acacias, bignonias, cedrelas, cecropias, rosewoods, Bombacaceae, Brazil nuts, rubber trees, figs, and palms. The only gymnosperms are climbing Gnetums and rare coniferous trees or bushes of the evergreen podocarpus.

Four levels of forest vegetation are distinguishable. The floor level is often covered by a mosaic of small herbs and ferns. The second level is formed by bushes and taller herbs—heliconias, hirtellas, eugenias, calliandras, and others. The third level contains shade-loving trees, such as mauritia, orbygnia, and euterpe. These trees are sometimes covered by orchids, ferns, bromeliads, cacti, and pipers growing on them. It is quite common to find dense groups of palms in the middle of the forest.

The four growth levels of the rain forest

Higher trees constitute a fourth, or ceiling, level. Generally, these forest giants—such as the silk-cotton, or sumaumeira (*Ceiba*), the para nut, or castanheiro (*Bertholletia*), the sapucaia (*Lecythis*), and the sucupira (*Bowidchia*) with its blue flowers—stand alone and are conspicuous against the skyline.

Beyond the margin of dense forest, the first and second level almost disappear, and one can walk between the trunks of trees and hanging lianas without great difficulty. The diffused lighting prevents the growth of thick underbrush. As the sun cannot be seen, it is easy to lose one's bearings. Few people have dared to penetrate the forest, and some have been lost on short trips. Even the Indians avoid entering the woods too far from their settlements near the riverbanks. They are afraid of legendary inhabitants of the dense forest and of becoming lost.

Fauna. To give an impression of the complete fauna of the Amazon is as impossible as it is to give one for the flora. More than 8,000 species of insects have been collected and classified. Myriads of mosquitos bother the traveller and may transmit malaria and yellow fever. The leaf-cutting ants (*Atta* and *Acromyrmex*) in a short time can destroy large parts of a plantation. The extermination, or at least the limitation, of these pests is of great importance to further habitation of the area. The most troublesome insects of all are the ubiquitous, small, black flies, called *pium*s, whose bite can itch for days. Fireflies, stinging bees, hornets, wasps, beetles, cockroaches, cicades, centipedes, scorpions, ticks, red bugs, giant spiders, and butterflies are abundant. Sometimes hundreds and thousands of butterflies gather in the afternoon on humid sand near riverbanks or on islets, sipping moisture and often covering several square yards.

Fishes

The number of fish species is estimated at about 2,000. Some of the smaller ones, such as the neon tetra and discus, are wonderful aquarium ornaments. The best known of the larger ones are the pirarucu, the giant catfish, the stingray, and the electric eel. Piranhas have the reputation of being very bloodthirsty and aggressive, but many deaths are due to attacks of giant catfish in deep water, where piranhas only finish the victim off.

Other peculiar water animals are the manatee (sea cow, common in lakes), freshwater dolphins, caimans, turtles, and the anaconda (giant water snake), which easily grows to 20 feet in length. Turtles and their eggs furnish one of the principal food supplies. Parrots, toucans, hangnests (or caciques), ducks, doves, geese, perdizes (tinamous), herons, cormorants, gulls, ovenbirds, roseate spoonbills, scarlet ibises, and the *jabirú* (or great South American marabou stork), belong to the exceedingly rich birdlife of the forest. Hummingbirds are more frequent at higher altitudes than in the valley.

Frogs and toads abound. Poisonous snakes are not numerous but are feared. Of the nonpoisonous species, the boa constrictor reaches huge proportions but seldom grows as large as the giant anaconda. Monkeys of many species are hunted and eaten by the Indians. Big cats, such as jaguar and ocelot, are rare, although pumas may be found in larger numbers in the Andes. Several other carnivores, such as the *cuatí* (coati), *grisons* and weasels, are present. Rodents are numerous. *Queixadas* (peccaries) and, largest of all, *antas* (tapirs), are quite common where man has not extinguished them by merciless hunting. The same is true of armadillos, deer, and the giant anteater. Bats flourish by the millions.

Population. The population of the Brazilian Amazon was reported as 3,000,000 (1970 census). The population of the entire area probably does not exceed 5,000,000. Approximately 1,000,000 live in towns. The average density is less than two persons per square mile.

The number of Indians can only be guessed but they are believed to number no more than 100,000. The Indians of the forest live as fishermen and hunters, practicing nomadic agriculture on the riverbanks. Some tribes are still considered wild (*i.e.*, aggressive), but pacification and acculturation of their Stone Age way of living is speedily advancing. By 1980–90 little or nothing is expected to be left of their original way of life.

The Chibcha- and Kechua-speaking peoples have a higher cultural inheritance, which they have been able to maintain. Numerically, the aborigines form a minority against the descendants of Portuguese and other Europeans, Africans, and Japanese.

Exploration and transport. In early days the river was the only means of access into the forest. As mentioned above, Orellana was the first to travel on the mainstream from the Peruvian Andes to the Atlantic (1541). Nearly a century later, Pedro Texeira went from Belém to Quito, Ecuador and the region became more and more known through the explorations of the Portuguese. About 1751 Charles de La Condamine made the first astronomic–geographic survey and brought the deadly Indian arrow poison curare to Europe. At the beginning of the 19th century, the German explorer Alexander von Humboldt, accompanied by the French botanist Aimé Bonpland, mapped the connection between the Amazon and the Orinoco system through the Brazo Cassiquiare (Cassiquiare River), whose existence was insufficiently known before.

A number of scientists explored the region. Louis Agassiz, the Swiss-born United States naturalist, investigated its animal life. Many botanists, whose results benefitted the famous *Flora Brasiliensis* of Karl von Martius, studied the vegetation. Scientific exploration has been intensified, especially by the United States National Geographic Society, the Brazilian Petrobrás Company, and research institutes in Manaus and Belém, Brazil.

River transport and aviation

The first steamship line was authorized by the Brazilian emperor Dom Pedro II in 1850. One can get from Belém to Manaus and up to Iquitos in Peru by steamers; Manaus is served by ocean-going vessels. At the beginning of World War II, Brazilian civil aviation began to link Manaus, Belém, Óbidos, Pôrto Velho, and other cities to the rest of the continent. Air transport is of major importance to the region.

The location of the Brazilian capital, Brasília, near the southern fringe of Amazonia has inaugurated a new era. A road has been built from Brasília to Belém; and one from Cuiabá to Santarém is nearly finished. Pôrto Velho can also be reached by car from Cuiabá by a road that will eventually link that town to Manaus and Peru. The biggest enterprise of all is the construction of the Transamazonian Highway. Together with the neighbouring network of roads, it will link the Amazon with the rest of South America.

The economy. The resources of the Amazon are many and varied. The first explorers were attracted by gold and diamonds, and industrial diamonds are still mined on the Rio Araguaia. Precious gems are occasionally found there and elsewhere. The most important game animals, used and traded for human nutrition, are: the manatee (or sea cow), which is menaced by extinction; various turtles and their eggs; the tapir; the red deer; the capybara; and, among the Indians, several types of monkeys. Wild forest birds and waterfowl are common game. Fish is also an important food source, especially the huge pirarucu (*Arapaima gigas*). Its flesh is dried in the sun and sold at every marketplace as a durable staple food.

Timber and other forest resources

Excellent timber is furnished by the mahogany (*Swietenia macrophylla*), the Amazonian cedar (*Cedrella odorata*), and many other species. Other trees, such as the cumaru (*Coumoruna odorata*), furnish perfumes and medical ingredients. The louro-inhamuí (*Ocotea barcellensis*) produces a liquid that can be used as a fuel for internal-combustion engines. The big sorva (*Couma macrocarpa*) exudes a drinkable milk sap. The economic kings of trees, however, are the rubber tree (*Hevea brasiliensis*) and the para nut (*Bertholletia excelsa*). The rubber tree has been one of the most important objectives of penetration and exploitation of the forest. It gave rise to a period of great but temporary prosperity, especially to the city of Manaus. Grown in plantations, it continues to make an important contribution to the economy.

Other nutritional sources include the heart of palm (*Euterpe oleracea*), the vanilla orchid, and the guarana (*Paullinia cupana*). Guarana, black pepper, and vanilla have become important agricultural crops. Guarana seeds are used to prepare a stimulating Brazilian soft drink, which may find world markets. Rice is cultivated in some areas. Agriculture and animal husbandry must be conducted carefully and scientifically, however, to preserve the very delicate ecologic balance that sustains the fertility of the soil.

Geological exploration is producing new knowledge of mineral resources. Oil has been found, and tin mining has been carried on for some time. Other valuable minerals are also extracted.

The greatest challenge to Brazil and the surrounding nations is to integrate Amazonia without destroying its future. Brazil is making a strenuous effort in this direction through fiscal incentives, mobilization of public opinion, development projects, scientific research, and road construction. It is possible that Amazonia may become a major source of raw materials for the entire world.

BIBLIOGRAPHY. The *Anuário Estatístico do Brasil* is the official Brazilian yearbook of statistics, containing much information pertaining to Amazonia. HERMAN LENT (ed.), *Atas do Simpósio Sôbre a Biota Amazônica* (1967), is the most important modern collection of original scientific papers in Portuguese, Spanish, and English on the geology, anthropology, limnology, botany, zoology, and conservation of the Amazon, with excellent bibliographies. H.W. BATES, *The Naturalist on the River Amazons*, 2 vol. (1863; 3rd ed. abridged, 1873, reprinted 1962); and A.R. WALLACE, *A Narrative of Travels on the Amazon and Rio Negro*, 2nd ed. (1889, reprinted 1969), are both historically interesting. P. LECOINTE, *L'Amazonie brésilienne*, 2 vol. (1922), is a classic account of Amazonia as it was 50 years ago. H. SCHULTZ, *Isto é a Amazônia!* (1964), comprises a collection of excellent photographs of several Amazonian topics. The same author's *Hombu: Indian Life in the Brazilian Jungle* (1962), is perhaps the best modern collection of original photographs with scientific text concerning the Amazonian Indians; practically indispensable for readers with ethnographical interests.

(A.R.S.)

Amazon River

The Amazon (Portuguese, Amazonas) is the greatest river of South America and the largest river in the world in volume and in area of its drainage basin. Although the name Amazon is popularly employed for the whole of the main stream, in Peruvian and Brazilian nomenclature it is properly applied only to sections of it. In Peru, the upper stream (fed by numerous tributaries flowing down from sources in the high Andes) down to Iquitos, is called Marañón (Portuguese, Maranhão) and from there to the Atlantic, Amazonas. In Brazil the name Solimões is used from Iquitos to the mouth of the Rio Negro and Amazonas only from the Rio Negro to the sea.

The first European to explore the Amazon River, the Spanish soldier Francisco de Orellana, gave it its name after reporting pitched battles with tribes of female warriors. The length of the great river that he partially explored is almost 4,000 miles, slightly less than the Nile but still the equivalent of the distance from New York City to Rome. Its westernmost source is high in the Andes, within 100 miles of the Pacific Ocean; and its mouth is in the Atlantic. It is estimated that about one-fifth of all the water that runs off the Earth's surface is carried by the Amazon. At its mouth, 150 miles wide, there is a mean discharge of 170,000,000,000 gallons of water per hour into the ocean. This is some ten times the

amount carried by the Mississippi River, and it turns the ocean's water from salt to brackish for over 100 miles from shore.

THE NATURAL ENVIRONMENT

There are more than 1,000 known tributaries of the Amazon that flow into it from the Guiana Highlands and from the Brazilian Highlands, as well as from the Andes. Seven of these tributaries—the Japurá (Caquetá in Colombia), the Juruá, the Madeira, the Negro, the Purus, the Tocantins, and the Xingu—are more than 1,000 miles long; and one, the Madeira, is more than 2,000 miles from source to mouth. The largest ocean-going steamers can ascend the river 1,000 miles to the great city Manaus. Freighters and small passenger vessels drawing 18–20 feet of water can sail to Iquitos, some 2,300 miles from the Atlantic, at any time of the year.

The Amazon Valley

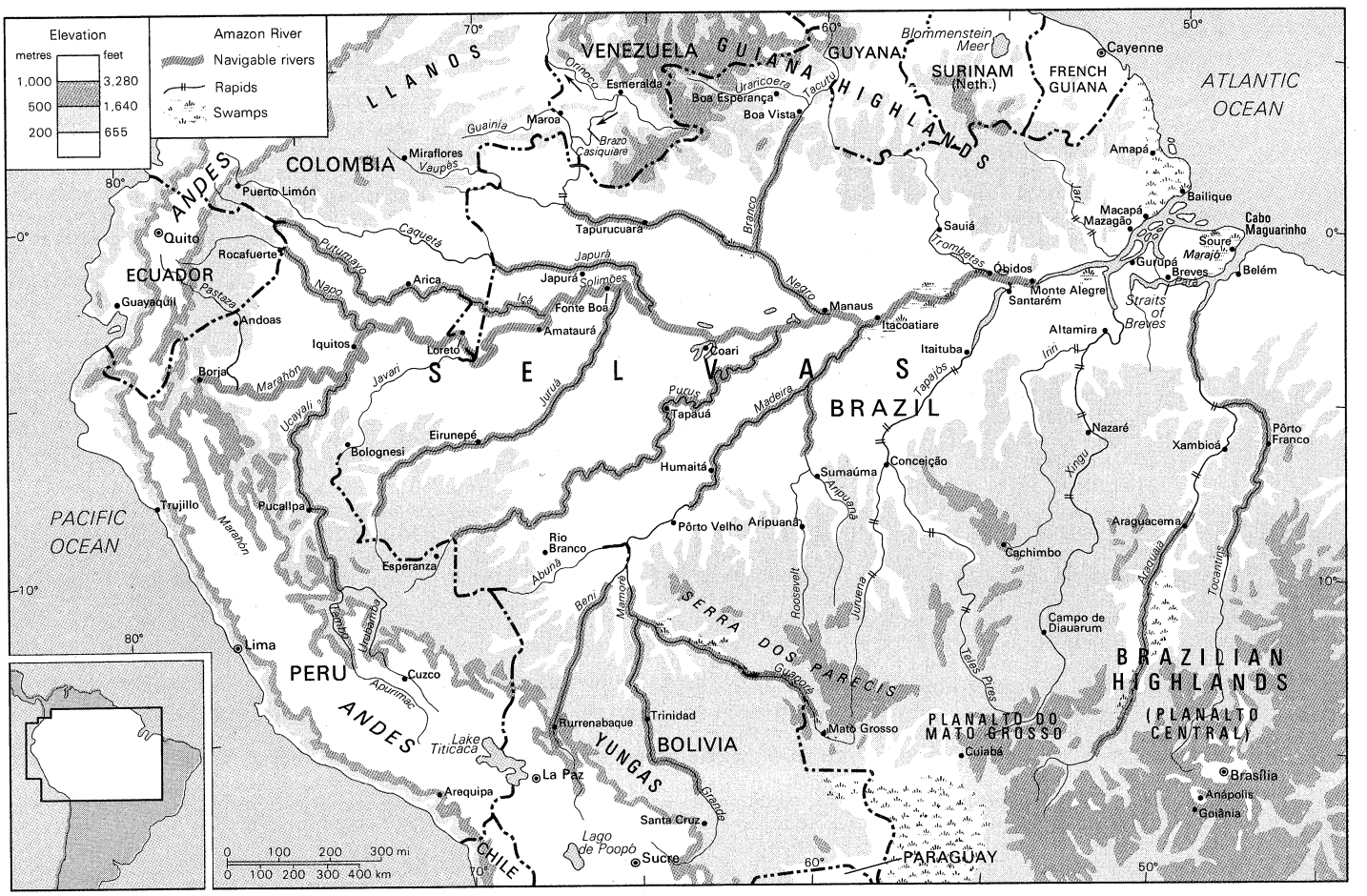
The Amazon Valley is the largest lowland in Latin America. The drainage basin of the Amazon, about 2,722,000 square miles (7,050,000 square kilometres) in extent and nearly twice as large as that of any other river, stretches over 25 of the Earth's 180° of latitude. It includes the greater part of Brazil and parts of Venezuela, Colombia, Ecuador, Peru, and Bolivia. The Amazon Valley is not an immense swamp; most of Amazon country is terra firma, well above the level of the annual inundations. Going by the river from Belém to Manaus, the tidal forests of the delta are soon left behind; after leaving the Straits of Breves, ranges of hills several hundred feet high begin to appear in the north. At Monte Alegre the hills come down almost to the water's edge; beyond them extends a wide expanse of open *campos*, or grasslands, bounded by ridges along their northern edge. These natural *campos* are characteristic of the Amazonian landscape in many places—the largest of them, the *campos gerais* of the Rio Branco, far to the west, being several thousand square miles in extent.

The Amazon over the millennia has meandered freely over a wide flood plain, on which there appear a series of meander scars, oxbow lakes, and recently abandoned channels. Whenever the deposition of silt or alluvium is great enough to lessen the strength of the main current of the river, it will, during high water, overflow its temporary natural levee and carve out another channel. The new channel will in turn gradually silt up over a period of years or even decades, and the stream will then again shift its course.

The areas that are annually flooded, the *várzeas*, are rejuvenated each year by the deposit of fertile silt left as the waters recede. It is estimated that these valuable soils occupy some 25,000 square miles, nearly double the area of The Netherlands.

Climate. The climate of Amazonia is warm, rainy, and humid. The length of day and night is equal on the Equator (which runs only slightly north of the Amazon River), and the usually clear nights favour relatively rapid radiation of the heat received from the sun during the 12-hour day. There is a greater difference between the noonday and midnight temperatures than between the hottest and coolest months. Hence, night is the winter of the Amazon. To be sure, there are occasional cooler periods, especially during the Southern Hemisphere's winter months when an especially large and powerful air mass from the south polar region pours out over the Amazon Plain and causes a sharp drop in temperature, known locally as a *friagem*, or cold snap. At any time of the year several days of heavy rain will be succeeded by clear sunny days and fresh nights, when the temperature-humidity index is low. In the lower reaches of the river, cooling trade winds blow most of the year.

More important than temperature is the amount of precipitation. Moisture-laden winds blowing from the Atlantic Ocean across South America are forced to rise when they reach the Andes, where they are cooled and, in cooling, lose their moisture through condensation to feed the great rivers that flow east from the Andes and



The Amazon River Basin and its drainage network.

form part of the Amazon River system. In the lowlands themselves, extensive convectional storms account for much heavy precipitation. There are at least three distinctive climatic types or subregions within Amazonia, based on the rainfall regime: (1) Near the mouth of the Amazon and the western portion of the plain there is an average year-round precipitation of more than 80 inches, fairly well distributed throughout the year. During some years, twice this amount has been recorded, whereas in others there have been prolonged dry seasons. (2) In the rainfall regime that prevails over most of Amazonia there is a season of noticeable slackening of precipitation, although the decrease in rainfall is not severe enough to hinder plant growth. (3) Along the southern edge of the region, the climate of Amazonia grades into that of west central Brazil, with a more distinct dry season during the southern winter.

Three climatic regions of Amazonia

The prevailing winds blow between east-northeastward and east-southeastward during the dry season. These winds are moderate in July and August but fresh during the rest of the season, when gusty winds, known locally as *marajós*, sometimes reach great force. This season is the best time for ascending the river. Sailboats descending the river drift with the current.

Normally, the influence of the tide is felt as far as Óbidos, situated about 600 miles above the river mouth. A tidal bore, locally called *pororoca*, occurs at times in the Amazon Estuary prior to spring tides; it consists of a wave from five to eight feet high, the crest of which breaks and spreads over the shallow waters of the river and its tributaries.

The forests of Amazonia

Soils and vegetation. From an airplane, the great treescape of central Amazonia looks like a green carpet with holes or worn spots in the form of oxbow lakes, bright green sloughs, and extensive swamps. There are giants of the forest with white trunks that stand out like clumps of broccoli, flowering trees of many hues, and a great variety of palms. There is in the forests of Amazonia a bewildering diversity of species. There are no pure stands, and lumbering has taken on the characteristics of prospecting. Vast areas have been prospected for hardwoods, and those trees near the rivers have in many places been harvested. It is anticipated that new techniques in the making of paper, plastic, and cellulose may result in the cutting over of large areas of forest, regardless of composition.

The assai palm produces berries from which refreshing drinks and sherbets are made. Various palms are tapped for their sap, which is a potent wine when fermented. The heart of palm, or palm cabbage, produced by several forest palms is a fine-textured vegetable with the best features of lettuce and cabbage. Palm nuts are regularly gathered to be used as chicken and hog feed or for the extraction of oil. A delicious fruit, *uvas silvestres* (literally, "wild grapes"), about the size of a plum, is also harvested.

In the 19th century there was a burgeoning rubber industry, an economy based on the collection of a valuable product from a variety of wild trees native to the Amazon—chiefly the Para rubber tree or *Hevea brasiliensis*, a tall tree of soft wood with high-branching limbs and a large area of bark. When the trees were domesticated in the Far East in the late 19th and early 20th centuries, however, the product from the wild trees of the Amazon could not compete, and the boom collapsed.

The soils of a large part of Amazonia's uplands appear to be capable of crop and livestock production. Once it becomes economically feasible to introduce efficient land-management practices and to add chemical or animal fertilizers, areas at present in bush may pass into the arable class. No less important than native-soil fertility are legal access to the land, proper financial and technical assistance, and transportation facilities for the movement of marketable produce. In the seven years from 1960 to 1967, for example, scores of fattening pastures were established in the newly created municipality (*município*) of Paragominas alone in the forest lands of Pará state (*estado*) that were opened up by the Belém–Brasília highway.

The vast reserve of fishes

Fauna. *Fishes.* The Amazon and its tributaries, together with the swamps and oxbow lakes, are in reality a vast sea of freshwater, most of it slowly flowing, in which live millions of fish. One of the best fish, reputed to be the biggest freshwater fish in the world, is the pirarucu, taken with a harpoon, not on rod and reel. It is a welcome source of protein to river dwellers. Thousands of tons of giant catfish, *gamitama*, *paitsche*, and many others are to be seen in the markets along the river. One of the most dangerous of river fish is the small, eight- to ten-inch piranha, ferocious and sharp-toothed, which will often attack any animal that enters the water, or even man; its razor-sharp teeth cut out chunks of flesh, and a school of these fish can do lethal damage in a short time. With the rapid means of transportation afforded by jet airplanes, there also are world-wide markets for tropical fish, such as silver carp, neon tetras, and even piranhas, as well as for aquatic plants.

Electric eels, giant water snakes such as the anaconda, and freshwater stingrays are native to many sectors of the vast river system. The flesh of the stingray is coarse-fibred but very palatable. Crocodiles are hunted for their skins, and turtles and their eggs are considered a delicacy.

Mammals. The capybara, largest rodent in the world, looks like a giant guinea pig and may weigh up to 100 pounds; it lives along the Amazon and its tributaries, feeding on the grasses, reeds, and rushes that grow on the banks, and taking refuge in the water when attacked; it is an excellent swimmer and diver; its flesh, though not considered a delicacy, is used for food and is frequently dried for shipment to market; its hide is made into very good leather.

The manatee, often called the sea cow, attains a length of six to eight feet, and a large specimen may weigh up to 2,000 pounds or more. It is hunted for its flesh and its oil, but it is so rare as to be almost extinct. Great efforts are being made at present to capture the animal alive, for shipment to zoos around the world.

The nutria, or tropical otter, is also to be found in less frequented areas and is valuable for its pelt. The tapir, which attains a weight of several hundred pounds, is an expert swimmer and diver and is much sought after for its excellent flesh. The peccary, or wild pig, and many kind of deer are natives of Amazonia.

Water buffalo, brought to the Ilha de Marajó (Island of Marajó) many years ago, have gone wild and now live in the more remote, swampy parts of that ranching area. The Brazilian government has recently introduced domesticated water buffalo and is studying them at an experimental station not too far from Santarém. They seem ideally adapted to such an area and may in time become the basis for a dairying industry, as they are in the brackish inlets and swamps north of Naples, Italy.

Arboreal fauna. One of the most common animals is the small, agile squirrel monkey, in great demand for use in laboratories. There are howler monkeys, woolly monkeys, capuchin monkeys—those that so often accompanied the organ-grinder—and a host of others. The decapitated bodies of monkeys, split and blackened over a charcoal fire, are used extensively for food and are frequently seen for sale in local markets.

In more remote areas is found the sloth, a harmless animal that spends its life hanging from the limbs of trees, the leaves of which are its favourite food. The iguana lizard, often attaining a length of some five feet and looking like a prehistoric monster, climbs trees with the facility of a squirrel and is much sought after because of its delicious flesh.

MAPPING THE AMAZON

Early exploration

Since the great voyage of discovery by Orellana in 1541–42, the Amazon has been the object of many surveys, both reconnaissance and detailed. The drainage basin is so large that many travel books as well as detailed monographs have been written about various parts of it. The German naturalist Alexander von Humboldt wrote of conditions along the Rio Negro. The English naturalist Henry Walter Bates spent the years from 1848 to 1859 along the Amazon, collecting thousands of species of

animals and writing up his notes of animals, human beings, and natural phenomena in a charmingly objective manner. His book, *The Naturalist on the River Amazons*, originally published in two volumes in 1863, was an immediate success and is one of the great classics on the Amazon River. An official United States expedition was sent to Amazonia in the middle of the 19th century; in 1854 in Washington William Lewis Herndon published as a public document the report that he and Lardner Gibbon, lieutenants in the U.S. Navy, had made to Congress under the title of *Exploration of the Valley of the Amazon*. They reported in fascinating detail the possibilities for power navigation in various parts of the river system, the resources available, health problems, social conditions, and possibilities for further settlement and development.

The 20th century has been a period of exploratory and scientific expeditions. In 1913–14, former U.S. president Theodore Roosevelt and Brazilian Col. Cândido Rondon headed an expedition that explored the Rio Roosevelt (originally Río Teodoro), a tributary of the Madeira, and made natural-history collections and observations. An expedition sponsored by Harvard's Institute of Geographical Exploration did important scientific work between the years 1910–24. The American Geographical Society and the American Museum of Natural History have undertaken geographic explorations and sent collecting expeditions into Amazonia over the years. The American Geographical Society compiled data for and published 1-to-1,000,000 scale maps of this vast area. These are only some of the highlights of exploration during the first three decades of the 20th century. Hardly a year passes that enthusiastic scientists and adventure-seeking hunters do not write books on this area.

During World War II the United States government sent dozens of American scientists and technicians into Amazonia to make regional surveys and to study the actual physical and human resources as well as to draw up plans for future development. Since the war, expeditions have been sponsored both by the American Geographical Society and by various United Nations organizations. Brazilian scientists have increasingly concerned themselves with the Amazon region. This is especially true of scientists associated with the Instituto Agronômico do Norte and the Museu Paraense Emílio Goeldi (Goeldi Natural History and Ethnographical Museum), both located in Belém, the Conselho Nacional de Geografia, and the Instituto Nacional de Pesquisas da Amazônia, established in 1952, in Manaus.

Increasing Brazilian surveys and explorations

HUMAN ECOLOGY

The aborigines. At the time of the Conquest, the bottom lands of the Amazon and its major tributaries supported relatively dense, sedentary populations who practiced intensive root-crop farming, abetted by fishing and by the hunting of aquatic mammals and reptiles. The *campos*, the higher areas between the rivers and their floodplains, were, and still are in some of the more remote sectors, inhabited by small, widely dispersed nomadic tribes of Indians. These groups rely largely on hunting of animals, large and small, and the gathering of wild fruits, berries, and nuts, and on some small patch agriculture of low yield.

The Amazonian Indians devised means of making the poisonous bitter manioca edible, the end product, *farinha*, being a food staple widely used today over most of tropical America. Amazonian Indians had perfected the use of quinine as a specific against malaria and extracted cocaine from the leaves of the coca tree. They also collected the sap of the Para rubber tree. The Amazonian Indians were skilled rivermen in their dugout canoes and had invented and used the hammock, an ideal bed in which to sleep in the humid tropics. Their ancient arrow poison, curare (*Chondodendron tomentosum*), has been used in the therapy of a host of paralyses and spastic disorders, such as multiple sclerosis, Parkinson's disease, and St. Vitus's Dance.

The early explorers of the Amazon, such as Orellana and Orsua and Aguirre, provisioned themselves from the food supplies of the Indians they met, commandeered their canoes, and took into slavery those that they wanted. The result was a complete breakdown in native life and an appalling decrease in the Indian population. As late as 1906 there were reports of the wholesale capture of wild Indians who were enslaved in order to tap rubber, which was plentiful, with a high price on the world market but which was difficult to get at because rubber trees were sparsely scattered over a huge area, and increasing toil was necessary to locate and work them. Women and children were wantonly butchered, as were any men who became ill or who could not gather their assigned quota. More recently, in the late 1960s and the 1970s, there have been further reports of oppressive treatment of Indians by traders and land speculators.

Trade and modern populations. The vast network of rivers is today organized under the control of a hierarchy of traders. Since travel away from the rivers is still extremely difficult, the people scattered in small communities along the riverbanks are dependent for their connections with the outside world on the launch of the trader to whom they are in debt. The largest trading companies are at Belém and Manaus; many smaller traders are controlled by the big companies, each of which occupies a strategic point at a river junction and controls all upstream communications. From many small villages comes a considerable movement of goods downstream, including such products as Brazil nuts, rotenone, cabinet woods, skins, and some quantity of rubber. In return, the traders supply foodstuffs, tools, and firearms.

Dominance of trading companies

The Amazon Basin is one of the great demographic vacuums of the world. The Brazilian government estimated the population of the six federal units making up the Brazilian Amazon to be 3,518,000 in 1970, of which about 565,000 lived in Belém and another 250,000 in Manaus and vicinity. The population of that part of Amazonia falling within Bolivia, Peru, Ecuador, Colombia, and Venezuela probably does not exceed 400,000, with 74,000 of these in Iquitos, capital of the Peruvian oriente. The relative emptiness of the Amazon has intrigued people for many years, and the increasing effectiveness of malaria control, improved diets, and sanitation measures, and the increased ease in transportation are all making the Amazonian region more attractive for human settlement.

The construction of new highways and the immediate development of agriculture along them, the construction of factories and an oil refinery, and the recent development of the tourist industry, all have injected a new dynamism essential to the economic development of the Amazon and its vast hinterland.

Relations between Amazonia and other sectors of Brazil. The opening of highways between Brasília and Belém and between Brasília and Pôrto Velho has meant that manufactured goods from southern Brazil arrive with fewer delays and in many cases at freight rates much lower than formerly, when shippers had no competition. New settlements and agricultural developments have appeared along these roads and are themselves generating new passenger and commodity traffic. The new road under construction between Manaus and Pôrto Velho will doubtless have a profound impact on Amazonia and Manaus when it is completed. Along the relatively short road (170 miles) from Manaus to Itacoatiara, Japanese colonists have organized agricultural cooperatives and have introduced modern agriculture techniques to produce vegetables, chickens, and eggs for the local markets, and black pepper for export to southern Brazil and the world market. The federal government is allocating millions of cruzeiros to provide lines of communication and supply, mainly by river, for the construction of the Transamazonian and Cuiabá–Santarém highways. Furthermore, plans are being made and moneys allocated to resettle thousands of families from drought-stricken northeastern Brazil to farms along these new highways.

The importance of new highways

The Manaus market is short of fresh beef most of the year, yet there are an estimated 285,000 head of cattle

grazing on the savannas of Roraima, some 300 miles to the north. A road will soon be finished that will give these grasslands access by truck to Manaus. The Roraima ranchers, centred in the Boa Vista area, are well aware, however, that roads are being constructed to Manaus from the south which will put that city closer to more advanced sectors of the country, and that along these roads forest lands will be cleared for fattening pastures that will be much closer to Manaus than are the grazing lands of Roraima Territory. Roraima ranchers are already making contact with buyers in the potential markets of Venezuela, Guyana, and Surinam.

The highway engineers seek to build the north–south roads as much as possible on the higher, better drained land between the rivers tributary to the Amazon. The Manaus–Pôrto Velho highway follows in general the interfluve area, frequently grass-covered *campos*, between the Purus and Madeira rivers. The Cuiabá–Santarém highway will follow for the most part the higher interfluve area between the mighty Tapajós and Xingu rivers. The east–west road, however, cuts across the valleys of the north-flowing rivers. The first stretch of the TransAmazonian Highway, from Pôrto Franco in Maranhão to Itaituba in Pará, some 850 miles, was completed in early 1972. The second stretch of some 600 miles, from Itaituba to Humaitá in Amazonas, on the Rio Madeira, was finished almost a year later.

Development of the tropical frontier

The future. For the overall development of Amazonia it has been suggested that (1) the road building program be continued; (2) regular service of shallow-draft diesel-powered barges on the rivers be introduced, to eliminate exorbitant charges of middlemen; (3) property frontage per owner on all roads and rivers be restricted in order to insure more rapid spatial spread of settlement in and intensive development of the hinterland; and (4) the central government provide at least minimal educational facilities and public-health measures, as well as short-term credit and property titling services for the agricultural population.

The vast tropical lowlands of Amazonia are a frontier zone that by its very nature is not yet able to live a life of its own, somewhat like the United States Middle West a century ago. As long as the Middle West led a kind of colonial existence vis-à-vis the Eastern Seaboard, it could not work out its own regional salvation. Once it was settled by an industrious, agricultural people and was crisscrossed with railroads and motor roads for intraregional as well as interregional trade, it could and did develop its own complementary industrial society. With the introduction of adequate roads and transportation facilities, education, and public-health measures and permanent rather than nomadic agriculture by a vigorous people, the Amazonian lowlands may indeed experience an evolution that is similar in many respects.

BIBLIOGRAPHY. EARL PARKER HANSON, *South from the Spanish Main: South America Seen Through the Eyes of Its Discoverers* (1967), almost one-third of this excellent volume devoted to a discussion of various facets of Amazonia; ADRIAN COWELL, *The Heart of the Forest* (1960), the story of an expedition to Indian tribes in the Xingu river basin; GILBERTO FREYRE, *New World in the Tropics: The Culture of Modern Brazil* (1959), a discussion of the historical factors that are playing a significant role in incorporating the Brazilian part of the Amazon basin into industrialized Brazil; LILO LINKE, *People of the Amazon* (1963), an account of a journey in which the author followed in the footsteps of the 16th-century Spanish explorer Francisco de Orellana; YVES MANCIET, *Amazonie, terre inachevée* (1961; Eng. trans., *Land of Tomorrow: An Amazon Journey*, 1964), a highly readable account of a voyage up the Amazon; HELEN and FRANK SCHREIDER, *Exploring the Amazon* (1970), a bird's-eye view of the river and its basin, with colour illustrations.

(R.E.Cr.)

Ambartsumian, Viktor A.

Viktor Amazaspovich Ambartsumian, a prominent Soviet astronomer, is best known for developing a theory concerning the origin and evolution of stars and stellar systems. He is also the founder of the school of theoretical astrophysics in the Soviet Union.



Ambartsumian, 1970.
Tass—Sovfoto

Ambartsumian was born on September 18, 1908, in what is now Tbilisi, Georgia, U.S.S.R., into an Armenian family. His father, a teacher of literature, encouraged the development of his outstanding aptitude for mathematics and physics. In 1925 he entered the University of Leningrad with the intention of devoting his life to research in astrophysics and in the following year published a paper on solar activity, the first of ten papers he published while an undergraduate. After graduating in 1928, he became a graduate student in astrophysics under A.A. Belopolskii at Pulkovo Observatory near Leningrad.

From 1931 to 1943 he lectured at the University of Leningrad, where he headed the Astrophysical Department. In 1932 he advanced his theory of the interaction of ultraviolet radiation from hot stars with the surrounding gas, a theory that led to a series of papers on the physics of gaseous clouds. His statistical analysis of stellar systems in 1934–36, in which for the first time their physical properties were taken into account, was found to be applicable to many related problems, such as the evolution of double stars and star clusters. These accomplishments were recognized with his election as a corresponding member of the Academy of Sciences of the U.S.S.R. in 1939 and his appointment as deputy rector of the University of Leningrad in 1941–43. His theory of the behaviour of light in a scattering medium of cosmic space, put forward in 1941–43, became an important tool in geophysics, space research, and particularly astrophysics, such as in studies of interstellar matter.

Research on star formation

In 1943 Ambartsumian joined the Armenian Academy of Sciences in Yerevan, the capital of Soviet Armenia, and began teaching at Yerevan State University. In 1946 he organized the construction near Yerevan of the Byurakan Astronomical Observatory, now one of the most important observatories in the U.S.S.R., where he began another successful period of activity. In 1947 he discovered a new type of comparatively recent stellar system, which he named stellar association. The most important result of his study is the conclusion that the process of star formation in the Galaxy continuing our solar system still continues and, specifically, that most stars have their origin in changing systems of groups of stars.

Later, Ambartsumian studied the phenomena in the atmosphere of stars that are changing in physical characteristics, such as luminosity, mass, or density. He saw these changes as being connected with the direct release of interstellar energy in the outer layers of the stars. He also investigated nonstationary processes in galaxies. These investigations are of great importance, both for the problem of the evolution of galaxies and for the study of still unknown properties of matter.

Theoretical Astrophysics

His textbook, *Theoretical Astrophysics* (1958), has gone through many editions and translations. It contains examples of his unique and fruitful approaches to stubborn

astronomical problems. In addition, he studied radio signals coming from outside our galactic system of stars. He was led to conclude that these radio signals represent not colliding systems of stars, according to a widely-accepted interpretation, but the subatomic process of fission within galaxies. Therefore, according to his view, "radio galaxies" may represent systems of stars, interacting in close proximity, that were formed from superdense formations of stellar material. In support of this view, he pointed out the presence of jets, condensations, and streamers that are bluish in colour; found around certain galaxies, these are characteristic of an early stage in stellar development.

Ambartsumian's invariably thought-provoking manner of presentation drew large audiences to his lectures at international symposia. He enlivened even his most abstrusely mathematical lectures with quotations from classic and contemporary poets.

Many governmental decorations and awards were presented to Ambartsumian. In 1947 he was elected president of the Armenian S.S.R. Academy of Sciences and member of the Parliament of Soviet Armenia and, in 1950, to the Supreme Soviet of the U.S.S.R. In 1953 he was elected to full membership in the Academy of Sciences of the U.S.S.R. In 1948–56, he was vice president and, in 1961–63, president, of the International Astronomical Union. In 1968 he became president of the International Council of the Scientific Union, and he participated in activities of many foreign academies and scientific societies. For his many contributions to Soviet science, his country awarded him two Stalin prizes, among many other orders.

Ambartsumian, a popular figure among Soviet Armenians, was honoured in 1968, when his 60th birthday was celebrated as a national event. Armenians of other countries also share the pride in his achievements. His two daughters and two sons all embarked on scientific careers in mathematics and physics.

BIBLIOGRAPHY. For further information, see A. MASSEVITCH, "Viktor Amazaspovich Ambartsumian: A Portrait," *ICSU Rev.*, 5:29–33 (1963).

(A.G.Ma.)

Ambrose, Saint

Ambrose of Milan was a foremost churchman, critical interpreter of the Bible, and preacher of the early Christian Church. As Bishop of Milan (374–397), his energy and authority ensured the dominance of Catholic Christianity among the governing class of the western half of the late Roman Empire. His relations with the emperor Theodosius I provided a model for medieval ideas on the relations of church and state.

Early career. Second son of the commander of the emperor's Gallic body guard, Ambrose was born about 339 at Augusta Treverorum (Trier), where his father had his official palace. His father probably died soon afterward, for Ambrose grew up in Rome, in a palace frequented by the clergy, with his widowed mother and his elder sister Marcellina, a nun. Duly promoted to the governorship of Aemilia-Liguria in *c.* 370, he lived at Milan and was unexpectedly acclaimed as their bishop by the people of the city in 374. A popular outsider, chosen as a compromise candidate to avoid a disputed election, Ambrose changed from an unbaptized layman to a bishop in eight days. Coming from a well-connected but obscure senatorial family, Ambrose could be ignored as a provincial governor; as bishop of Milan he was able to dominate the cultural and political life of his age.

Church-state relations

Ecclesiastical administrative accomplishments. An imperial court frequently sat in Milan. In confrontations with this court, Ambrose showed a directness that combined the republican ideal of the prerogatives of a Roman senator with a sinister vein of demagoguery. In 384 he secured the rejection of an appeal for tolerance by pagan members of the Roman senate, whose spokesman, Quintus Aurelius Symmachus, was his relative (Letters 17, 18). In 385–386 he refused to surrender a church for the use of Arian heretics. In 388 he rebuked the emperor Theodosius for having punished a bishop who had burnt a Jewish synagogue. In 390 he imposed public penance on Theodosius for having punished a riot in Thessalonica by a massacre of its citizens. These unprecedented interventions were palliated by Ambrose's loyalty and resourcefulness as a diplomat, notably in 383 and 386 by his official visits to the usurper Maximus at Trier. In his letters and in his funeral orations on the emperors Valentinian II and Theodosius—*De obitu Valentiniani consolatio* (392) and *De obitu Theodosii* (395)—Ambrose established the medieval concept of a Christian emperor as a dutiful son of the church "serving under orders from Christ," and so subject to the advice and strictures of his bishop.



St. Ambrose, detail of a fresco by Pinturicchio, 1480s. In Sta. Maria del Popolo, Rome.
Alinari

Literary and musical accomplishments. Ambrose's relations with the emperors formed only part of his commanding position among the lay governing class of Italy. He rapidly absorbed the most up-to-date Greek learning, Christian and pagan alike—notably the works of Philo, Origen, and St. Basil of Caesarea and of the pagan Neoplatonist, Plotinus. This learning he used in sermons expounding the Bible and, especially, in defending the "spiritual" meaning of the Old Testament by erudite philosophical allegory—notably in the *Hexàémeron* ("On the Six Days of Creation") and in sermons on the patriarchs (of which *De Isaac et anima* ["On Isaac and the Soul"] and *De bono mortis* ["On the Goodness of Death"] betray a deep acquaintance with Neoplatonic mystical language). Sermons, the dating of which unfortunately remains uncertain, were Ambrose's main literary output. They were acclaimed as masterpieces of Latin eloquence, and they remain a quarry for students of the transmission of Greek philosophy and theology in the West. By such sermons Ambrose gained his most notable convert, Augustine, afterward bishop of Hippo in North Africa and destined, like Ambrose, to be revered as a doctor (teacher) of the church. Augustine went to Milan as a skeptical professor of rhetoric in 384; when he left, in 388, he had been baptized by Ambrose and was indebted to Ambrose' Catholic Neoplatonism, which provided a philosophical base that eventually transformed Christian theology.

Conversion of Saint Augustine

Ambrose provided educated Latins with an impeccably classical version of Christianity. His work on the moral obligations of the clergy, *De officiis ministrorum* (386), is skillfully modelled on Cicero's *De officiis*. He sought to replace the heroes of Rome with Old Testament saints as models of behaviour for a Christianized aristocracy. By letters, visitations, and nominations he strengthened this aristocratic Christianity in the northern Italian towns that he had once ruled as a Roman governor.

In Milan, Ambrose "bewitched" the populace by introducing new Eastern melodies and by composing beautiful hymns, notably "Aeterne rerum Conditor" ("Framer of the earth and sky") and "Deus Creator omnium" ("Maker of all things, God most high"). He spared no pains in instructing candidates for Baptism. He denounced social abuses (notably in the sermons *De Nabuthe* ["On Naboth"]) and frequently secured pardon for condemned men. He advocated the most austere asceticism: noble families were reluctant to let their daughters attend the sermons in which he urged upon them the crowning virtue of virginity.

Evaluations and interpretations of Ambrose. Ambrose's reputation after his death, in 397, was unchallenged. For Augustine, he was the model bishop: a biography was written in 412 by Paulinus, deacon of Milan, at Augustine's instigation. To Augustine's opponent, Pelagius, Ambrose was "the flower of Latin eloquence." Of his sermons, the *Expositio evangelii secundum Lucam* ("Exposition of the Gospel According to Luke," 390) was widely circulated.

Yet, Ambrose is a Janus-like figure. He imposed his will on emperors. But he never considered himself as a precursor of a polity in which the church dominated the state: for he acted from a traditional fear that Christianity might yet be eclipsed by a pagan nobility and Catholicism uprooted in Milan by Arian courtiers. His attitude to the learning he used was similarly old-fashioned. Pagans and heretics, he said, "dyed their impieties in the vats of philosophy"; yet his sermons betray the pagan mysticism of Plotinus in its most unmuted tints. In a near-contemporary mosaic in the chapel of S. Satiro in the church of S. Ambrogio, Milan, Ambrose appears as he wished to be seen: a simple Christian bishop clasping the book of Gospels. Yet the manner in which he set about his duties as a bishop ensured that, to use his own image, the Catholic Church would rise "like a growing moon" above the ruins of the Roman Empire.

BIBLIOGRAPHY

Collected editions: The edition of the complete works given in J.P. MIGNE (ed.), *Patrologia Latina*, vol. 14–17 (1845), is gradually being superseded by that of the "Corpus Christianorum Series," *Sancti Ambrosii Mediolanensis opera* (1957–). English translations of Ambrose's letters and various works are given in vol. 26, 42, and 44 (1954–63), in the series "The Fathers of the Church."

Life: The contemporary biography, written by Ambrose's former secretary Paulinus of Milan, was edited by MICHELE PELLEGRINO as PAOLINO DE MILANO, *Vita di S. Ambrogio* (1961). The most comprehensive biography is F. HOMES-DUDDEN, *The Life and Times of St. Ambrose*, 2 vol. (1935). A shorter account is ANGELO PAREDI, *S. Ambrogio e la sua età*, 2nd ed. (1960, Eng. trans., *Saint Ambrose: His Life and Times*, 1964). Ambrose's current position is dealt with in J.R. PALANQUE, *Saint Ambroise et l'Empire romain* (1933); his relation to Neoplatonism is considered in PIERRE COURCELLE, *Les Lettres grecques en Occident, de Macrobe à Cassiodore*, rev. ed. (1948; Eng. trans., *Late Latin Writers and Their Greek Sources*, 1969).

(P.R.L.B.)

American Indian Peoples, Arts of

To achieve an understanding of American Indian art requires at the outset a willingness to discard many long-standing preconceptions and judgments based upon an evaluation of Western art. Because the basic aesthetic tenets and artistic goals are different, any art critic employing the usual Western ground rules in an attempt to comprehend Indian art is bound to be unsuccessful. At best, it is difficult for non-Indians to evaluate Indian art. Although it is not necessarily true that one must be an Indian to appreciate their art, it is true that the finer subtleties and depths of significance that are so obvious to the Indian are often lost to the non-Indian observer. Finally, it is deeply to be regretted that most if not all of the major commentaries on Indian cultures have been recorded by those who were natural enemies; the white man, however sympathetic he might be, simply cannot ignore the fact that he is, after all, evaluating the product of a culture that he has successfully suppressed. This inevitably introduces emotional and psychological elements that result in tremendous distortions. Prior to the 20th century, Indian cultures were considered essentially savage, hardly meriting Western attention. It is only in the past half century that Indian art has enjoyed serious, balanced consideration, and even this has been in meagre terms.

Today one of the obstacles to a realistic evaluation and appreciation of Indian art is the belief held by many that the mere fact that the primitive artist occupies a place so close to nature somehow endows him with a privileged status insofar as "natural art expression" is concerned. To regard a product as artistic simply because it was Indian-made defies reason. While it is a fair statement that no Indian tribe existed that did not develop some degree of art—and many were responsible for aesthetic masterworks—it is not true that all expressions were equally impressive. The quality and beauty of these arts varied widely; for, just as in the Western world, there were good and bad artists, and the fact that a native craftworker turned his hand to a given task in no way guaranteed success. Another obstacle to a full appreciation of the arts of the Indian peoples is the belief held by many sympathetic viewers of aboriginal art: that there are eternal truths which can be found in all cultures, and that to the truly sensitive eye these aesthetic verities exist wholly independent of the particular cultural milieu, the purpose of the art expression, or the relative level of the cultural development. The position taken in this article is that since the literature, music, dance, and visual arts of the Indian peoples did not evolve in isolation from the sociological, religious, and political milieus of those peoples, an understanding of the latter cannot be divorced from an appreciation of the former. The article will deal with the arts of the North, Central, and South American Indian peoples and the Eskimo, in relation to the culture and subcultures of each of these peoples and in relation to the Western world. (F.J.D.)

Importance of the sociological, political, and religious background of Indian art

This article is divided into the following sections:

I. Literature
  General characteristics
  Oral literatures
    North American cultures
    Central American cultures
    South American and Caribbean cultures
  Written literatures
  Study and evaluation
II. Music
  Dominant native style
    Melody and rhythm
    The significance of song texts
  Musical instruments
    Drums: membranophones
    Rattles: idiophones
    Wind instruments: aerophones
    Stringed instruments: chordophones
  Regional customs: North America
    The Far North
    The Eastern Woodlands area
    The Great Plains
    The West
    The Southwest (Arizona and New Mexico)
  Regional customs: Mexico and Middle America
    Northern Mexico
    Middle America
  Regional customs: South America
  Studies and publications
    Musical transcriptions and analysis
    Other fields of Indian musicology
III. Dance
  General characteristics of American Indian dance
    Extent of dance forms
    Patterns of participation
    Socially determined roles in dance
    Religious expression in dance
    Patterns and body movement
    Foreign influences
  Regional dance styles
    The Eskimos
    Eastern Woodlands
    The Great Plains
    The Northwest coast
    The northern desert and California

The Southwest
Mexico and Middle America
South America
Study and evaluation
IV. Visual arts
Nature and elements
The role of the artist
The function of art
Materials
Regional styles of American Indian visual arts
North America
Mexico and Central America
West Indies
South America
Arts of the American Indian peoples in the contempory world

I. Literature

GENERAL CHARACTERISTICS

Folktales have been a part of the social and cultural life of American Indians regardless of whether they were sedentary agriculturists or nomadic hunters. Gathered around a fire at night, the hardworking Indians could be transported to another world through the talent of a good storyteller. The effect was not only derived from the novelty of the tale itself but also from the imaginative skill of the narrator, who often added gestures and songs and occasionally adapted a particular tale to suit a certain culture. One adaptation he often used was the repetition of incidents. He would repeat an incident a specific number of times (depending upon the ceremonial number in the culture, three, four, five, or seven), with the hero killing that many monsters or that number of brothers who had gone out on the same adventure. This type of repetition was very effective in oral communication, for it firmly inculcated the incident in the Indians' minds much in the same manner that repetition is used today in radio and television commercials. There was an aesthetic value to the rhythm gained from repetition and also a dramatic effect, for the listener knew that when the right number of incidents had been told, some supernatural character would help the hero, sometimes by singing to him. For this reason oral literature is often difficult and boring to read. Oral literature also loses effect in transcription because the reader, unlike the listener, is often unacquainted with the world view, ethics, sociocultural setting, and personality traits of the people in whose culture the story was told and set.

Because the effect of the story depended so much on the narrator, there were many versions of every good tale. Every time a story was told, it varied only within the limits of the tradition established for that plot and according to the cultural background of the narrator and the listeners. While studies have been made of different versions of a tale within a tribe, there is still much to be discovered, for instance, in the telling of the same tale by the same narrator under different circumstances. These gaps in the study of folktales do not indicate lack of interest but rather the difficulty in setting up suitable situations for recordings.

Use of the terms myth and folktale

The terms myth and folktale in American Indian oral literature are used interchangeably because in the Indian mind the difference between the two is a matter of time rather than content. If the incidents related happened at a time when the world had not yet assumed its present form, the story may be regarded as a myth; but even if the same characters appear in the "modern" present, it is considered a folktale. Whereas European fairy tales traditionally begin with the vague allusion "once upon a time," the American Indian myth often starts with "before the people came" or "when Coyote was a man." To the Eskimo, it is insignificant whether an incident occurred yesterday or 50 years ago—it is past. Thus, while the mythology of preliterate people has a historical development of its own, it cannot be admitted to the present-day classification of historical because it lacks documentation.

American Indian mythology can be divided into three major cultural regions: North American cultures (from the Eskimos to Indians of the Mexican border), Central and South American higher civilizations, and the Caribbean and South American simpler cultures. Though each region exhibits a wide range of development, there are recurrent themes along the cultures, and the importance of mythology itself within each culture varies. In North America, for example, stories are not involved with a panoply of divine beings, as are those of Middle America and South America, where the tales bear some resemblance to the complicated mythologies of the surrounding higher cultures and are quite confusing with many hybrid monsters and giants. In North America many mythologies (like "al Churinga" of the Australian aborigines) deal with a period in the distant past in which the world was different and people could not be distinguished from animals. These mythologies are related to the concept that all animals have souls or spirits that give them supernatural power. Because man has subsequently been differentiated from the animals, they come to him in visions, and in the stories they help the hero out of trouble. When there are many tales involving a single character, such as Raven, Coyote, or Manabozho, the transcriptions are linked together today and called cycles. Whether this cyclic presentation of tales is or is not indigenous to the Indian mind remains questionable, but it is used by folklorists who have been accustomed to legendary cycles of medieval European literature, such as the tales of *Beowulf*, *La Chanson de Roland* (*The Song of Roland*), or even *The Canterbury Tales*.

The body of American Indian folklore does not include riddles as found in African folklore, for example, nor does it include proverbs, though there are tales with morals attached.

Importance of mythology within a culture

The importance of mythology within a culture is reflected in the status of storytellers, the times assigned to this activity, and the relevance of mythology to ceremonialism. In many instances the relationship between mythology and ceremonialism is very close. The Navajo ceremonials, like the chants, are based entirely on the characters and incidents in the mythology. The dancers make masks under strict ceremonial control, and, when they wear them to represent the gods, they absorb spiritual strength. The Aztec ceremonials and sacrifices placate the gods who are the heroes of the mythology. If, however, the mythology is made up primarily of animal tales and stories of personal and social relationships, the actors and characters involved in the stories are also an index to the beliefs and customs of the people.

ORAL LITERATURES

North American cultures. *Eskimo.* Eskimo culture can be divided into two major subgroups: one culture extending from Greenland to the Mackenzie River and another west from the river to the Pacific Ocean. The Eskimo culture embodies simple stories of hunting incidents in which the heroes are sometimes helped through supernatural power. Other stories include themes in which people ascend to the sky to become constellations: maltreated children become animals, and an orphan boy becomes successful. Still others surround the exploits and priestly magic of the shamans. In the region from Greenland to the Mackenzie River, Sedna is the highest spirit and controls the sea mammals; the Moon is a male deity who lives incestuously with his sister, the Sun. When she discovers he is her brother, she seizes a burning faggot and rushes away into the sky, the Moon pursuing her.

There are many stories involving family life and others dealing with the feuds between Eskimos and the Indians south of them.

Stories of the western Eskimos

The western Eskimos along the Pacific and Arctic coasts have the Raven cycle, a series of tales centred around Raven, a protagonist whose role ranges from culture hero to the lowest form of trickster. Many of the same plots and themes also occur in tales of the Northwest Coast culture. Around St. Michael and other coastal villages, a story about a flood that took place in the first days of the earth is told. Many stories are especially intended for children and stress proper behaviour. They are often told by young girls to younger ones and are illustrated by incising figures in the snow or on the ground with an ivory snow knife. On the lower Yukon, a migra-

tion legend is told about a long journey from east to west. The usual incident breaking up this party of travellers is a quarrel, after which they divide into two groups, occupying separate villages, and for years make constant war on each other. Tales of hunting begin as personal adventures but become stylized with supernatural characters and events.

Northwest Coast. There is greater similarity in the mythology of the various tribes along the Northwest Coast than in other regions of North America. Collectors of folktales have gathered a long series of stories told in the region from the mouth of the Columbia River through southeastern Alaska and have gathered them into a Raven cycle. The protagonists of these stories—from south to north, Coyote, Mink, and Raven—again vary from culture hero to the lowest form of trickster. In each subarea the stories elucidate the origin of a village, a clan, or a family and are regarded as the property of these clans or families. Thus, they can only be used by others through permission or, sometimes, purchase. Examples of this type of myth are Bungling Host, Dog Husband, and Star Husband. In Bungling Host, Trickster, after seeing his host produce food in various ways (*e.g.*, letting oil drip from his hands), fails to imitate the magic methods to procure food and barely escapes death. In Dog Husband, a girl has a secret lover who is a dog by day and a man by night. When she gives birth to pups, she is deserted by her tribe. She then destroys her children's dog skins, and they turn human, becoming successful hunters. In some versions parents lose all their sons to a monster, and when a new baby is born, it grows rapidly, kills the monster, and restores the brothers. Star Husband, another widely known myth, relates the story of two girls sleeping outdoors who wish the stars would marry them. They ascend to the sky, marry the stars, and experience a series of remarkable adventures.

Mythology of the Kwakiutl and Salish-speaking tribes

Among the Kwakiutl of Vancouver Island, the mythology is represented in an elaborate series of dances that illustrate characters and incidents with masks, puppets, and other mechanical devices. The principal events during the winter ceremonial season, these ceremonies include initiation into the secret societies, the highest of which is the Cannibal Society. Less elaborate forms of this winter ceremonial are found among the southern tribes who base their activities on the quest for the guardian spirit and on the return of the spirits to those who have seen them in visions. In order to exorcise these spirits, their songs must be sung and their dances performed. The Salish-speaking tribes of southern British Columbia and Washington have less complicated costumes for this ceremonial, but their dancing is very interesting and vigorous. The attitude of the Indians of the Northwest Coast toward animals is expressed in rituals such as the first salmon ceremony and the ceremonial attitude toward the bear. When the first salmon of the spring run is caught, it is ceremonially cleaned and placed on a clean mat or a bed of fern leaves. It is welcomed with an address of thanks and promised good treatment. The entrails are wrapped in a mat and thrown into the river so that they can return to the land in the west where the salmon can tell how well he was treated. The salmon is carried to the house by a selected group—children, women only, or the family of the successful fisherman—and is roasted and eaten, also by the selected group. The bear is never killed wantonly. When seen, it is addressed in terms of kinship, expressing an attitude that is known sporadically from the Ainu of northern Japan to many places in northwestern America.

California. The many small tribes of California exhibit more unity in their mythology than is present in many other features of their culture. In the north central area, the Kuksu cults enact the myths of the creator and culture hero with Coyote and Thunder as the chief characters. In southern California in ceremonies of the Chungichnich cults, contact with the highest god is achieved by smoking datura or Jimsonweed, which produces hallucinations of animals. The boys initiated into the cults regard the animals as their guardian spirits. This concept relates the cult activity with the most fundamental feature of American Indian religion, the concept of the individually attained guardian spirit.

In general, the Californian tribes have a very confused mythology, which was disintegrating even before the white man settled in their territory. Though some animal stories and a few themes about vague characters existed, little of it has been recorded and carefully studied.

Southwest. The Indians of New Mexico and Arizona, along with a few small tribes related to them in southern California, have cultural traditions with some features in common. In the folklore of the Southwest, the emergence and migration myths show the Indians emerging from an unpleasant underworld up to earth at the time when it is not yet completely formed. They start a long trek southward, some looking for a sacred spot, others looking specifically for the centre of the earth. In some instances they are lead by a pair of culture heroes, the Twins, also called the Little War Gods, who help stabilize the surface of the earth and teach the Indians many features of their culture, including ceremonials. When the people were weary during the migration, the gods, the kachinas, came and danced for them until someone made fun of their peculiar faces and insulted them. They allowed the people to copy their masks and costumes, and then they returned to their home under the earth. Since that time the men from the kivas, the ceremonial chambers to which all the men belonged, have made these costumes and masks and have performed the dances necessary to stimulate and protect the harvest, bring rain, and promote the general welfare.

The Twin Gods of the Pueblo villages are a combination of the helpful god and the trickster. They sometimes behave like unruly children and tease their grandmother to death. Coyote, in the Pueblo literature, is not the trickster he often is in the folklore of cultures farther north. Though he does share some of the trickster plots and is always sly, he is often caught in his own wiles. A group of very crude and vulgar tales about him exist, which have been transcribed from Spanish days, surrounding two characters who travel together named Djos and Ley, a corruption of "God" and "King." Certain European tales, such as the Cinderella stories, have been added to the collections of Pueblo folklore.

Ceremonials of the Navajo and Apache

The Athabascan-speaking tribes of the Southwest are the Navajo and Apache. Nowhere in America are mythology and ceremonial more closely associated than among the Navajo, where the myths are poetically expressed through great "chants." The principal characters are the gods of the wind, the rain, the dawn, the sun, the semiprecious stones, the sacred plants, corn, tobacco, squash, and the bean. The ceremonials are intended to cure sickness, both mental and physical, and to protect people on dangerous missions, rather than to inspire any sense of worship. All the arts are combined in the ceremonies: the story itself, the poetic expression of it, the painting of the masks, the beautiful combinations of feathers in the headdresses, the sand paintings illustrating the story while the chants are sung, and, finally, the dancing of the characters who wear the regalia. This is one of the most inspiring ceremonials devised by the American Indian.

The other Athabascan-speaking people, the Apache, are divided into several groups, of which the Lipan are particularly interesting. The southernmost of North American tribes, they live partly across the Mexican border. They have an emergence myth and share with all southern Athabascans the culture hero known as "Killer-of-enemies" and his younger brother "Child-of-the-Water" (comparable to the Twin War Gods of the Pueblos). One of the monsters in the tales is Big Owl, a destructive cannibal in the form of a large owl. The story of the man seeking spiritual power from the gods who goes down the Colorado River in a hollow log to reach the holy places where the spirits live is almost identical to its Navajo version. There is a Lipan Coyote cycle, but there are no Spanish-derived tales.

The White Mountain Apache tell stories only between dusk and dawn and during cold weather. They have two

major cycles, the creation myths, in which something is created out of nothing, and Coyote myths surrounding the trickster *par excellence* of that name. Two minor cycles centre around Big Owl and Gain, a supernatural being who lives in mountains and caves and underground worlds. The White Mountain Apache learned some European stories from captives in Spanish and Mexican towns in the late 18th and early 19th centuries. The Jicarilla Apache divide their didactic stories into sacred and secular, telling them at night during the winter with a break at midnight for a festive meal. The Jicarilla are seriously absorbed in their mythology, into which they inject objects of modern life such as the telegraph and the automobile. They also have creation stories and the legend of the man who went down the Colorado River in a hollow log. Among the trickster stories, the Coyote cycle is well developed. The Colorado River tribes, which are closely affiliated with the people of southern California, comprise the last division of the Southwest cultures. These tribes were Christianized by missionaries so early that little of their mythology remained to be recorded. It is known that Coyote played an important part in their sacred stories, though he was also portrayed as a deceitful trickster. Like the Pueblo tribes, the Luiseño also migrated, looking for the centre of the world, where their god, Wiyot, had died. His death was the first among the people, and they lost their immortality. Wiyot later returned as the new moon. There are many stories about the stars, which were regarded as the souls of the dead. The Chunguchnich cult was also known here but may have come within the mission period.

Eastern Woodlands. The northeastern Algonkins were the first groups of American Indians north of Mexico to have protracted contact with Europeans, so their own ways of living were disrupted at a very early date. Some of their culture traits can still be found among the Central Algonkin to the west, and some of the most elementary stories are known to all groups in this region. This mythology centres around a culture hero known as GlusKap to the Micmac and as GlusKabe among the Algonkin, whose consistently altruistic character and human appearance distinguish him from many other culture heroes. He carries out the usual exploits, one of the most popular being the episode in which he kills a Monster Frog who has been impounding the water. Though he revels in the trickster adventures of all American Indian characters, he appears somewhat exempt from the crude buffoonery of other culture heroes.

To the west, the Central Algonkin developed the Midewiwin or the Grand Medicine Society—shared by the Eastern Sioux—whose activities revolved around the quest for a vision that would bring them in direct contact with supernatural beings who instructed them in curing ceremonies. The members of the society were not shamans, had no individual powers, and were only effective when they acted together. In its use of certain mnemonic devices containing a series of symbols used for instructing initiates, the society foreshadowed an approach to writing.

Folklore of the Iroquois

The Iroquois, who developed one of the great confederacies of American Indians, had a strong religious and mythological background for their folklore. In their creation story, which is the basis of their religious beliefs, they acknowledged a supreme being "beyond the conception of man." This "being" sent from heaven is a woman who in her descent fell on a big turtle imbedded in mud. She gave birth to a daughter conceived by the turtle who bore two sons. The good one was born first; the other was born through the mother's side and killed her. The sons grew up and helped their grandmother finish the formation of the earth. The Iroquois had curing societies similar to the Midewiwin; their members were not shamans, and they cured in a group rather than as individuals.

Many tribes in the Southeast exhibit cultural systems very similar to those of the northeastern tribes; but others, especially in the lower Mississippi Valley, had a more elaborate religion and mythology that divulged a definite relationship to the higher cultures of Meso-America.

Plains. The expansive area of North America between the Mississippi River and the Rocky Mountains, extending from the Gulf of Mexico to the sub-Arctic Indians of Canada, embodied many cultures whose various rites and ceremonies emerged from a common background. Many tribes were seminomadic and depended more on buffalo hunting than on agriculture for their living. The more sedentary groups, the Eastern Siouan, included the Mandan and Hidatsa. Marginal groups, which seem to have continued an older form of Plains culture before the advent of the horse, lined the borders of the Plains area.

The Sioux narrate the following creation story: the Old Man, Waziya, lived beneath the earth with his wife. Their daughter married the Wind and bore four sons, the winds North, East, South, and West. Together with the Sun and Moon, the winds were the "Controllers" of the universe, and a series of very involved stories tells of their powers. As the world was being formed by the invisible being, Iktoma, the trickster made trouble wherever he could. The usual plots are found in this collection of trickster stories. In order to reach the supernaturals or "Controllers," rituals and ceremonies had to be conducted. The most important ritual was the Sun Dance, because the Sun was one of the principal powers. A medicine man was responsible for the whole ritual, which had to be recited in letter-perfect fashion, because great harm could be inflicted on the whole group if the performance was not exact.

Creation myth of the Crow

In contrast to the Sioux, the Crow are a bit more lighthearted about their approach to the universe. Their culture speaks of a creation myth in which Old Woman's Grandchild, the son of an Indian woman and the Sun, destroys monsters. He then goes to the sky and becomes Morning Star. The genealogy of this character very closely resembles the Navajo myth of Changing Woman, the Sun's mistress who bore the children Monster-slayer and Child-born-of-Water. This concept of change into an astral body is quite widespread in the Plains. In a Cheyenne version of the Dog Husband story, the mother and her children go to the sky and become the Pleiades constellation. The Crow liked to express themselves poetically, and often they recited in song. The military societies have many songs that express their high aims and others that are songs of bravado. In many of their amusing stories, there are plays on words that are often difficult to translate.

The last of the Plains tribes, the Comanche, believe that the Great Spirit created some people, but that there were white people existing before them. A flood washed these white people away, and they turned into white birds and flew away. A secondary spirit was then sent to create the Comanche. but they were not perfect at first; therefore the spirit came a second time, giving them intelligence and showing them how to make everything. There are the usual trickster stories with Old Coyote as the central figure. Since the Comanche were a very aggressive tribe, there are many stories based on war exploits.

Central American cultures. The cultures of Central or Middle America comprise the highest developments of the New World and have been considered comparable to some of the oldest cultures of the Mediterranean. Included are the Aztecs of Mexico, the Maya of Central America, and the Inca of Peru.

The Maya, who will be mentioned again below, have a very complicated creation myth relating the several stages at which man did not satisfy his creator. After each stage he was destroyed, and another attempt to create a more perfect being was made. Since Maya culture in its highest form can only be reconstructed from the art and the hieroglyphics, little is known except the mythology that can be derived from the religious ceremonial art (see below *Written language* for details).

Mythological eras of the Toltec period Aztecs

The Aztecs of the Toltec period had four mythological eras: those of (1) the Water Sun, which was destroyed by flood; (2) the Sun of the Earth, which was destroyed by earthquake; (3) the Wind Sun, which was destroyed by a giant and only Quetzalcóatl, the feathered serpent, left, prophesying the destruction of the Earth by wind and the

evolution of humans into monkeys; and (4) the present Sun of Fire, which will end in a general conflagration. Quetzalcóatl, the survivor of the age of the Wind Sun, brought civilization to the people. This mythology, which was the basis of the ceremonial life, was maintained by the ceremonial priest, but there were also simple folktales that resembled those of North America.

Much more information about the Aztecs exists than about the Maya, because their civilization was still functioning at the time of the Spanish conquest; whereas Maya culture had completely changed, and the old traditions were almost unknown. What is available is the result of painstaking scholarly work in the analysis of the hieroglyphics.

The Inca civilization of Peru has been added to the higher cultures of Meso-America because it resembles them more closely than it does its neighbours, the simpler tribes of South America. As far back as mythological history can be traced, the Incas have worshipped Viracocha, the creator. He was the omnipotent being who took part in every mythological incident. He created man from painted stone dolls, a specific way of saying that man evolved from the living rock. He also had the primitive capacity for infinite self-multiplication, and some of his offspring also became local gods. Another cycle of creation in which Viracocha functions bears some resemblance to the periods listed for the Aztec. The "Origin People" came out from their caves, and the creator organized the process of living for them. They became the ruling class. Then the "Wilderness People" came from other caves and became the common people, who increased rapidly. Many diverse languages and cultures developed. The next cycle produced the "Wartime People," who placed a premium on being ruthless, strong, and cunning. They took whatever they wanted and forced people to move to more unfavourable places; in this way, they spread the population. In all of these myths the flood is present, which requires the re-creation of man after each incident.

Since Peruvian mythology covers a large and difficult terrain, local cults developed in many places utilizing the same characters in different incidents. When all the incidents are assembled they seem very confusing and contradictory.

South American and Caribbean cultures. There are so many cultures among the Indians of South America, and so little is known about many of them, that a selection was made of four tribes, some of whose mythology has been published. The Aymara in northern Chile share the culture hero creator Viracocha with the Peruvians. According to the Aymara, he rose from Lake Titicaca, created the earth, the sky, and man, and then resubmerged. Man was disobedient to him, and he therefore led man to Tiahuanaco and turned him into stone. Viracocha created the sun, moon, and stars because man, whom he created again, lived in darkness. Another mythological character of the Aymara culture is Thunnupa, a bearded white man from the north who opposed polygamy and chicha, the national drink. Animal tales are also very common in this culture, some having Aesop-like plots. Fox is the comical character in tales, and, because of this European aspect, it is often postulated that the tales might have been brought in with the Spanish conquest.

The Mapuche culture, also in Chile, relates tales characterized by fairly long narratives about such supernatural characters as Shooting Star, who may also be a cannibal, hybrid monster, winged serpent, ghost, or apparition. Again, the wily Fox is the principal character in the animal stories, though he is often outwitted. Folktales are told at night and are accompanied by mimicry and gestures.

The many folktales of the Cágaba, who inhabit the Sierra Nevada de Santa Marta in Colombia, are religious in nature, having supernatural characters who arrange the world for man and who try to control the demons that plague him.

Finally, the Chibcha, who live in Colombia north of the Orinoco River, have a body of mythology that reflects the lack of organization found also in their culture. The stories are cosmological and ritual but lack all perspective of time. Many are concerned with Bochica, a culture hero who is one of the major gods of the Chibcha pantheon.

Mythology of the Caribbean Indians

Because many islands in the Caribbean were populated by people who came from the northern parts of South America, the scanty information available concerning their folktales and mythology is included in this section. The Arawak were the first Indians encountered in the New World by Columbus at Hispaniola. A typically Caribbean type of Indian who migrated from the tropical forests of South America, they combine the concept of a guardian spirit with fetish worship and fabricate idols representing plants, animals, and human spirits. The Sun and the Moon, who are connected with a myth about human emergence from a cave, together with various astral beings and a culture hero, are typical characters of Arawak mythology.

The Arawak Indians were soon decimated by Spanish invaders in the Caribbean, and by 1535 only about 500 were left at Hispaniola. The Spanish therefore brought Negroes and other Indians to work on plantations. This situation also occurred on Puerto Rico and Jamaica, where the Indian cultures were totally changed because of the Negro influx and subsequent mingling of the races. One outstanding collection of tales from this region consists of stories about Spider, a culture hero of West African folklore.

WRITTEN LITERATURES

One of the great scholars of Maya culture, J. Eric Thompson, has said:

> In the New World only Maya culture extends to us the privilege of sharing its thoughts and its struggles, its triumphs and its failures, for in the glyphs the dead past has left a chart to guide the living present along the corridors of time.

Maya hieroglyphs

In the 16th century, the area of hieroglyphic writing did not coincide with that of Maya speech. It appears that the hieroglyph originated in such lowland languages as Olmec and Zapotec. Though people may have used hieroglyphs at an early date, they seem to have never passed the rudimentary stages. Maya hieroglyphs on stone and wood are confined to the Classical period (AD 300–900). Inscriptions on monuments record the passage of time, often the close or decline of certain periods, and invoke images of the gods, the "rulers" of each day, to which they could bring fortune or disaster. Written codices comprised books of divination and a mixture of prophecy and history that were similar to the historical codices of Mexico. Three codices remain extant: Dresden Codex, Paris Codex, and Madrid Codex Tro-Cortesianus (in two parts), each named for the cities in whose museums the codices are currently exhibited. Maya hieroglyphic writing manifested a calendric system and also certain religious concepts of the Maya culture that reveal a mythology of surprising richness, demonstrating the independent growth of New World civilization.

The Maya culture is divided by scholars into four periods: first, the Formative or middle culture horizon in the second half of the millennium before Christ. During the latter part, the earliest known carving has been found, dated in terms of the Maya calendar at AD 320. Second, the Classical period, from about AD 300–900, saw the development of the great "cities," architecture, sculpture, and ceramics. At this time a series of sophisticated deities appeared that were no longer directly related to the soil or the elements. The culture became divided into two levels, a theocratic government and priesthood and a lay culture that remained simple and agricultural, with home industries, a simple family organization, and a religion built around the personification of powers of nature, which was served by a nonprofessional priest. Toward the end of this period, influences from Mexico made themselves felt. The third period, or the Early Post-Classic period, from 900 to 1200, followed a transition period when metal appeared. The first gold working area was Panama and Costa Rica. In this period, the Mexicans or the Chontal Maya conquered and settled in several

large cities in Yucatán, including Chichén Itzá. Itza, the conqueror of Chichén Itzá, introduced Mexican architecture and religion, including the cult of Quetzalcóatl, the feathered serpent god, as well as militaristic organizations such as the fighting orders of the Eagle and the Jaguar. Influences from Tula also modified Maya culture. According to later records, some mural texts and codices were made during this period, but they no longer exist.

Concurrently with the Classical period of the Maya, the Mexicans were also developing a written language, which was not so highly sophisticated as that of the Maya and could more correctly be called picture writing. The pictographic writing of the Aztecs was too simple to record literature, offering no way of making general statements or expressing abstract ideas. Though there was no alphabet in this writing, a picture of an object or an animal could be combined with another and given a new meaning. This writing was taught by the priests who were entrusted with the education of the young boys.

Development of pictographic writing

Pictographic writing developed in several areas, including the Mixtec-Puebla region and Texcoco. Meanwhile, the Aztecs were becoming more powerful along the outer borders of a highly civilized region, and about 1200 they moved closer to the centre of activity. As the government became more centralized, reports had to be submitted, and pictographic writing provided a satisfactory medium for this task. Even after the Spanish conquest, these reports were still presented in the same manner and form. Even when the writing was scrived by the Spaniards, Indians continued to do the illustrations.

After the conquest, historical accounts were written that reiterated the past history of the principal Aztec regions. Much of what is known today about the early history of the Aztecs is derived from these works. A method of recording Nahuatl, the language of a large portion of Mexico, was combined with Spanish to supplement the graphic records. It is believed that some of the graphic records represent oral traditions possibly learned in chants that were recited on ceremonial occasions.

Historically, the only recognized attempt at writing an Indian language (other than by linguists) was made by Sequoyah, a Cherokee born at Taskigi, Tennessee, about 1760. The son of the mixed-blood daughter of Chief Echota and a German father, he became a trapper and hunter until crippled in an accident. In 1809 he became interested in writing and printing, and in 1821 he gave a syllabary of his language to the chief men of his nation, which they accepted. Two years later he visited the western Cherokee in Arkansas, where he took up residence. In 1824 parts of the Bible were written in his script and others followed in 1828. The *Cherokee Phoenix*, a weekly paper, was published, to which Sequoyah, contributed, and he exerted much influence in the reorganization of Indian Territory by the government. He set out to find the origin of his language but died en route to Mexico.

STUDY AND EVALUATION

Many large collections of Indian folktales exist that are historically important, though they lack information necessary for a modern study of the works. To make such a study, the folklorist must have a biography of the raconteur and the circumstances and exact date of the collection. If the study is to encompass literary style as well as theme, the folklorist must know whether the tale was originally told in a European language or, if an interpreter was necessary, his relationship to the raconteur and his experience as an interpreter.

Anthropological and linguistic studies

In the 1920s and 1930s anthropologists experimented with a type of ethnographic recording in certain tribes, which served two purposes: it explained cultural activities and attitudes of a culture and supplied the anthropologists or folklorists with a new vocabulary not found in the transcribed folktales. The folktale has served as a source of study for linguistics scholars, and tales recorded in an accepted phonetic code have always been a great asset to their study also. In order to get more text of this kind, the famous anthropologist Franz Boas taught a Kwakiutl to write phonetic text, which was then translated. Anthropologists also realized that the folktale reflects the culture in which it is told and sometimes keeps up with culture changes or, conversely, retains historic patterns of the culture. Again Boas pioneered in a study of Tsimshian mythology, recording the pattern and customs of Tsimshian life as revealed in the myths.

Students of mythology in the 19th century were interested in distributional studies or in following a plot or a group of motifs around the world to discover how myths spread and how plots disintegrate and become motifs in other plots. They concluded that the life of a folktale among preliterate people depended upon how often it was heard and remembered, and that each place it was found it became part of its history.

Study of the folktale has engaged psychologists in the last half century, and studies are now made of personality traits of a culture as expressed in their tales, and scholars now search for symbols that articulate human experience in a culture. A raconteur's analysis of the ethical code expressed in a story often reveals the ethics of a particular society. Such studies have been made by Margaret Lantis in "Nunivak Eskimo Personality as Revealed in the Mythology" and by Melville Jacobs in his analyses of the Chinook myths *The People Are Coming Soon.*

In the second half of the 20th century, Indians have attempted to collect and publish stories of their own tribes. So far they have not been very successful, because they have made too great an effort to conform to the literary style and taste of the non-Indian reader, and hence the tales have lost much of their vitality. (E.Gu.)

II. Music

Native musicians of the New World have preserved many traditional songs and, at the same time, have also been receptive to European influence; thus, any examination of their music raises the problem of distinguishing native from hybrid music. This consideration of American Indian music will involve an outline of the dominant features of native styles, such as melodic and rhythmic characteristics and the meaning of song texts, and will include a description of the principal musical instruments used. There will follow an examination of the stylistic features of various Indian cultures within North America, Central America, and South America.

DOMINANT NATIVE STYLE

The definition of the native American Indian musical style is complicated by variety between tribes and within tribal repertoires. It is feasible to approach the problem by contemplating the songs in prominent native rituals of four different North American tribes—Iroquois midwinter rites, Dakota war (grass) dance ceremonies, Pueblo plaza dances, and Navajo medicine ceremonies.

An evident feature of the music is male dominance. Equally evident is the limitation of music to song, performed without harmonization and with percussion accompaniment. The singing voice is well supported from the diaphragm and has a pulsating quality caused by rhythmic expulsion of the breath. Yet within this basic vocal-tone quality the tribes show considerable variation. The voice of the Dakotas and other Great Plains Indians is shrill, loud, and ornate, compared with the full-toned vocalization of the other three tribes—especially the deep, rich Pueblo quality.

Melody and rhythm. Despite exceptions among other tribes, melodies are sparse, sometimes even monotone. Only rarely do songs expand beyond the notes of the five-tone (pentatonic) scale, for example, c–d–e–g–a–(c′). In Navajo songs there is a preference for the intervals of the triad, that is, for two successive rising or falling thirds (the third is the interval, or distance between notes, encompassed by three notes of a European major or minor scale plus the octave, as in bugle intervals, thus c–e–g–c). The tones of the melody most commonly spread over the range of an octave (as, from c to c′; eight notes of a Western scale), although special song types may have narrower or wider ranges. Melodies characteristically descend from a high note, frequently by a se-

quence of repetitions of the theme on ever lower pitch levels. Sometimes they end with the maintone as the lowest note; sometimes they dip below the maintone but conclude on this note.

The melodic themes vary in complexity. They may include syncopations (accents not coinciding with the normal musical beat) and melismata, flourishes with many notes to a syllable of text. The musical metres are almost always in irregular patterns, in contrast with the regular patterns of two, three, or four beats found in most European folk songs. The theme may be repeated or may be varied in pitch level and length, or it may lead to a second theme or a free development.

Drummer and singer coordination

Although some chants flow in free timing, most melodies adjust their accents to the duple beat (alternation of two beats, one strong, one weak) of the accompanying drums or rattles. The melody may lag behind the accompaniment or may syncopate the instrumental beat; but it holds to the same tempo, or basic speed, as the accompaniment, whether speeding up, slowing, or maintaining a steady tempo. This occurs whether the singer himself is the percussionist or whether he follows a separate drummer. If the percussionist uses a tremolo in certain song sections, the singer continues with the basic beat. At all times the musical pulse of the singer reinforces that of the instruments.

This vocal-instrumental pulsation guides the dancers, especially in the self-accompanied Iroquois stomp dances with their bouncy step. At all times the song structure guides the dancers. The structure as a whole may consist of a series of short songs—from six to 100 or more. Perhaps there is an increase in tempo and force, as in the Iroquois bear dance. Perhaps there is a succession of paired songs, slow and fast, as in the Pueblo harvest dance. Perhaps each song is split into two parts, one marked by vibrato, the other by duple beat, with corresponding different dance motions, as in the Iroquois eagle dance and the Pueblo buffalo dance. Since the tunes are set—be they ancient or newly composed—the dancers know exactly how to follow the musicians. In Iroquois rites the dancers often engage in some improvisation within the basic framework, although the singers do not improvise.

Within these four North American groups and among other tribes as well, these definitions would require modification. Some tribes, as in British Columbia, use microtonic scales (*i.e.*, having intervals smaller than a European half step) and complex percussion rhythms. To the south, flutes and other instruments are used. In Latin America hybridization of European and Indian styles presents the challenge of disentangling native elements from imported elements.

The significance of song texts. The language and the meaning of the song texts provide evidence of the extent of European influence. Many texts consist of repeated syllables formalized within each tribe, *e.g.*, Iroquois "yo yowine," Plains "wida weha wehiyo." The meaningful words in native rituals may address the supernaturals, the animal spirits, the deities of the four cardinal points of the compass, or the sacred mountains. Navajo texts are especially rich in symbolic imagery.

European influences on text and tune

European influence manifests itself in various ways. Songs of the Peyote cult, or Native American Church, may introduce the name of Jesus. Hymns of Great Lakes Christian Indians fit native words to European folk tunes or Wesleyan, harmonized hymn tunes. On a more secular plane, Oklahoma social round dancers sometimes include amourous English texts; yet the tunes are clearly indigenous. By contrast, Mexico's Aztecs have invented Aztec poetry to go with tunes in Spanish style to honour the Virgin Mary or the beloved. As the final stage of de-Indianization, the dance songs of the Zapotecs of Mexico combine love poetry in Spanish with languorous waltz-like tunes.

MUSICAL INSTRUMENTS

Musical instruments may be clearly native, clearly imported, or ambiguous in origin. Percussion instruments —drums and rattles—are the mainstay of New World instrumentation. In Arizona, Sonora (Mexico), and areas farther south, wind instruments, native or imported, add to the tonal richness. Stringed instruments have a tremendous vogue in Latin America. Despite their adaptation to native ceremonies and musical styles, they reveal their Iberian origin.

Drums: membranophones. As in many Asian tribal societies, the single-headed shaman's drum with its narrow, round frame is widely distributed. More unusual are the square drums found in British Columbia in Canada, Veracruz State in Mexico, and Guatemala. Other hide-covered drums are more localized. The Iroquois use a water drum made by stretching hide on a small, wooden keg partly filled with water. The same type of water drum also accompanies the incantations for the Midwestern Midewiwin medicine rite. A similar pottery drum serves the singers for the peyote rites of many tribes. The gentle tone of these drums contrasts with the power of the Great Plains powwow drum, a double-headed boomer suspended horizontally from a frame. In the drum society ceremonies, war dance ceremonies, and secular Oklahoma war dances, four men surround the drum and beat it relentlessly as they shout their aggressive tunes. The ceremonial drum of Pueblo dances, played by one drummer, is equally forceful but has a different, vertical shape; four or five drums can produce a volume balancing that of the 70-odd singers. These drums are always struck with a stick rather than by the finger.

Drums peculiar to Middle America

As exceptions, the Aztec upright, wooden, carved drum, the *huehuetl*, and the Maya pottery *kayuum* are finger drums. Skillful manipulation of the head produces various tones. The *huehuetl* is paired with the horizontal, wooden *teponaztli*, which is a slit-drum, a hollowed-out, carved log struck with sticks on various parts of the resonant body. The *teponaztli* is not a membranophone but an idiophone, an instrument whose hard body vibrates to produce sound. Another curious idiophone is the ancient *ayotl*, a turtle shell that was scraped with bone sticks in pre-Hispanic rites of the Olmec and Zapotec peoples of Mexico.

Rattles: idiophones. The forms and materials of rattles, which are classified as idiophones, are derived from the natural resources at the disposal of various cultures as well as from the purpose of the instrument.

The jingler. The jingler, or rattle, attached to a dancer's body or clothing, is probably the oldest form of rattle. Prominent in nomadic and hunting tribes, it frequently consists of animal hoofs or of imitations. Alaskan Eskimo dancers shake mittens with attached puffin beaks and belts with animal teeth. Deer hoofs are fastened to the belts of both the *maso*, or deer dancer, and the *chapayekas* dancers of the Yaqui tribe of Arizona and Sonora, Mexico, and also to bands worn across the shoulders of the peyote (*hikuli*) dancers of the Huichol, a tribe of northern Mexico. From northern Mexico to South America deer hoofs are worn on the girdles and ankles of curing shamans, or medicine men. Deer, peccary, or tapir claws are worn on the legs or ankles among Brazilian tribes such as the Kamakán, Apinayé, and the Amazonian Huitoto. Antelope, elk, goat, ox, or buffalo hoofs may augment or substitute for the deer hoofs, as on Iroquois dancers' knee bands in the knee-rattle dance. Among eastern tribes in the United States, they are used by the Shawnee and Delaware and in the Penobscot and Wabanaki leading dance. In modern times, metal cones in hoof shapes have been fastened onto the dress of the adolescent girl in the Apache puberty rite and onto dresses of women dancers among the Havasupai of the Grand Canyon. Siberian shamans wear iron, cone-shaped jinglers on their drums and kilts and collars, the iron itself conferring power over demons. Aztec dancers fastened gold, copper, or shell bells on ankles or breast. In South America snail shells or nuts are commonly substituted, as among the Brazilian Bororo and Amazonian Jivaro. The Incas of ancient Peru used anklets of fruit shells or gold.

Shaken rattles of animal origin. Deer-hoof rattles in clusters are shaken by the Yuma Indians of Arizona and

by kachina dancers of the Rio Grande Pueblos; by the leader of the California Diegueño *keruk*, or mourning rites; and by Brazilian Kamakán dancers. Bunches of hoofs are fastened to short sticks in the rattles of numerous men's societies of the Great Plains, particularly the dog societies of the Crow, Hidatsa, and Kiowa and the corresponding Oglala and Iowa Mawatani society. Among the Flathead of Montana such a rattle was used in the hysterical cures of the Bluejay shaman's dance and in bison-calling rites. Rattles of hoof clusters spread east to the Sauk, Fox, and Iroquois Indians, west to the California Hupa dancing doctors, and southwest to the shaman-singer of the Apache puberty rite. In California and South America the deer-hoof rattle is commonly associated with girls' puberty ceremonies. Among the Maidu and Karok of California, the girl herself shakes the rattle. Usually, older women surround the girl and thump long, hoof-tipped poles on the ground, as among the California Klamath, Tolowa, Shasta, Achomawi, and Maidu mountain tribes, as well as among the Chaco tribes of South America.

Deer-hoof rattles for girls' puberty rites

Sand-filled butterfly cocoons were fastened on sticks manipulated by the leader of boys' initiation rites of the Yokut, Maidu, Pomo, and Miwok of California. In the rain ceremony of the Arizona Papago tribe, the singer-dancers wear cocoons, along with bells and shells, and the Yaqui native dancers all tie strings of cocoons around their ankles.

Metal rattles. In recent times the Indian has taken with delight to the sleigh bells and metal-pellet bells (jingle bells), as well as to metal rattles introduced by the white man, preferring them to natural materials that are harder to obtain. The dancers and runners of the Hopi and Tiwa Pueblo Indians at Taos, New Mexico, and, generally, Oklahoma festival dancers tie them around their ankles, below their knees, around their waist, and in a strip from belt to knee.

Turtle shells. Another animal source for both worn and shaken rattles is the turtle or tortoise. With attached deer hoofs, a turtle shell is worn behind the right knee of members of Hopi men's societies and Snake societies and commonly by Hopi kachina dancers. Small tortoise shells are worn at the knees of the leading woman in Oklahoma stomp dances. In the eastern United States, Delaware reciters of visions shook a tortoise shell, and (a borrowing from the Cherokee) the Iroquois society of women planters passes such a rattle from singer to singer. In the Great Lakes region Huron shamans are reported to have shaken a tortoise shell, filled with pebbles, near the ears of their patients. Cure motivates the most spectacular use of the turtle rattle, namely by the therapeutic rituals of the Iroquois, Delaware, and Cherokee-Shawnee. The shells are cleaned and filled with cherry pits, and the neck is reinforced by inserting a wooden handle. In the Iroquois False Face society ceremony, the huge turtles furnish instruments for "doorkeepers," the 10-inch ones for the singers, and the little ones for the False Face beggar boys.

Gourd rattles. Agriculture introduced the gourd rattle, which is always shaken by the dancer or singer, never worn. Its shapes and sizes are as varied as those of the squash and gourd, and it is ornamented with ingenuity. The Arizona Papago shaman, sometimes the Iroquois and Cherokee, leaves the stem on as a handle. Usually a wooden handle is inserted, perhaps decorated with beads or feathers or with carvings. Perhaps the largest ones are the pair used in the *maso*, or deer dance, of the Yaqui of Arizona and Sonora, Mexico. The smallest gourd rattles are the tufted rattles of the Iroquois eagle dancers. In between are the flat Hopi and the egg-shaped Cherokee rattles and the small round *sonaja* rattle common throughout Mexico, particularly in the *Matachini* ritual dance. These are an inheritance from the ancient Aztec *ayacachtli* rattle used, for instance, in the springtime ceremony. The designs of perforations that covered these calabash rattles have spread to northern Brazil and to California and have even been imitated in wood by the Tsimshian of British Columbia. Symbolic painted designs may cover the surface, particularly of the Hopi kachina rattles. Ritualistically, they should be filled with maize kernels, a few squash seeds, and, in the Southeast, grains of wheat. The Iroquois prefer chokecherry seeds. Nowadays unsymbolic materials are inserted: buckshot, pebbles, even macaroni.

Ornamentation of gourd rattles

Other varieties of rattles. Where gourds are not available, other materials are pressed into service. Tribes of the North Pacific Coast have developed a fine art of carving wooden rattles into animal and bird shapes. North American Plains Indians and South American Patagonian tribes make them of hide, in various shapes: spherical, doughnut-shaped, or flat disks attached to drumsticks. The Chippewa and Naskapi (Eastern Woodlands tribes) shaman's *cicigwan* resembles a small drum. The Iroquois and Cherokee-Shawnee have introduced cow-horn rattles, and the Chippewa construct cylinders out of birchbark. The Yaqui *Pascola* (clown dancer) shakes a wood-and-metal *sena'asom*, a form of sistrum, an ancient type of rattle consisting of small metal bars hung with copper rings and set loosely in a small, oval, wooden frame. Peruvian pottery and Argentine Araucanian basketry rattles add to the wealth of variety of materials used, but they by no means exhaust the list. For again civilization has introduced harsh substitutes in the tin can, spice box, and almost any old box that is handy: even for the sacrosanct Midewiwin ("medicine ceremony") of the Chippewa and Menominee of the Great Lakes region.

Wind instruments: aerophones. In North America, flutes are limited to courtship, except for a whistle in an Iroquois medicine rite and in a gens (kinship-group) ceremony of the Meskwaki. In Mexico and southward, wind instruments provide colour and symbolic significance. At least two dances feature a one-man drummer-flutist who holds a flute in his left hand and plays a drum with his right. These are the *Pascola*, or clown dance, of the Yaqui (performed with a cane flute and 15-inch [36-centimetre] drum) and the *volador*, or flying acrobat, dance of eastern Mexico. Such flute-drum combinations resemble the pipe and tabor combination found in Europe, raising the question of importation. Yet it is probable that this is a coincidence and that the one-man drummer-flutist of Middle America is native. In contrast, the penetrating, oboe-like *chirimía* of Oaxaca, Mexico, originated in Spain.

Flute-drum combination

Wind instruments have a long history in Arizona, Middle America, and South America. Archaeological repositories have proved that the Aztecs had many kinds of flutes of diverse materials, including double and triple flutes capable of playing harmonies. They also paired trumpets in harmony.

In Bolivia, Peru, and Ecuador the ancient panpipe is still in use. Its harmonized, plaintive, pentatonic melodies provide an enchanting accompaniment to ritual and secular dances. In South America, the long *botuto* horn of Venezuela resembles the alphorn of the Austrian Alps and the gigantic Tibetan trumpet. Among the Araucanians of Argentina the horn appears in two forms, both fashioned of bamboo—the doleful ten-foot *tutruka* and the more melodious, three-foot *llolkin*. Particularly in Central and South America, conch-shell trumpets are important ritual instruments.

Stringed instruments: chordophones. With one exception, stringed instruments used in Indian music are of Spanish origin. Despite their nativistic ideology, even Central Mexico's Indian performers of the *concheros* society manipulate their armadillo-shell lutes in guitar harmonies and with a brushing technique. In the *matachines* dance-dramas of the Pueblos and Yaqui, the accompaniment of fiddle and guitar or fiddle and harp uses tunes resembling those of England's Morris dances—harmonized, foursquare, in diatonic (seven-note) major scales. Picturesque stringed instruments include the fiddles of the Tarahumare of northern Mexico. Small ukuleles, called *jaranas*, are plucked by the *viejitos* (little old men) dancers of the western Mexican Tarascan tribe. The only native stringed instrument is the Tarahumare earthbow. Serving as a percussion instrument, it consists of a bow held vertically upon a gourd set over a small

hole in the ground. The bowstring is struck with a stick, the gourd and hole acting as a resonator.

The list of instruments gives some idea of their rich variety. Their playing techniques and their materials and forms show tribal and functional variety. So does the relation to melody and dance movement.

REGIONAL CUSTOMS: NORTH AMERICA

North American cultural areas include the Far North, the Eastern Woodlands, the Great Plains, the West, and the Southwest.

The Far North. *Eskimo music.* The Eskimos are rapidly losing their ancient songs for shamanistic cures and animal dances. They retain a shadow of the creative contests in the "drum dance" exhibits of individual songs with asymmetrical, complex rhythms and narrow-ranged, erratic melodies. All around Hudson Bay, the acculturated tundra dwellers delight in the hymnody of the Catholic and Anglican missions, singing "What a Friend We Have in Jesus" in the Eskimo language.

Decline of Eskimos' ancient songs

Far northern tribes. The far northern Indians of Canada—*e.g.*, the Catholic Dogrib near Lac la Martre, Northwest Territories, and the Anglican Cree near Rupert House, Quebec—have also become experts in plainchant and in native versions of "Nearer, My God, to Thee." Among the few native retentions are the Dogrib songs for the Hand Game and for the drum dance and the tea dance, social round dances. These songs are remarkable because they show similarities between these Indians of the Athabascan linguistic group and their linguistic relatives, the Southwestern Navajo and Apache. The Dogrib prefer triad melodies and vocal pulsations, carrying pulsation to extremes of speed and complexity. The few reports on the music of other northern tribes indicate a substitution of hymnody in the native language for native songs, even among the isolated Mistassini Indians.

The Eastern Woodlands area. In the Eastern Woodlands, all along the northern Atlantic Coast and inland to the Great Lakes (even to Lake Winnipeg in Canada), some older people deplore the young people's lack of interest in the Indian hymns, from the 17th-century chants of the Catholic Penobscot of Maine to the vigorous, 19th-century Chippewa revival songs, with piano, organ, saxophone, guitar, and male and female vocal harmonies.

The Algonkian musical renaissance. Algonkians old and young, however, manifest a resurgence of racial pride. They usually show their enthusiasm in public programs, in New Brunswick and Maine for St. Anne's Day (July 26); in Cape Cod, Rhode Island, and Long Island at the ancient harvest time, August. One Narragansett matron, Princess Red Wing, has been reconstructing the traditional, monthly festivals for public presentation. Women have played a prominent role in musical resurrections at various locations, on Manitoulin Island, Ontario, and in Michigan.

The Algonkian song-and-dance repertoires contain a few local vestiges of earlier tribal music, besides introductions by visiting Iroquois, Oklahoma, and Pueblo Indians, and some inventions. They always feature corn dances, war dances, calumet (peace-pipe) dances, and antiphonal (alternating two groups of singers) stomp dances.

Survivals of antiphonal singing

Collections of surviving music of the Abnaki (Wabanaki) and Penobscot tribes suggest a traditional prevalence of antiphonal singing throughout the Eastern Woodlands as far west as the Mississippi River. This singing style survives among the Cherokee of North Carolina; the Seminole of Florida; and most potently among the Creek and Yuchi tribes, now removed to Oklahoma; and among the Iroquois of New York state and Six Nations Reserve, Ontario.

Traditional Iroquois music. In their scattered reservations, the Iroquois preserve a rich tradition of native, ritually functional music. They provide for individual creativity in the variations between various renderings of ancient songs and in new compositions for women's dance and the social fish dance. But they insist on tradition, admit no European influence on their music, and have been reluctant to accept the pan-Indian songs from Oklahoma.

Their vast and varied repertoire defies categorization. Most of their songs, however, reveal four styles, each with a separate historical origin. Some shamanistic medicine songs and individual men's songs recall Eskimo chants in their wavering, limited melodies, although they use much simpler, repetitive rhythms. Another style, for war and animals, fits into the definition at the beginning of this article, with pentatonic, sequentially descending melodies. These songs have two repeats, contrary to the quadruple repeats common in songs of the Great Plains Indians. In some ceremonies, women pick up the tune and render the second repeat during their dancing. The somewhat similar third style, exemplified by the eagle dance and several other rites, provides the most terse examples of the dual, Iroquois musical structure. Paradoxically, these songs probably emanate from the Central Plains. The moiety system, dividing the tribe into two groups of matrilineal clans, dominates the cast of singers and dancers. Each song has two repeats, within an extended series. Each melody—that is, song half—contains two parts: one with drum vibrato and tremulous movement and the second with a rhythmic beat and dance progression. The song cycle starts with a free, introductory invocation, and it ends with a more complex finale.

The fourth type is the antiphonal, self-accompanied song; it prevails in agricultural rituals. One distinguishing feature is the "rise" in the middle of the song. That is, after several repeats of the antiphonal theme, it reappears on a higher tone, with change also in the dance movements; then it subsides to several repeats on the original level. Thus, there is a three-part, not a dual, pattern. The Iroquois also use a more ancient and primordial type of antiphonal singing. Thus, there are numerous responses sung at the end of certain songs. These responses, primarily monotone with a temporary rise of one step, have in performance an insistent power that unites the song-dance leaders and the chorus of people, who sing the response.

The Great Plains. The Indians of the Great Plains have spread their influence far and wide—to the Eastern Woodlands Iroquois and the Pueblos of the Southwest. Their influence on the Algonkians of Wisconsin and Minnesota has been so strong that it has drawn them into the Plains rather than the Woodlands area. Thus influences of the Dakota tribe have fashioned the present Algonkian drum society and war dance and have displaced the ancient Great Lakes animal songs with their simple, repetitive themes and imitative dance movements.

Reciprocal influence from Eastern Woodlands tribes

Conversely, the Plains tribes of North Dakota, Manitoba, and Saskatchewan have borrowed several Algonkian rituals. In powwows they feature war dances with duple drumbeat and round dances with a triple beat (beats in groups of three: strong-weak-weak), and with a shrill female chorus an octave above the men. Within the dominant song-dance style, the war-grass dance displays such special characteristics as high pitch, wide intervals in the melody, and wide melodic range. The singers vary the intensity. They attack notes with barks, slides, glottal stops, hiccoughs. On a second rendition of a chorus phrase, they shift from the duple beat to four or six strong beats; then they continue softly, tapping the edge of the drum, gradually increasing to the final strong, loud volume. In Oklahoman versions of this type of dance song, the dance responds to the musical changes, but in the Northern Plains there is no corresponding change of step or movement.

The more northwesterly Plains tribes, especially the Salishan Flathead and Algonkian Blackfoot, do not include Algonkian rites. They intensify some musical features of the Dakotas—the pulsation, strident vocal quality, and the wide range of musical scales in war dances. But they also retain some old medicine songs and sun dance songs with a less dramatic style. The Flathead have substituted such Plains traits characteristic of the Dakota for the wavering, more subdued musical qualities of the Salishan tribes.

The West. *The tribes of the Northwest.* The tribes of British Columbia share the dramatics of trancelike, Eskimo shaman dances, as well as their rhythmic complexity. But the British Columbia tribes have evolved a unique musical style, with elusive tonality, or musical orientation, around the maintone and clearly perceived notes of the scale. Thus, they use semitones, or half steps, and even microtones, with counter-rhythms of song and percussions, erratic musical metres, and complex song structures. Other special features are a drone (sustained note heard against a melody), which provides a rudimentary harmony; occasional use of simple flutes; and an extensive vocabulary of musical patterns and performance techniques, especially among the Nootka and Kwakiutl tribes. Their flamboyance differs from that of the Plains Indians and even more from the other western tribes in the deserts of California and Arizona and the Rocky Mountain Plateaus.

Complex melodies of British Columbia

Music of desert and plateau tribes. The tribes of the deserts and plateaus can be grouped together because of the simplicity of their aboriginal cultures and musical styles. Unifying factors are ritualistic preoccupation with visions, cures, and death (as in the elaborate Yuma mortuary rites) and long, mythological ballads. They also share a nonpulsating, relaxed vocal quality, an undulating, narrow-range melodic contour, and simple rhythmic patterns. Variations among tribes include instruments like the Ute scraping stick and the Pima upturned-basket drum.

A special Yuman feature is termed the "rise," a higher pitch level in the middle of a song that recalls a "rise" of the Woodland stomp-dance song but without its antiphonal singing style. These simple people, however, have become very susceptible to Euro-American culture and religion and are losing their traditions and accepting Mormon hymns and jazz dances.

The Southwest (Arizona and New Mexico). Geographically, Arizona tribes such as the Pima, Papago, and Havasupai can be included in the heterogeneous Southwest, but their subdued musical styles differ from the styles of the vigorous nomads—the Navajo and Apache of the Athabascan language group and the richly endowed Pueblo-dwellers, with similar cultures but diverse languages.

The Navajo and Apache. The Navajo and Apache live miles apart and have been subjected to diverse influences, but they retain common qualities in their consistent preference for melodies with "bugle" intervals and their pounding percussion beats. Their social round-dance songs and chants of the Peyote cult are related to the Plains Indians. The Navajo corn-grinding songs and the masked *yeibichai* spirit dance suggest Pueblo influences, although the eerie falsetto voice and light rattling used in the *yeibichai* dance songs are distinctive. The Apache songs for the masked crown dance fit the incisive musical beat to the angular postures and gestures of the four mountain spirits.

Pueblo music and ritual. The Pueblo peoples, from the Arizona Hopi and Zuni to the Tewa of the Rio Grande Valley, have in the course of hundreds of years evolved analogous musical arts with local differences, such as greater complexity in the westernmost villages. The most sacred songs for the appearances of the masked kachina rain gods resemble songs for unmasked, public, summer rain and corn dances. These songs (specifically as studied along the Rio Grande) combine terse, traditional musical forms with elusive tonality and rhythm. In the rain dances and so-called basket dances, the men provide their own rattle-and-song accompaniment. In corn dances they and the female dancers are supported by numerous separate singers and a chorus. The male singers start on a low, growling pitch, sometimes increasing in speed from a slow tempo. In the course of three or four melodic themes, repeats, and interpolated melodic fragments, they raise their voices an octave and a half. Finally, they subside to the growl, marking time all the while with the right foot on the powerful, pulsating accents. While the dominant beat is duple, the metre shifts at times to triple beat or to a slower beat, which they call *t'a*, or "pause," then returns to the earlier pulsation. The rain dances are simple and stationary; the performers of the corn dances move in elaborate designs within an allover pattern of a slow and a fast dance.

Pueblo music in sacred dances

Two-part song structures and elaborate dance patterns govern a dance type that shows influence from the Kiowa and Comanche tribes of the adjacent Plains. The buffalo and game animal dances, eagle dance, and Comanche dance precede the slow-fast parts with an entrance song of evident Kiowa origin. The dance songs resemble the typical American Indian style of pentatonic, descending melodies, in contrast to the pyramiding, rise-and-fall melodies and elusive tonality of the agricultural songs. The fast dances tie in with the Plains Calumet dances and Iroquois eagle dances, having a repeated pattern of a drum tremolo, followed by a section with a regular duple beat. The wintertime animal dances follow diverse Pueblo patterns, with a repeat of the dance in several plazas and with four appearances during the day. These dances require traditional songs for specified occasions and new creations within the frame for other occasions.

In addition, various Pueblo tribes perform an assortment of unclassifiable dance-songs, such as the San Juan Deer Dance with its Apache musical characteristics. For public shows, Pueblo Indians of Taos and other northern communities show their expertness at secular, clearly Kiowa war dances, hoop dances, and round dances. Most Rio Grande Pueblos also display Spanish influences. On saint's day fiestas they hold a Catholic mass, with Roman Catholic chants and hymns in English or Spanish. At Christmas and at other times they perform a *matachina* dance (a ritual dance for men), usually with fiddle and guitar accompaniment. At Santa Clara and Jemez, New Mexico, the accompaniment is with voice and drum, and the dance movement is in native style. The *matachina* dance relates the Pueblos to Mexican peoples—the Yaqui of Arizona and Sonora, Mexico. The polytheistic religion of the Pueblos resembles that of the ancient Aztecs and Maya. But their indigenous styles of music and dance appear very different from the aboriginal and contemporary styles of Latin America, even from those of the adjacent tribes of northern Mexico.

Relation of the Pueblos to Mexican peoples

REGIONAL CUSTOMS: MEXICO AND MIDDLE AMERICA

The cultures in Middle America include both the tribes of northern Mexico and peoples of southern Mexico who are grouped with those of Central American countries such as Guatemala.

Northern Mexico. *Yaqui music and dance.* The Yaqui of Arizona and Sonora have had much contact with Southwestern tribes. They relate to the Pueblos in their ceremonial *matachina* dance, using the same kind of longways formations (two parallel lines of dancers) and instrumentation (fiddle and harp) but without the Pueblo farce of a bull killing.

At Eastertime this dance forms part of a festive complex in church and atrium. A fantastic, Jesuit-Indian Passion Play bypasses the drama on Calvary. It consists of stations-of-the-cross processionals, with plainchant wailing by women; surging groups of black-robed *pilatos* (Pontius Pilates), and wriggling, jumping, masked *chapayekas* clowns to spastic fiddle scratchings. It also features the native dances of the deer and his pursuers, the *pascola* clowns, next to a group of coyote impersonators.

Yaqui Easter ceremonies

For their animal drama the Yaqui have evolved their own native orchestra—for the deer an ensemble with rubbing of mechanically precise notched sticks and with a drum consisting of a half gourd floating in a bowl of water, along with subdued singing of ancient, symbolic deer chants; for the *pascolas*, first a fiddle and harp in jig tunes, then a one-man flute-and-drum in irregular, pentatonic tunes; for the coyotes a low, gurgling chant to drum and the dancers' brushing across a bow that they straddle. The repetitious tunes with their limited scales and the medium-loud production seem to contradict the free and forceful dance style. By their restraint they blend with the subtle sounds of the dancers' rattles, deer-hoof belts, gravel-filled cocoons, and bells.

The various passive and active rattles of the deer dance

Interplay of dancers' rattles and music

interplay with the musical accompaniment. The cocoons swish with every step. The deer hoofs descend with the knee and foot in thumping steps, but in leaping steps their momentum causes them to syncopate—to work in opposition to the dance movements—to fly apart as the dancer descends and to strike together as he rises. Add to this the circular swish of one gourd and the alternately accented beat of the other. The *pascola* dancer shows a similar complexity, in the syncopated rhythm of the *sena'asom* rattle against the left hand. At times he vibrates it back and forth in front, with a motion of sowing seeds. He is probably replacing a native rattle by this brass-and-ironwood sistrum known in Mesopotamia and ancient Egypt.

The Yaqui have been given much attention because of their accessibility and their extraordinary ceremonial arts. They seem stylistically related to other, less publicized tribes of northern Mexico.

Other northern Mexican peoples. Northern Mexico's Uto-Aztecan tribes share with the Yaqui a medium-loud, nonpulsating song style, with repetitious, undulating melodies emphasizing the tonic and third in a scale using only the fifth, fourth, third, and tonic notes of a European major scale (as, g–f–e–c). The Seri Indians of coastal Sonora state show their affinity with Arizona's desert tribes in this singing style and also in their long, legendary narrative songs. East of the Sierra Madre Occidental Range, the Cora, Huichol, and Tarahumara have retained some examples of music in this same style. One might expect a link across the United States border in the *hikuli* ("peyote") songs of the Huichol, the alleged originators of the widely diffused Peyote cult, yet the Huichol songs are quite distinct from peyote songs in the United States.

Middle America. The northern boundary of Middle America is rather vague, lying somewhere in the central Mexican state of Jalisco, where the Indians are largely acculturated to Mexican society. The area from Jalisco to the Panama Canal displays a bewildering network of native musical styles in mountain and jungle, hybrid and modern music, and rural and urban styles. It would be convenient to assign the flute and drum music to native rites and the stringed music to Spanish fiestas, but there are exceptions. Flutes, not stringed instruments, accompany the European-derived dances of *moros y cristianos* ("Moors and Christians") and *conquista* ("conquest") dramas with clearly medieval tunes. These and other dances, such as the *santiaguitos* and the *matachines*, are derived from European ritual dances for small groups of men. Cacaphonous flute ensembles hail the birth, marriage, and death rites of the acculturated Tarascan Indians in Michoacán state. On the other hand, stringed instruments appear in Carnival and Lenten festivals, in the armadillo-shell lutes of the *concheros* dancers, and in the fiddle accompaniments for the *arrieros* ("mule-drivers") dance. The shrill, oboe-like *chirimias*, probably originally from the Mediterranean island of Majorca, accompany fireworks fiestas in Oaxaca. Again, the delightfully sour village bands present clearly Euro-American marches and polkas.

Every fiesta is a conglomerate, with, perhaps, a native *volador* (flying acrobat) performance; a variant of the *santiaguitos-matachines-moros y cristianos* dances; morality plays with chants and fiddles; and, finally, *huapangos*, lively couple dances with fiddles in jig tunes.

To further confuse the issue, contiguous tribal groups manifest contrasting styles. In southern Chiapas state, city dwellers delight in the Indian *chiapanecas* girls' dance and its music. The mountain Tzotzil Indians of Chamula and Zinacantán, Chiapas, have devised their own versions of stringed music, with handmade guitars and harps, and, again, contrapuntal combinations (*i.e.*, with two or more simultaneous melodies) of trumpet, reed flute, and drum. In the nearby swamps, the Lacandon, an isolated branch of the Maya, maintain their ancient, shamanistic chants, their conch-shell trumpets, and prehistoric drums. Again, descendants of the ancient horizontal slit-drums of the Maya survive in the city of Mérida, Yucatán.

A special oddity is the marimba, a wooden xylophone probably derived from Malaya via Africa, with primitive forms extending even into South America. The well-constructed marimbas of Mexico and Guatemala serve various purposes, from Guatemalan religious dances to regional and jazz dancing in plaza socials of cities like Mazatlán, Salina Cruz, Tehuantepec, in Oaxaca state, and Zacatecas in Zacatecas state.

Though it would be a vast simplification, it is not impossible to set off the native musical styles retained in mountains, jungles, deserts, and generally in rural settlements as distinctly different from the music encountered at urban political fiestas and tourist attractions.

REGIONAL CUSTOMS: SOUTH AMERICA

The same overgeneralization of rural–urban styles fits South American music, with qualifications. Native music survives among remote tribes of jungle and desert, but the highest Indian musical cultures remain in the Andes. Geographically, the distribution contrasts with that of Mexico, where native survivals surround the centralized cultures of the mestizos. In South America, simpler tribes have occupied the interior. The rural–urban distinction is similar on both continents, with Indian influence stronger in the villages and Iberian secular styles in the cities, especially on the coasts.

Musical traits of Amazon peoples. Amazonian tribes have proved to be fairly tenacious because of their inaccessibility. Some, like the Chaco tribes, have been losing their aboriginal arts. But other Amazonians display complex music for voice, percussion, and flute. The songs of two of the best known tribes—the Motilone of Venezuela and the Jívaro of eastern Ecuador—are medium-fast and generally medium-loud and show a preference for the interval of the third, as in songs of northern Mexico. Some Jívaro songs are pentatonic. The range varies from two tones to an octave. Frequently, the maintone lies in the centre. The rhythms are often complex and irregular, with triplets (three notes where two would normally lie) and dotted notes (two-note figures that either divide the beat into long–short values instead of notes of equal length or divide the beat into a snapped short–long rhythm known as the Scotch snap). The Motilone have some heterophonic songs, that is, songs in which the singers simultaneously vary the same melody in different ways. Such songs serve for amusement rather than for ritual.

Songs of the Motilone and Jívaro tribes

Andean music. Andean tribes, across the mountains from the Jívaro, also favour pentatonic scales, but they evolved a more sophisticated musical style during prehistory and in the period of Spanish colonization. They retain fragments of ancient ceremonial music, such as the invocation to the rain god on the Island of Taquila in Lake Titicaca—a repetitious tune for whistle and drum. The Pampa Hash, a district in Canas Province near Cuzco, Peru, perpetuates an ancient agricultural rite, with the flutes and chants both crossing each other and sounding in unison, or with chant accompanied by a background of two chords played on the *chillador*, a tiny, lutelike instrument. The elaborate, composite fiesta of the *chunchus* and *collas* near Cuzco combines native and hybrid components. A native prologue for side-blown flutes and drum is played on the eve of this continuation of the ancient Inti-raymi, the Inca festival of the solstice on June 24. A drama of *chunchus* ("Spaniards") and *collas* ("Indians") ends with the defeat of the *collas*, as in the Mexican *conquista* dramas. The music combines voices, flutes, conch-shell trumpet, and *chillador*. There is social dancing in the streets, and, finally, the *kacarpari* farewell music, for flutes and drum.

Among social dances, the *wayno* (or *huayno*), is preeminent. This paired longways dance (*i.e.*, in a double line of couples) in duple time sometimes uses a background of *sicuris*, or panpipes, their tuning corresponding to the Western G major scale with F natural. Highland mestizos prefer a combination of the *charango*, a five-stringed ukulele, with harp and other strings, sometimes with shifting metres, from $\frac{6}{4}$ to $\frac{4}{4}$. In either case, the melodies in pentatonic scales shift continually from a tonality suggesting a European major key in the higher

register to a cadence, or stopping point, in a tonality suggesting a European minor key. The harmonies change correspondingly.

Uncertain major–minor tonality of Andean music

This uncertain major–minor tonality pervades the Andean music. It appears in the *chinguinada* for church festivals, with a bass ostinato (repeated melodic-rhythmic pattern); in the mestizo *yaraví* serenade and in the mestizo *marinera* couple dance, with its metre alternating $\frac{3}{4}$ and $\frac{6}{8}$ patterns. Other, more urban dances, such as the *vals criollo*, use European major and minor scales. They fall in line with regional couple dances of Venezuela, Colombia, the Pacific coast through Chile, and inland to the pampas.

Mestizo music of the coast. The Creole people of the coast and plains have been replacing the indigenous music of the mountains with their own music, especially with the ubiquitous *cueca*, or *zamacueca*, couple dance. Their string ensembles, analogous to the Mexican mariachi bands, play to fit waltz steps or *zapateados* ("heel strikes"), with syncopations and shifts of metre. At times the instruments provide counter-rhythms with the gay songs in major scales. The lively, popular *pericón* and *firmeza* dances reveal very little Negro influence in their music, in contrast with dance music of Brazil and coastal Colombia; but they owe much to the *contradanza*, or longways couple dance, of Spanish colonial times.

Although North America's tribes did not mix native and European styles, the natives and mestizos from Middle America to the tip of South America have created infinitely varied hybrid musical styles.

STUDIES AND PUBLICATIONS

Musical fieldwork, recording, transcription, analysis, publication in books, articles, and records has been prolific since the turn of the century. The early musical transcriptions taken down by ear were not too accurate and musical scores based on phonograph-cylinder recordings leave much to be desired. The improved quality of disk recordings about 1930 greatly helped fieldworkers and analysts. The tape recorder, especially in its transistorized form, has boosted the studies of tribal repertoires before their extinction and during changes and revivals.

At first the field collector and the musical analyst were frequently different people, anthropologists enlisting musicians to transcribe their field recordings, sometimes with success.

Musical transcriptions and analysis. The techniques of musical transcription and methods of analysis are in flux. Various scholars have employed machines for tonometric and sonometric transcriptions. Such devices, based on sound-wave frequencies, show melodic intervals more precisely than normal musical notation. Nevertheless, most transcribers have found it convenient—and generally easily readable—to use the conventional European musical notation, with various diacritical marks to show divergent pitches and ornaments. Some transcribers have copied the scores to show analogous musical phrases next to each other, for convenient comparison. Usually they include the corresponding texts with translations.

In the realm of musical analysis there have been many experiments. The most useful method for melody analysis is the long-standing system of "weighted scales," according to which the scales used in given melodies are presented in terms of frequently and infrequently used initial and final notes and important intervals. Presentations of music in the form of graphs have been useful in the determination of tribal styles and in the comparisons of tribal and European tonalities.

Other fields of Indian musicology. One aspect of the study of Indian music still underexplored is the relationship of musical rhythms and structures to their associated dances, though some workers have included descriptions and diagrams of dances. The process of music and dance coordination is laborious, little used; yet it is important for the future.

Another aspect, the harmonization of native tunes and the composition of Indian ballet music, has been exploited but with questionable success. As in tribal performance of traditional music, modern study and use of Indian elements in contemporary compositions combines a knowledge of traditions with creative imagination.

(G.P.K.)

III. Dance

The dances of the American Indian peoples are comparable in many ways to the folk dances of Europe. They represent forms passed down over centuries and modified through interaction with foreign and other Indian cultures. The origins are similar, lying in religious rite; in attempts to invoke magic and thus cure illness or assure success in food production, hunting, and warfare; and in such life-passage rites as birth, puberty, and death.

Nowhere among the many diverse Indian cultures is there a sense of dance comparable to that expressed in the staged art dance of the West. A number of influences from European dancing passed into the American Indian dance, but in many places the descendants of the aboriginal inhabitants of the Western Hemisphere have retained the traditional dances.

The treatment of American Indian dance in this article is meant to focus first on certain general features of dance and their manifestation in a number of areas. The diversities existing within this larger framework then become apparent through consideration of the dances of the several cultural regions or tribal groupings. (More detailed comparisons between American Indian dance and the folk dance of European and other peoples will be found in the article FOLK DANCE.)

GENERAL CHARACTERISTICS OF AMERICAN INDIAN DANCE

Among the essential factors in an overall picture of American Indian dance are the diverse kinds of dance, the organization of the dances in terms of participation, and the relations of man and deity expressed in the dances. In addition, a variety of other stylistic considerations are relevant, as are the foreign influences that have been absorbed.

Extent of dance forms. Many themes, typically the celebrations of life crises, developed in the New World during millennia of residence, migration, and exchange. These were most prominent in the marginal cultures of California, Venezuela, and Tierra del Fuego in the most southerly reaches of South America. Mortuary rites were prominent in the northland and the deserts. War and hunt dances have had different degrees of prominence, their greatest development being among the hunters of the Great Plains of North America. So-called animal dances varied according to the local fauna, a tiger mime belonging to tropical peoples and a bear cult reaching across the northern part of the New World and into Siberia.

Religious magic, or shamanism, practiced by societies or individual priests is probably derived from such Siberian tribes as the Tungus and Chukchi. In various forms and used usually for cure, it extends to southeastern Brazil but is most potent and most trance-oriented among the Eskimo. From Mexico, and probably earlier from Peru, agricultural rites fanned out into the southeastern woodlands and the Southwest. More recent than the other rites, their forms show enrichment from Iberian rituals.

Patterns of participation. A distinction between performer and spectator has long existed in American Indian dance, though it is not the artificial separation that characterizes much of Western stage dancing. This latter condition has occurred only with the performance, largely in North America, of dances for tourists and during Indian participation in folk dance festivals or regional "powwow" gatherings.

Dancer-spectator interaction

Spirit impersonations, including maskings and noise, were used in widely separated areas to frighten nondancers. Specific instances of such practice included the puberty rites of the Yaghan and Ona of Tierra del Fuego; and among the Kwakiutl Kusiut of British Columbia in Canada, similar ceremonies were held in dance houses with a definite performing area. Except for a few specialized rites like the eagle and False Face dances, the change



Masked Iroquois dancer of the Cayuga tribe performing False Face Society dance, 1907. A turtle rattle is held in the right hand.

By courtesy of the Museum of the American Indian, Heye Foundation, New York; photograph, M.R. Harrington

of roles among spectators, dancers, and musicians is characteristic of the sacred ceremonies of the Iroquois longhouses in the eastern woodlands region of North America. White visitors are welcomed, especially into such dances for the Creator as the great feather and drum dances; and all Indians, from the aged to mothers with babies in arms, are expected to join in.

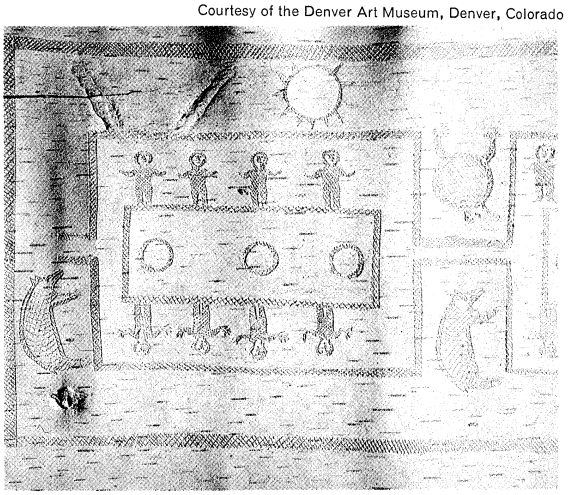
Among the Pueblos of the United States Southwest, the dancers remain separate because they require special rehearsals and ritual blessings. When they emerge from their sanctuaries, or kivas, onto the dancing plaza, they dance to invoke rain, health, and other blessings for the people from the supernatural spirits. After the ceremony the Indians often join in less formal social dances uniting all participants and observers. Though they have religious connotations, as among the Iroquois, these dances are secular, and anyone may enter or drop out at will.

Socially determined roles in dance. Vistors may not perceive the patterns of social organization reflected in the dances. It is clear that men or women alone begin some dances and the other sex may then join in, and that men monopolize some dances, women others. Less clear are the relations, especially complex in the longhouse dances of the Iroquois, between the moieties, the complementary divisions of the tribe based either on kinship or on ceremonial function. In all Iroquois dances, specific traditions decree the nature and degree of male and female participation and whether they dance simultaneously but separately or in various sexual pairs or other combinations. The leader of the dance and song and his helper, however, must be of different moieties, whether they lead from the floor or from the sidelines. When women enter a dance line, singly or with another, they must pair with a moiety opposite, or "cousin."

The Iroquois moiety pattern is crossed by another comprising various public or secret societies whose members are bound together for life, often joining the society during illness or other catastrophe. These societies perform such dances as the False Face curative rites, the female mortuary dances known as *ohgiwe*, and the dances of the sexually integrated Bear and Buffalo medicine societies. Elsewhere, religious dance societies were based on ages, as in the male warrior societies of the northern plains.

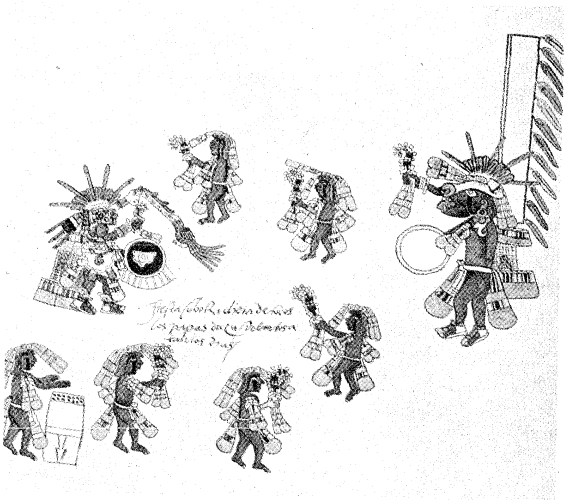
Some of these societies crossed local and even tribal boundaries, as in the extremely complex organization of Mexico's *concheros*, whose intertribal hierarchy runs from a *capitán general de la conquista de Tenochtitlán* through various local commanders and military ranks with specific duties to the attendant devil, sorcerers, and mythological figures. The *concheros'* claim to an Aztec heritage is given considerable credence despite some Spanish mixture, but, compared to that of the pre-Columbian Aztec, their ceremonial organization appears to be limited.

Courtesy of the Denver Art Museum, Denver, Colorado



Detail of Ojibwa birchbark scroll showing a moiety pattern ceremonial dance, c. 1875. In the Denver Art Museum, Colorado. Detail 46 x 41 cm.

By courtesy of the Newberry Library, Chicago



Aztec round dance for Quetzalcóatl and Xolotl, detail from a facsimile Codex Borbonicus (folio 26), c. 1520. Original in the Chamber of Deputies, Paris. 76 x 20 cm.

Early Aztec dances

In the organization of, and participation in, dance according to priestly or social status, the earlier Aztec dances were images of the completely theocratic government. A far cry from the democratic customs of the Iroquois, the circumscribed ceremonial roles of Aztec actor-dancers reflected the social structure comprising priests, nobles and warriors, commoners, serfs, and slaves. The priestly and noble-warrior classes took active roles in the many festivals of the Aztec calendar, and the priests trained noble youths for the priesthood or in dancing and singing. Warrior youths performed ceremonial combat, and the warrior orders of the eagle and of the ocelot fought captive slaves during certain festivals. Both commoners and serfs constituted the audiences, the former sometimes doing serpent dances with the nobility, the latter sometimes ceremonially attacked and routed by the priests. There were age roles and sexual roles as well, but the slaves, captives of sacred war with other city-states or purchased in the marketplace, as victims, had a passive role in the ceremonial activities.

Opposites played dramatically against one another in these rites: nobles and commoners, old and young, male and female. The warrior orders symbolized the clash of

the sky and light with the earth and darkness, and, as aggressors against poorly armed captives, they enacted the drama of sacrificer and victim. It was the priests and the passive slaves, however, who played the supreme moment of the ritual. The circle of social gradations was closed as the highest and the lowest ranks performed together the most crucial act of the Aztec dance-drama, human sacrifice.

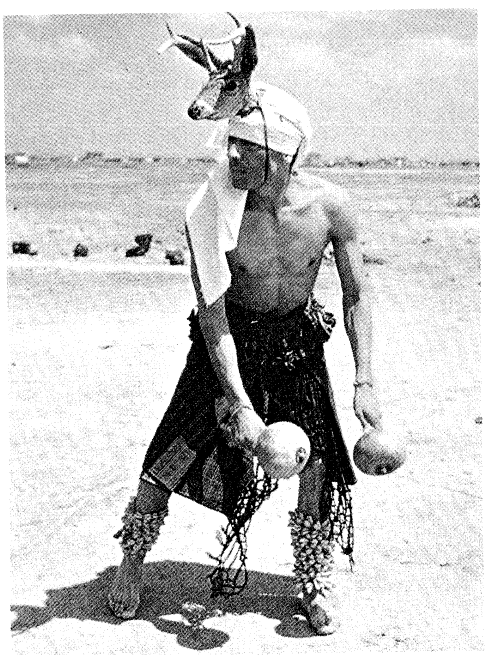
Religious expression in dance. Religious symbolism is significant even in the human interactions of the dance. Men often symbolize phallic, aggressive supernatural beings and rain-bringing deities, whereas women symbolize actual fertility. In Iroquois ceremonies, women represent the "Three Life-giving Sisters"—*i.e.*, the spirits of corn, beans, and squash, with no mimetic representation. Similarly, Pueblo women promote plant and human fertility by their symbolic dancing.

With no mimetic elements, the basket dance of the Tewa Pueblo rites includes invocations for plant growth and for the transmission of the gift of human life. The ceremony symbolizes the woman's central role in sustaining of the life of the pueblo.

Degrees of stylization of animal representations

In the animal realm there are also separate roles for men and women. Ottawa and Winnebago women imitate the winged flight of wild swans and geese, whereas the Iroquois and Pueblo men represent eagles. Both sexes join in the mime of supernatural bears and buffalo in ceremonies of the latter tribes, more realistically in Iroquois dances. In the Southwest, especially in the New Mexican pueblos, male representations of supernatural deer show gradations of stylization ranging from the naturalistic portrayals of Taos pueblo to the semistylization in the Santa Clara, San Ildefonso, Cochiti, and San Felipe pueblos, in which sticks replace forepaws, to the abstract, upright deer dancers of San Juan pueblo and masked, unreal deer in the kachina dance of the Hopi. The solo deer dancer of the Arizona and Sonora, Mexico, Yaqui, always a man, is relatively realistic, with mime of the hunt and killing.

Miguel Salgado



Yaqui deer dancer from Sonora, Mexico.

On the whole, in both Americas agricultural dances tend to be abstract, and animal dances are usually decidedly mimetic. The animal maskers of British Columbia are terrifying portrayals of supernatural beings. In Venezuela, masked beasts of the former Maipure puberty dance, *mauari*, threatened a pubescent girl and her cortege and had to be subdued magically.

Here and there the man–deity relationship is expressed in hand gestures. The Kwakiutl of northwest North America evolved codified ceremonial sign languages, as did the Pueblos, Aztecs, and Maya. In the San Juan Pueblo of New Mexico, the appearance of the rain gods is heralded by two ceremonial clowns using traditional gestures. Looking for the rain gods in the clouds, one of the clowns claps from his hands the ashes representing a cloud. He looks upward, shading his eyes to indicate his attempt to see into the distance. This gesture is always used whenever the clown speaks of what he "sees." The clowns repeat this action toward the four points of the compass, continuing to see the approaching rain gods, who bring with them the rain cloud. Similar performers may appear in the pueblo's plaza, outside the kiva, or sanctuary. Dancing, unmasked clowns enact motions of luring rain, of sowing seeds, of digging, and of gathering the plants as they rise from the ground. Clowns also appear in the men's spring dances and in the summer corn dances. After their entrance with a large group of male and female dancers, the corn dance singers station themselves in an arc near the drummers. They fit gestures to tunes and texts that are composed for each occasion but follow a traditional pattern and trend of ideas, beckoning to the rain gods in their cloud homes in the north, west, south, and east.

Invocations to the directions survive in the Great Plains and Great Lakes area, especially in the pipe dance. A single man offers a pipe to the thunderbird in the east, south, west, and north, moving clockwise, then to the deities of the sky and earth. Similar invocations to the directions survive in Middle America as fragments of the rich gestural symbolism of the Aztec and Maya. There, as in New Mexico, counterclockwise patterns emphasize the cardinal points.

Patterns and body movement. This religious, nature-oriented concept of space differs from that of Western man's folk and art dance, which has only geometrical or emotional significance. The geometric ground plans, however, show similarities with Western practices. The circling dances are sunwise in areas of former hunting people and countersunwise, or widdershins, among agriculturalists. Serpentine line dances also prevail among agriculturalists, among the Iroquois, Pueblos, Middle American, and Andean peoples. Among the Iroquois, many round dances are open, with a leader, coincidentally resembling dances of the Balkans of southeastern Europe.

Nature-oriented concepts of space

Aboriginal line dances are quite simple, whether they are single file or double file. Spanish influences are apparent, however, in the elaborations used in the double-file dances of the Southwest and Latin America. Spanish and Austrian influences probably inspired the couple dances of Latin America, for aboriginal dances juxtapose male and female partners only rarely, and never in overt courtship mime.

In body movement, the Indian style differs from that of northwestern Europe but coincidentally resembles the posture and quality of Serbian dancers. Characteristic of Indian dancers is a slightly forward-tilted posture, forward raising of the knee, flatfooted stamp or toe-heel action, and tendencies toward muscular relaxation and restraint in gesture. This basic style of body movement varies not only from area to area or from tribe to tribe but also from dance to dance and even from one individual to another. The agricultural dances generally are performed with an upright posture and an easy manner. Male war dances may include complex gyrations and flexion of the torso, as do animal dances. Vision and clown dances may induce bodily distortion.

Throughout Europe and the Americas, the posture varies with sex. Women tend to be more erect than men, to lift their feet and knees less, and in general to perform in a more restrained manner. Except for the war dances, Indian women use the same steps as the men, within the stylistic restrictions. In the woodlands of eastern North America, everyone proceeds with the stomp step, a flat-footed trot. In the Pueblo area, where men and women use a similar step, the dancers also specialize in a foot lift and solid stamp. In certain dances, especially clown,



Dancer of the Hidatsa Dog society, showing style of body movement used by North American Indian dancers. Aquatint by Karl Bodmer, 1834. 38.7 x 53.3 cm.
By courtesy of the Rare Book Division, the New York Public Library, Astor, Lenox and Tilden Foundations

animal, and war dances and in some social round dances, individuals often invent variants of the basic steps. Sometimes the innovators borrow American ballroom steps such as those of the Charleston, though they adapt them to their own styles. The steps and formations of the Indian dance, as well as the overall structure of a dance or ceremony, follow the music closely. This connection is covered in more detail in the foregoing section on Indian music.

Foreign influences. Among the influences from the Old World, the dances of northern Europe and the Anglo-American dances have found little acceptance. The Longhouse Iroquois reject all white men's dances. Among the few influences are some Oklahoma jazzlike, war-dance steps, an Indian two-step danced by couples, a waltz in a Pueblo social dance, and numerous couple dances of Latin America.

Introduction of Iberian elements

Iberia, on the other hand, has not only loaned some steps but has metamorphosed the dances of Middle America and western South America to Argentina. The hybrid dances reveal every conceivable shade of stylistic adjustment.

Adaptations of mazurka, waltz, and other European dance steps occur in some ritual dances as well as in such secular couple dances as the Mexican *jarabes*. The European origin, reinforced by the Europeanized music, is obvious despite the subdued manner of performance. The most significant dances are the religious dance-dramas taken over from such medieval religious productions as "The Battles of the Moors and Christians" and the *matachina* dances—both for trained male societies.

Negro influences on Indian dance are scattered—the *huapango* couple dances of Vera Cruz, Mexico, the carnival dances of the Black Caribs in British Honduras, the *tamborito* of Panama, and couple dances of coastal Colombia. Except for the Indian-influenced *candomblé de caboclo,* a ritual of the *candomblé* sect (a variant of the voodoo cult), the religious dances of Brazil comprise only African and Portuguese elements. Such popular Latin-American ballroom dances as the samba of Brazil contain no Indian elements.

REGIONAL DANCE STYLES

The most distinctive tribal dance customs originated in response to animistic religious beliefs—*i.e.*, that all objects and living things have living souls. The customs changed with prehistoric and historic migrations, with intertribal contact, and, since European contact, with upheavals in the way of life and thought. Although many dances became extinct, some survived European influences; others are amazing hybrids or new creations of the period after European colonization.

To give an accurate understanding of the role of dance in traditional Indian society it is necessary to examine both dances that became extinct as European influences weakened tribal customs, and dances that have survived, with or without European modification.

The Eskimos. The traditional shamanistic exhibitions and masked animal rites are being replaced by Western-style square dances in the frozen north. The most prominent ritual figure was the *angakok*, the shaman who communed with spirits by the rhythm of a single-headed drum and by ecstatic dancing, usually inside an igloo.

Whaling dances of the Eskimos

Until recently, Eskimos held elaborate outdoor ceremonies for whale catches and similar events. In Alaska, preliminaries included the rhythmic mime of a successful whale catch, with a woman enacting the whale. A sprinkling of ashes on the ice drove away evil spirits, and there were incantations and songs when leaving shore, when sighting the whale, and before throwing the spear, all of them songs that the "great kashak" sang when he created the whale. As the whale was towed in, Fox Islands men and boys danced naked in wooden masks that reached to their shoulders. At Cape Prince of Wales on the Bering Strait, the whaler's wife came to meet the boat in ceremonial dress, dancing and singing, and boys and girls performed gesture dances on the beach. Then inside a circle of large whale ribs, the whaler's wife and children performed a dance of rejoicing. On the west coast of Hudson Bay in Canada, communal feasting, dancing, singing, games, and shamanistic performances took place within a circle of bones or one of stones. The men's motions consisted of vigorous and angular arm jerking and jumping, the women's of curving gestures and swaying, with the torso and arms in a seated or standing posture.

Eastern woodlands. In the area from the Atlantic coast to about the Mississippi River and across the Great Lakes and the St. Lawrence River, dream, medicine, plant, war, calumet (ceremonial peace pipe), and animal dances predominate. Among the northern tribes of this area, mortuary and hunting rites are dominant; among southern tribes, corn, bean, and squash rites are most frequent. The recurrent dance pattern is a counterclock-

Photo Trends



Moros y cristianos dance-drama from Guatemala. The Moor is on the right and the Christian on the left.

wise circling by large groups, with a running step or stomp to antiphonal singing (alternation of two groups or of leader and group). Medicine rites are often exclusively for female or male members of a society, but dances for hunting or agriculture admit men, women, and children. In winter and in war or hunting ceremonies, men are the organizers and leaders; in summer and in agricultural ceremonies, women are featured performers.

Iroquois ceremonial dances

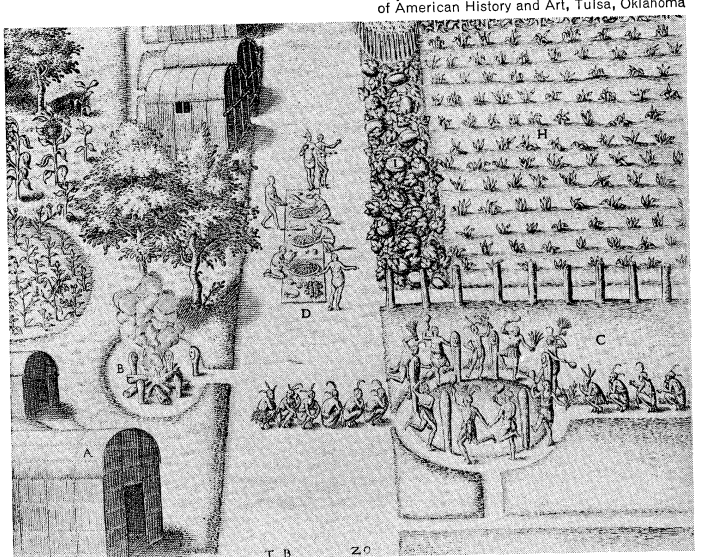
The Iroquois continue to maintain their ancient ceremonies and a large repertory of dances and songs, including rites for crises of life and for animals and plants. They also have acquired steps and dances from other tribes, especially those of formations in two straight lines. The Iroquois bear dance combines former hunting associations both with a clan-origin legend and with a curative society. When the bear spirit is displeased, he causes neurotic spasms in a person and must be appeased in a ritual at midwinter or in private summer ceremonies. The focal personnel consist of the patient and paired conductors, dance leaders, and singers from opposite moieties. Ceremonial songs and ritual offerings are followed by group dancing in which visitors and society members participate.

Although the Cherokee of the Smoky Mountains in North Carolina and Tennessee, a southern Iroquois people, have animal dances, they emphasize corn dance ceremonies. The Creek, Yuchi, Seminole, and other tribes of the southeastern United States greatly emphasize the summer green corn harvest ceremony, or Busk. Before the removal of many of those tribes to reservations in Oklahoma, they acquired few dances outside their own traditions. They carried the stomp circling to its utmost development by winding the line of dancers into a spiral or even into four spirals at the four corners of the dance ground. They differed from the northerners in preferring an outdoor dance setting.

Dance ceremonies of the Algonkian tribes

Among tribes of the large Algonkian family, survivors on the northeastern Atlantic coast of North America have long forgotten their tribal lore. The stomp dances performed until a few decades ago by the Penobscot of Maine and Narraganset of Rhode Island have experienced a strong revival. Algonkian tribes around the Great Lakes share many of the medicine and animal dance ceremonies known to the Iroquois, and the more

By courtesy of the Thomas Gilcrease Institute of American History and Art, Tulsa, Oklahoma

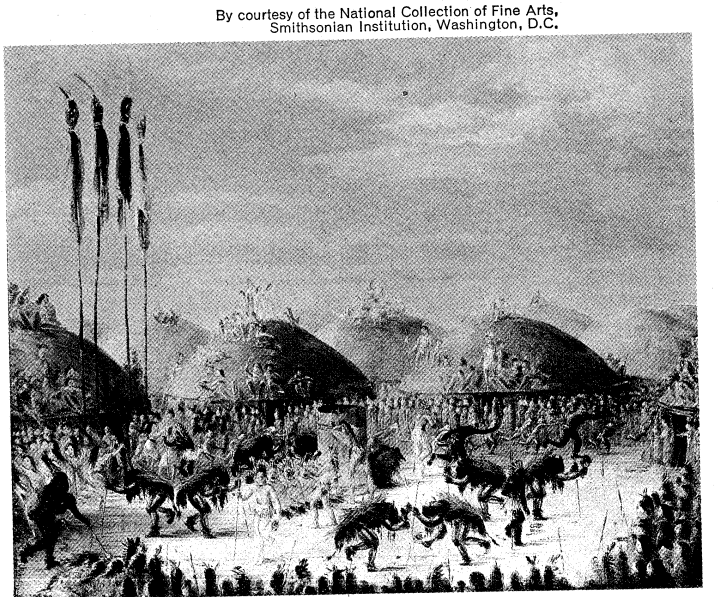


Crop fertility dance of an Algonkian tribe in Virginia, detail of an engraving by Theodor de Bry after a watercolour by John White, 1590. In the collection of the Thomas Gilcrease Institute of American History and Art, Tulsa, Oklahoma. 31 X 22 cm.

southerly groups hold corn dances. The Chippewa in the Upper Peninsula of Michigan and the Menominee and Winnebago of Wisconsin have maintained a hunting dance and a special wild-rice ceremonial danced in September when this crop is harvested. These groups show the influence of the adjoining Great Plains tribes in some of the circle dances, men's war dances, and buffalo dances.

The Great Plains. In the area extending from the Mississippi River to the Rocky Mountains and from Texas and Oklahoma into Canada, the dream dance ritual is made part of a visionary cult associated with boys' puberty and with a votive sun dance ceremony. During the one to four days' duration of the sun dance, usually held during the summer solstice, the participants abstained from food and drink. Dancers painted their bodies in symbolic colours and carried an eagle-wing bone whistle in their mouths. To a large drum and special songs, they circled in procession and saluted the sun with lamentation. They danced in place facing the sun, on their toes or the balls of their feet. Dancers continued until they fell unconscious or achieved their vision.

By courtesy of the National Collection of Fine Arts, Smithsonian Institution, Washington, D.C.

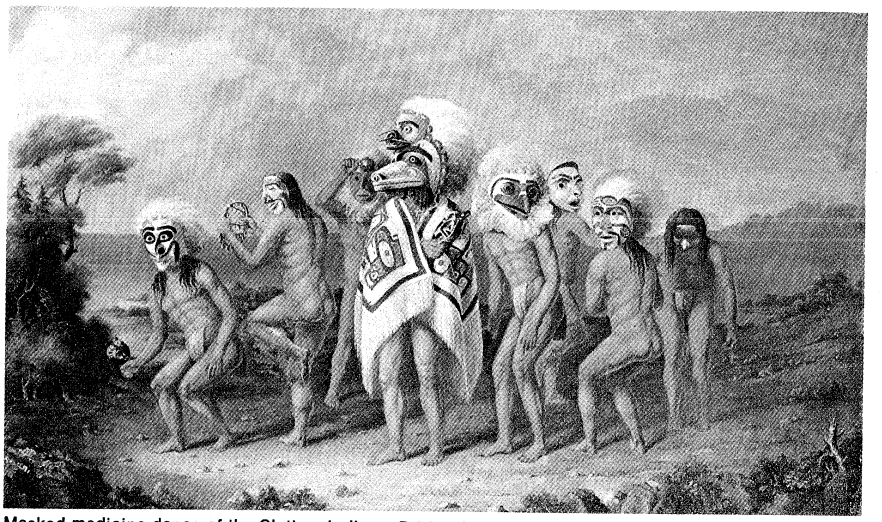


"Bull Dance of the Mandan Indians," oil painting by George Catlin, 1832. In the National Collection of Fine Arts, Washington, D.C. 70 x 57 cm.

The calumet (peace pipe) or peace dance originated in the tobacco rite of such northern Plains tribes as the Crow, Dakota, and other Sioux tribes. Its most elaborate development, however, was in the central plains ritual of the Pawnee and the neighbouring Omaha, Iowa, Ponca, and Osage. The war dance is organized into male war societies. Women have few such societies, although they perform a scalp dance. Animals are associated as tutelaries, or guardian spirits, both in the vision and war cults. The most spectacular hunting ceremonies, such as the bull dance of the Mandan Indians, developed from the economic significance of the buffalo herds. Buffalo rites merged with sun and war ceremonies and spread to tribes in other areas. The individual warrior, his prowess, and dancing skill were extolled as women progressed clockwise in a closed circle, with a sideward shuffle or bounce unlike the running step of the woodlands Indians.

The Northwest coast. Indian tribes along the Pacific coasts of Washington and British Columbia developed masked medicine dances and elaborate fishing ceremonies, such as that performed for a bountiful salmon catch. Their two most striking types of ceremonies are the potlatch, a feast and a dance for display and distribution of the host's wealth, and the midwinter initiation ceremony. Lasting several months in a special dance house, this rite initiates young men into a ceremonial society and includes many highly individual masked enactments of totemic spirits.

The northern desert and California. Southerly tribes, such as the Havasupai of the Grand Canyon and the related Yumans, developed agricultural dances. The Yuman Mohave stress cremation processions and cere-



Masked medicine dance of the Clallam Indians, British Columbia, Canada. Oil Painting by Paul Kane (1810–1871). In the Royal Ontario Museum, Toronto. 76 × 46 cm.
By courtesy of the Royal Ontario Museum, Toronto

monies, but, like the Navajo, they also have curative and animal dances with long song cycles. In this area, the vision quest ceremony is at its peak, and in southern California the Diegueño and Luiseño aided the vision by means of a narcotic, *Datura*. The Northern Plateau tribes, such as the Paiute and the Salishan, individually danced themselves into trances. In this area arose the Ghost Dance, a hypnotic circle dance that spread to the Great Plains in the 19th century. The ceremonies are frequently addressed to the spirits of the dead. There are also many two-line dances, especially among the Ute and southern Paiute. The innumerable small tribes of California shared some of the preoccupations with vision, cure, and death, as well as the seed and root gathering economy of the tribes adjoining them on the east. They specialized in elaborate masked ceremonies for the initiation of boys and less elaborate circle dances for girls' puberty rites. The more northerly groups also stressed exhibition of dexterity and costuming.

The Ghost Dance

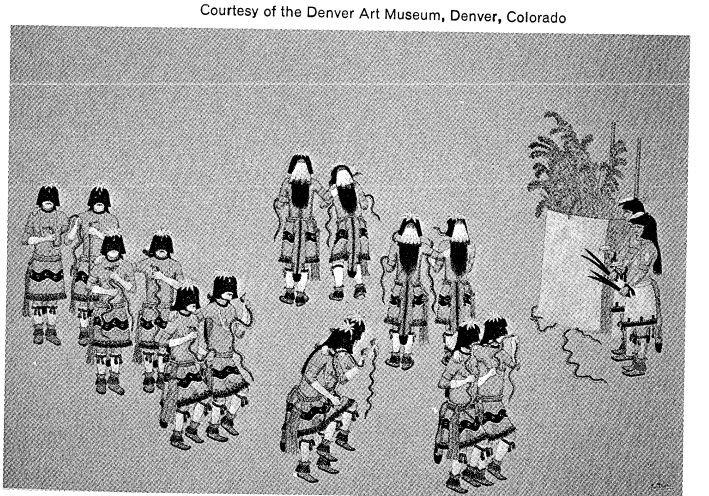
The Southwest. The semi-arid desert country from the Rio Grande west to the Mohave Desert of southern California and into northern Mexico and the southern Rocky Mountains is subdivided into three tribal areas: the Pueblo farmers along the Upper Rio Grande, the Zuni of New Mexico, and the Hopi of northern Arizona; the Navajo nomads, now turned shepherds; and the desert tribes that include agriculturists, such as the Pima, Papago, Yaqui, and former nomads, such as the Apache. The Pueblo dwellers of New Mexico and Arizona perform medicine rites and many winter animal and fertility dances. But the cycle of summer corn ceremonies and continuous prayers for rain form the core of their ceremonialism. The dances, organized by a male priesthood, are mostly well-drilled collective performances. Summer and winter clan groupings or moieties dominate ceremonies in alternation rather than through interaction as among the Iroquois. The most characteristic step is a stamp followed by a foot lift in a stationary line. This predominates especially in the very sacred dances held in the kivas, or sanctuaries. Semisacred dances in the village plaza add other steps and formations such as double lines, circles, and interweavings.

The most spectacular public dances of the Pueblos are the corn dances or *tablita* dances, after the women's tablet crowns with cloud symbols. They recur at various times during the spring and summer, with most pagentry after Easter and on the pueblo's saint's day. The Indians pay homage to the patron saint in an early morning mass and a procession to the plaza carrying his image, followed in the evening by a recessional to the church. By tradition each performance of the corn dance includes a slow and a fast dance. In the slow dance for entering the plaza, a chorus of seven to 70 older men shuffles across the plaza, singing and invoking the rain gods. A banner bearer leads a double file of 12 to 200 dancers, with a pair of men always ahead of a pair of women. For ten minutes they trot counterclockwise around the plaza. Following a pause, the singers form an arc, and the dancers line up face to face in two or four long files. They cross over, circle, and interweave in elaborate formations. Clowns meander in and out among the lines. The entire set is repeated at the other end of the plaza, and the group retires. The two moieties make alternate appearances. On the last appearance they combine, the two choruses singing simultaneously.

One of the most famous ceremonies is the snake-antelope dance of the Hopi in Arizona, a rite in which snakes are released in the four directions to seek rain. It includes swaying dancing to rattles and guttural chant, circling of the plaza with snakes, and ceremonial sprinkling of corn meal on the principal dancers by women of the snake clan.

The Hopi snake-antelope dance

Courtesy of the Denver Art Museum, Denver, Colorado



"Hopi Snake Dance," watercolour by Awa Tsireh, *c.* 1920.
In the Denver Art Museum, Colorado. 62 × 48 cm.

Masked dancers are a striking feature of Pueblo ceremonialism. The kachina dancers are sacred and represent the rain gods. Clowns with various names represent an ancient ritual heritage; in their black-and-white striped disguise of paint, they are eerie and also comical. Pueblo masking influenced neighbouring tribal dances such as the curative *yeibichai* of the Navajo. Curative ceremonies, with long song cycles, are emphasized by the Navajo, along with circular social dances, recalling those

of the Great Plains tribes. The Apache have developed a spectacular masked dance, called the *gahan*, to obtain cures but chiefly to celebrate a girl's coming of age. They also have rites for vision and divination, sometimes with the aid of a vision-inducing communal drinking ceremony. The male dance style is strong, angular, even acrobatic, while the women's style is subdued.

Mexico and Middle America. The triple teams of *pascolas*, or wooden-masked clown dancers, of the Yaqui Indians in Arizona and the Sonora desert in Mexico, descend from prehistoric clown-shamans. They dance without masks in semi-Spanish style, then, with masks, in an aboriginal, mimed deer hunt. By contrast, the *chapayekas* clown society recalls the Pueblo *tsaviyo* clowns in their antinatural behaviour and hide masks. The serious, vowed-membership society of the Matachini dancers ties in with the semi-Hispanic *matachina* dancers of the Rio Grande tribes and of the northern Mexican mountains. These dances are related to those further south such as the various types of *moriscas*. A fantastically hybrid Passion drama points toward Middle America.

A few indigenous dances survive in the mountains of Mexico. The circular *mitote* remains the ritual dance of the southern Tepehuan and other tribes of the Sierra Madre Occidental, like the Tepecano, Cora, and Huichol. Men and women skip in a counterclockwise circle, five circuits in one direction, then five in the other. A shaman accompanies with native songs, assisted by a musical bow on a gourd resonator. Formerly a deer dance followed the rounds.

The *hikuli*, or peyote dance, held in November, follows Huichol and Tarahumara pilgrimages for peyote. The dance of the Huichol is the more ecstatic. After consuming the trance-inducing peyote, men and women move in a counterclockwise progression, leaping jerkily and twisting their bodies.

The *rutuburi* is the typical ritual dance of the northern Mexican Tarahumara for three agricultural festivals—rain, green corn, and harvest—and for death and memorial rites. After triple invocations by a shaman, the women cross the dance space six times, then circle counterclockwise, holding hands and leaping with a stamp from left to right foot.

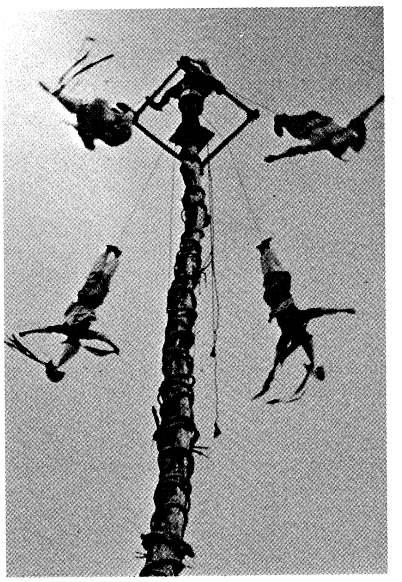
The *voladores*, or flying acrobats

Tribes of the Sierra Madre Oriental also engage in native survivals such as the *quetzales*, with great disc headdresses, and *voladores*, or flying acrobats. After ritual preludes to a fiesta, the flyers first dance around the pole with their musician and his flute and tabour. Traditionally there are four dancers, but the Otomí prefer six, including a man-woman, Malinche. They climb up a rope ladder and seat themselves on a small framework near the top, while each in turn dances on a two-foot central platform. During the flight songs they launch themselves into space with ropes tied around their waists and descend head down in 13 ever-widening circles till they reach the ground and land on their feet. The musician often performs special acrobatics, leaning back in an arc as he sits on the platform or jumping or pirouetting as he salutes the four directions. He may slide down one of the ropes amid acclaim.

The numerous clown dancers throughout Middle America are usually associated with various rituals. Some impersonate animals, like the *tecuanes* (tigers) of Guerrero state. Some, like the *viejitos* (little old men) of Michoacán state, combine native and Spanish features. Some, like the *catrines*, wear ragged modern outfits and poke fun at the bourgeoisie. Always connected with the dance-dramas held in conjunction with religious fiestas, the clowns are not simply comics but also social satirists.

The semi-urban Indian fiestas of Middle America include three types of European-influenced dance-dramas: the *morisca* and its variants, *moros y cristianos* (Moors and Christians), *santiagos*, *matachini*, and related plays; Passion plays; and *posadas*, *pastorelas*, and Guatemalan *loas*. These are often true dramas, with dance and music. The dialogue may be in Spanish or in an Indian language or a mixture of both.

Moriscas, ritual dances deriving from ancient European ritual dances, abound during carnival (the week before Lent begins) but also occur at Corpus Christi (second Thursday after Pentecost), Santiago (St. James Day, July 25), and on other feasts. In the drama of *moros y cristianos*, two factions mimetically tangle in arguments and battle with the ultimate victory of the Christians and the conversion of the Moors.



Voladores of Mexico.
By courtesy of the Museo Nacional de Antropologia, Mexico City

Posadas and *pastorelas* are danced episodes of the Christmastime *coloquio de los pastores*, or "Play of the Shepherds." Most popular in southern and central Mexico, New Mexico, and Texas, the *posadas* are generally processions by city boys and girls who go from house to house asking gifts and lodging and singing special hymns. On Christmas Eve, youngsters dance quadrilles of *pastores* and *pastoras* (shepherds and shepherdesses). The Guatemalan *loas* are short religious dramas presented at Christmas, during Holy Week, and on other holy days.

Secular couple dances

In form, the regional, secular dances are distinct from the ritualistic dances. Except for the circular *mitote* types, they are all for couples who do not touch. Among the best known are the *jarana* of Yucatán, the *zandunga* and *llorona* of the Zapotec Indians of Oaxaca state, the *chiapanecas* of Chiapas state, and the *huapangos* (called *fandangos* in some locales) of the east coast. The *jarabe* has many regional variants, as the *jarabe tapatío* (or Mexican Hat Dance) of Guadalajara, the *jarabes* of Tlaxcala and Michoacán states, and the *zarabanda* of Guatemala. Sometimes the theme of flirtation or female coyness blossoms forth in humorous interludes. Contests of improvisations to *la bamba*, widely danced in the Mexican Gulf Coast area, also contribute to the merriment of the Veracruz *huapango*.

These regional couple dances blend European and native steps and styles and use European or European-derived instruments and tunes. Even though they are mestizo dances, Indians enjoy them and give them a native flavour—*e.g.*, in steps or posture. The dances' European origins must usually be deduced from the style—Andalusia for the *zandunga*, the flamenco *fandango* for *huapangos*, and the *jota* of Aragon for *jarabes*.

South America. South American dances resemble those of Middle America. Native dance rituals have their last strongholds in the jungles and highlands of the interior. Iberian dance-dramas show their relationship with those of Middle America. Couple dances are prominent.

Northern South America. In Venezuela several tribes of the Orinoco River held masked puberty rites. For example, among the Maipure and Baniva tribes, Mauari, the spirit of evil, is impersonated by a dancer who is fully covered with red and black body paint, a face-covering of puma or jaguar pelt, and a crown of deer antlers. At the initiation of a youth or girl, he emerges from the

forest with maskers representing lions, tigers, deer, bears, and other wild beasts. Their bloodcurdling growls and howls mingle with the groans of the *botutos*, the sacred trumpets, to fill the night with a gruesome din. With wild leaps and contortions, they dance around the neophite and four shamans.

Ancient puberty ceremonies evidently had wide distribution, with distinctive features, in the Amazon Basin and in the Mato Grosso highlands of Brazil, the Gran Chaco region of Bolivia and Paraguay, and Patagonia and the Tierra del Fuego in Argentina. They formed part of a complex emphasizing shamanistic cures, death rites, and animal dances for the hunt.

Today the most prominent dances of Venezuela are the many versions of the *morisca* on Christian holidays and a dance of medieval devils on Corpus Christi. The *joropo*, a lively couple dance in waltz time, is the national dance of Venezuela.

Profusion of courtship dances in Colombia

Colombia has fewer religious celebrations and a greater profusion of courtship dances. The *joropo* extends into eastern Colombia. On the Caribbean coast the *bullerengue*, *lumbalu*, and the circular *cumbia* mingle Indian reticence and Negro ebullience. The Colombian *fandango* derives more from Spanish diversions. The national dance, the *bambuco*, originated in the Andean zone. Male and female partners, waving kerchiefs, enact a courtship mime of pursuing and flirting, combining dignity with sensuousness.

The Andean region. Along the Pacific coast and in Ecuador, Peru, Bolivia, and Chile, native dances have received Spanish influences. On Catholic holidays the northern Andean Indians perform vestiges of aboriginal animal rites for the vicuña, the tiger, and the condor by a solo mime within a large circle. Conveniently, Corpus Christi synchronizes with the Inca solstice ceremony, *intiraymi*, and presents an excuse for the reappearance of the native sun god in a huge, gold disc headdress.

The Araucanians of Chile, who resisted Incan influences, preserve a shamanistic harvest ceremony, the *ñillatun*, a combination of Christian ritual and an indigenous mass dance. During interludes two men mime ostrichlike rheas, with shawls as wings.

The highland fiestas of Andean Indians show more European influences than the *ñillatun*. Generally timed in accordance with Catholic festivals, the dances feature battles of Moors and Christians, clowns, demons in fantastic masks, and European animal characters. Some dramas ridicule the Spanish. The mountain fiestas often conclude with gay couple dances.

The coastal celebrations feature widespread couple dances of mixed Indian-Spanish origin. Only the *cumbia* shows Negro qualities. The mournful *huayno* (or *wayno*), *yaraví*, and *sanjuanito* of Ecuador, Peru, and Bolivia reveal Incan origins in their restrained manner and haunting music. The gay *marinera* of Peru and the headstrong *cueca* are more Spanish. The *fandango* and *pasacalle* are also at home in the urban ballroom.

The southern plains. Argentinians have developed such ritualistic mestizo (Spanish-Indian) dances as *las cintas*, a maypole dance, and the *sumamao* ("beautiful river") celebration. Argentina shares some Andean social dances, as the semi-Indian *carnavalito*, a collective circle dance. The richest repertoire of Argentina and adjoining Uruguay developed among the cowboys, or gauchos, of the Pampas. Their dances reveal more of the Spanish elements than those of the Andean regions. In the *pericón*, the dancers manipulate handkerchiefs; and in the *pericón*, *chacarera*, and *gato*, couples perform *zapateados* as groups. In the *bailecito*, *cuando*, *firmeza*, and *cueca*, they enact courtship mime with emphatic waltz steps. Other dances are urban, as the *milonga* of the lower classes and the sophisticated Spanish-French tango ballroom dance. Andeans and Argentinians have exchanged dances. In fact, the diffusion of dances is much greater in this area of South America than in Middle America.

STUDY AND EVALUATION

The secular pan-Indian dances of North America, such as the Oklahoma dances, the round and war dances of plains tribes, and the stomps of southeastern tribes all have in modern times spread from coast to coast—to the salvation of tribes who lost their heritage but want to perform something Indian, especially in urban gatherings. The most copious and reliable materials on these and other aboriginal dances are strewn through the works of anthropologists, folklorists, and a few musicians. General descriptions are often incorporated into anthropological studies and into notes on earlier observations by colonists, missionaries, and 19th-century scholars. The reports of Frank Speck on Cherokee dance are highly discerning. William Fenton collaborated with Gertrude P. Kurath in many Iroquoian field trips and publications. Kurath and James Howard collaborated on studies of the Manitoba Indians for the National Museum of Canada. Joann Kealiinohomoku, a young anthropologist with extensive dance experience, is studying the Hopi dances of Arizona, and William Powers is an expert on Siouan dances.

Lacking a major specialist, Middle American Indian dances are in large part unstudied in a scholarly manner. But in South America, a dancer, Delia Zapata Olivella, is working on the dances of Colombia, with a musicologist, George List. In Peru a folklorist, Mildred Merino de Zela, is organizing a study of native and creole dances. In Argentina, Carlos Vega has published a monumental work, *Las danzas populares argentinas*.

Essential to future studies is an examination of the arts in their cultural context. It is equally important to recognize the dance as an expressive art, to learn and analyze the movements, and to present them in dance notation alongside musical scores. Such presentation facilitates intertribal and intercontinental comparisons. The materials must stem from field work, but they can be supplemented by the many motion pictures in college archives and museums and in repositories such as the Wenner-Gren Foundation and the American Philosophical Society.

(G.P.K.)

IV. Visual arts

NATURE AND ELEMENTS

The role of the artist. The very use of the word art suggests one of the basic differences between Western and American Indian concepts. For not only did few Indian groups allow art to become a major way of life, as in the West, but many Indian languages even lack a term for "art" or "artist." If one wished to refer to a beautiful basket or a well-carved sculpture, it was usually necessary to rely upon such terms as well-done, effective, or perhaps powerful (in the magical sense). And the concept of an artist was largely of a person who was simply better at the job than was another. Generally, artists were accorded special significance only where wealth was a major factor in the culture. Wealthy chiefs, particularly on the Northwest Coast, hired stables of carvers for their personal aggrandizement, giving rise to a class of professional artists. Yet, the fact that the talented artist was in such demand did not set him apart as a person of undue importance; rather, his role was important at the moment his product was in demand—and if he happened also to be a religious, political, or war leader, so much the better.

Although Indian people, unlike Westerners, may not consider artistic skill in terms of a superior individual attribute, the difference between a well-woven basket and a careless piece of work or a particularly well-designed carving and a crudely made example did not go unnoticed. Fine workmanship commanded a premium, even in pre-European times, and with the advent of the dollar system, it was even more highly prized.

Collective versus individual art. The basic role of the American Indian artist is the same as that of the artist in any culture: to arouse an emotional response in his audience. In Indian cultures, the artist's ability to communicate successfully with his own people depended largely upon his recognition of the force of tradition. The social organization of the various tribes allowed less latitude for experimentation than Western cultures and usually compelled him to work in familiar channels.

Role of tradition

(The frequent failure to communicate with a non-Indian audience does not necessarily mean that the Indian artist has failed as an artist but that many of the cultural, racial, social, and political traditions involved in his work have proved to be almost insuperable barriers to understanding.) Yet, within this rigid framework of tradition, there was sometimes a surprising degree of freedom of expression. There are recorded instances of individuals having made considerable changes in the art (and the economy) of their tribes. Perhaps the most striking have been the careers of Nampeyó (1859?–1942), the famed Hopi potter, and María (1881?–) and Julián Martínez (1897–1943), of San Ildefonso Pueblo. Through sheer individual talent these people achieved a personal triumph by developing a style that not only was copied by other artists but in time also was regarded as "traditional" in that particular village. Although there is no way of knowing how often this happened in the past, there are suggestions that it occurred at Mimbres, among the Haida slate carvers, and quite possibly in some areas of the so-called Mound Builder cultures of the Southeast.

Origins of Indian decorative designs

The origins of most Indian decorative designs cannot be traced accurately today; most of them are lost in antiquity. Many obviously came from natural forms, while others are simple developments of geometric or lineal motifs. Some have become so interwoven with alien concepts—Western, after the advent of the European, for example—that it is impossible to completely unravel their sources. There is evidence, however, that some of the original forms were creations of individual artists and were often inspired by what the Indians called a vision. The vision is closely akin to a dream. To the Indian, the dream world was a mysterious, supernatural land, where his soul, which had left his body, participated in many strange activities and saw many strange sights. Since many of the designs or creatures that he saw were regarded as protective forms or beings, sent as signs from the spirit world, he would carefully re-create this dream world during his waking hours. If he were not an artist, he would occasionally describe his dream creature so that the proper individual could record it on hide, in wood, or in stone. But since these supernatural vision designs were extremely personal, they were usually recorded by the individual himself; hence, they vary tremendously in aesthetic quality.

Because art designs were regarded as personal property, an artist could buy a design or receive it as a gift from its creator, but to appropriate and use it for his own purposes was taboo. This custom amounted to quasi-religious and social copyright.

Importance of utilitarian considerations

The function of art. Since the early Indian peoples travelled extensively, their household goods had to have the maximum utility; thus, Indian art objects are basically intended to perform a service—for example, to act as a container or to provide a means of worship. Only in later years could these peoples afford the sophisticated luxury of household possessions that did nothing but "look nice."

The particular utilitarian form that Indian arts take often reflects the social organization of the cultures involved. Political and military societies seem to have found their major art forms in the world of weaponry, pageantry of costume, and panoply. This is most pronounced in the Plains, Aztec, and Inca civilizations, all of which reflect the dominant warrior cult forces active in their arts. Those cultures in which life was heavily governed by religion—the so-called theologies of America—tended toward a greater degree of ceremonial art than those in which life was less ritualized. All of the aesthetic expressions that have come down from the Maya, for example, obviously reflect the considerable weight of theocracy that existed in their world.

Generally, but not necessarily, the best of Indian artwork was applied to those objects intended to please a deity, soothe the angry gods, and placate or frighten the evil spirits. Through such means, the Indian sought to control his environment and the human or supernatural beings that surrounded him.

Some specific articles were reserved solely for religious uses, and some were for secular needs alone. Decoration does not always provide a clue as to these uses. Some of the most highly revered religious articles are completely plain—in fact, may be rather ugly—while others are highly embellished. Sometimes plainware bowls were used for food preparation, while other peoples did the same chore in polychrome bowls. Many objects served a dual function: normally, they were used for everyday household purposes, yet under a different set of circumstances they could fulfill a religious function. Beneath the surface, there was a magic at work, and, in initiated hands, a mundane article might release its supernatural power, calling upon unseen forces to aid the owner. This power might be visually evident in the form, shape, or decoration of the object, or might simply be "believed in" no matter what the physical state or appearance of the object might be.

Supernatural powers of art objects

The aim of the Indian artist was not merely to set down realistic records but to create the semimagical designs so common in primitive art. He quickly realized that he could not draw a tree as perfectly as it could be made by the Creator; so, with common sense, he did not try. Instead, he sought the spirit or essence of the tree and represented this in his design. Carvings, paintings, effigies, or "realistic" portraits are not simply pictures *of* people or objects; they embody the essence of that particular subject as well. In a sense, this can be called a form of witchcraft, since it captures the soul of the object or the person. This semimagical character of Indian art is difficult for the Western mind to understand. Not infrequently, the non-Indian will ask, "what does that design mean?"—a question that makes little sense to the artist. He might reply, were he so inclined, that it meant what he wanted it to mean. Also confusing to the questioner is the fact that the Indian often attaches names to designs, largely for convenience. When asked, an Indian will frequently call a given design a "leaf," or an "arrowhead," when what he actually means is that the design is "leaf-like," or "leaf-shaped," and so on. But the inquirer immediately translates this to mean that the design signifies a leaf or an arrowhead and tries to impart a "story" to the overall visual concept that may have no basis in reality insofar as the original artist is concerned.

Ritual was often interwoven into the very process of creating Indian art. The Westerner, whose attention is devoted primarily to the product and not the process of art, is satisfied to define the Zuni war god Ahayuta as wood sculpture of a religious nature; other details are simply added ethnographic data. To the Zuni, however, it is vitally important that the wood came from a tree that had been struck by lightning and that it was ritually prepared for carving. This is equally important in the instance of the False Face mask, which must be carved from the trunk of a living tree—hence the term "live mask." The tree is ritually addressed before the carver begins, and the mask and the tree are "fed" tobacco before separating the two. Such prescribed ritual is of equal, if not more, importance than the artistic skill employed in the production of the work. If the ceremonial acts were ignored, the article would lose its efficacy—and might even prove dangerously counteractive. This ritual aspect of Indian art, which permeates most of the ceremonial paraphernalia, is extremely complex and must be considered throughout the creation of the work of art.

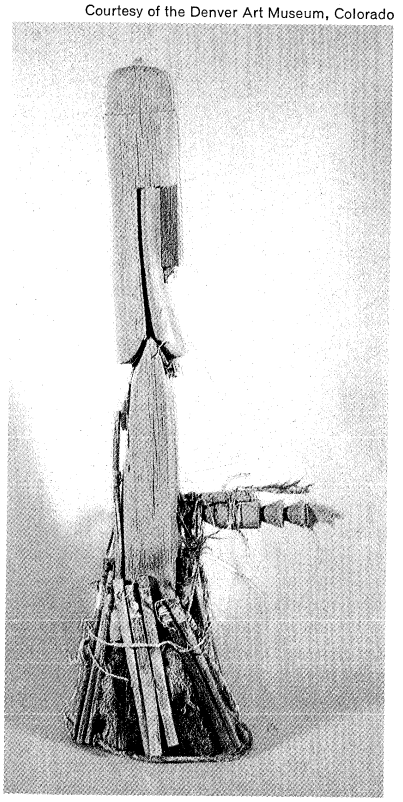
Not all Indian art, however, was religious or political. There was also a considerable amount of mundane, humorous, and even profane art produced by most cultures. Although much of the eroticism has disappeared in the Puritan fires that continue to burn the Westerner, sufficient examples remain from prehistoric and recent times to indicate a wholly relaxed freedom of expression reflecting a healthy, naturalistic outlook.

Materials. Working in the materials natural to their respective homelands, the various Indian cultures produced art that reflected their environment. Those peoples living in heavily forested regions, for example, inevitably became gifted sculptors in wood; those for whom clay was a major resource became skillful potters; and those living in the grasslands became fine basket

weavers. There is virtually no medium, however, that has not been explored and mastered by the Indian: jade, turquoise, shell, metals, stone, milkweed fibre, birch bark, porcupine quills, deer hair, llama dung, sea lion whiskers—all were used by the artist to lend colour or texture to the finished product. In many instances, such materials became desired commodities, to be traded over great distances; for certain objects were not regarded as "official" unless they were manufactured from a prescribed material, and, for religious purposes, a substitute could not be tolerated. Often, in such cases, the materials

Role of art materials in trade

Courtesy of the Denver Art Museum, Colorado



Wooden figure of the Zuni war god Ahayuta, *c.* 1900. In the Denver Art Museum, Colorado. Height 70 cm.

became, in a sense, currency, for they achieved a standard value within the economy, with ready acceptance wherever they were in vogue.

The relationship between material and design in Indian art was quite different from that in Western art. The Western painter usually imposed a design on the artificially limited surface of a flat, rectangular canvas; and the sculptor, following predetermined spatial arrangements, imposed a shape on his material. On the other hand, the Indian painter and sculptor were less likely to force their materials to conform to a design. They tended instead to adapt their design to the natural outlines of their materials, which often happened to be a complete and therefore irregular buffalo hide, a tree branch, or a stone. This naturalism is one of the most pleasing aspects of Indian art and often demonstrates the artist's remarkable ability to incorporate the natural form into his composition.

REGIONAL STYLES OF AMERICAN INDIAN VISUAL ARTS

The term "Indian art" covers an extremely broad category, encompassing all art expressions of the original inhabitants of the Americas and their cognate descendants. It thus includes not only varied and completely disparate cultures but also spans great time sequences—from the mid-20th century back to prehistoric times. (Surviving artifacts clearly demonstrate that ancient man was already possessed of considerable aesthetic ability; flint, for example, was carefully flaked into attractive, well-balanced forms; and stone carving and pottery were capably handled.)

Although the dissimilarities between the artistic expressions of different cultures and different times are great, there are also similarities; for the borrowing of art forms from distant and occasionally alien peoples was a common practice. Objects in museum collections reveal, for example, that prehistoric merchants travelled thousands of miles seeking and bringing back from other tribes such materials as feathers, shells, jade, and turquoise. This far-flung trade expanded the limits of tribal styles, for the Indian merchant returned with new ideas as well as materials. In time, new designs and motifs became part of the stylistic concepts and traditions of his own people. Intertribal marriage, too, affected regional styles. While in some tribes marriage within the group was required, in others it was forbidden. In the latter case, the girl—who was often the artist—took her own art traditions to the new group, into which they were subsequently incorporated.

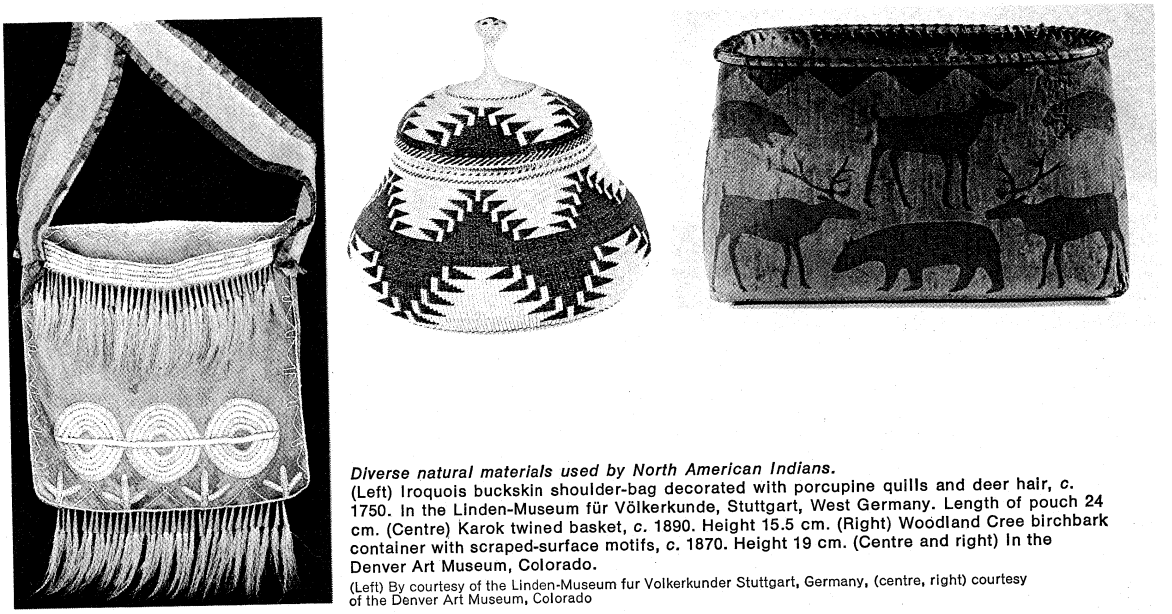
It is becoming increasingly evident that there were common forces at work, even if widespread in time and space. There are certain symbols that are widely encountered, and some would seem to have had similar significance over a wide area. It is likely that trade routes or political hegemony levied the major influences upon this phenomenon, and it may be said that the depth and spread of symbols is relatively equal to the migration of the peoples involved. In Middle America, for example, the so-called Plumed Serpent motif is to be found in one form or another in almost every culture, and this motif extends even into the United States, where it is encountered in visual form as well as in legend. The existence of the feline deity in South America, from the south up into the northern Andean region, is another instance of the travel of an idea and a visual element. Certain customs also have enjoyed wide acceptance; for example, the role of trophy heads, the use of masked personations, and winter solstice New Fire ceremonies. And each of these customs was accompanied by related visual art expressions.

Despite the similarities between the art forms of different cultural groups and different times, one cannot speak of Indian art as though it were a single concept. Just as there were several hundred native tongues, dialects, and speech forms, so were there an equal number of tribal styles, motifs, and design forms. In trying to establish a common aesthetic bond, the well-schooled researcher generally finds as many differences as he does similarities.

When two completely different peoples move into a common area, such as the migration of the Athabascan Navajo into the Pueblo Southwest, the eventual result may be a melding of cultures, the loss of certain ancient individualities—since each contributes to the new expression—and the emergence of new aesthetic qualities. It is not certain just how skilled the Navajo weavers were when they arrived in the Southwest, but the Pueblo people were highly developed in that art. Subsequently, the Navajo not only learned new weaving techniques and designs but in time also improved upon the Pueblo methods, transferred the sex role of the weaver from male to female, and matured as far more sophisticated craftspeople. On the other hand, under the same circumstances, surprising differences can sometimes be found; for example, while the Hopi and Zuni people live almost side by side and under similar cultural conditions, it is quite possible to identify the art products of both without great difficulty. This is equally true of cultures in ancient times, such as the Aztec and Mayan, or in another time and another region, the Sioux and the Crow.

It is in those tribes or cultural entities that at one time were part of a whole but have subsequently split off that one most often finds common themes, art elements, and cultural patterns so similar as to be confusing.

To most readily understand some of the artistic impulses active among the Indian tribes of the New World, it is convenient to take them in geographical sequence, north to south.



Diverse natural materials used by North American Indians.
(Left) Iroquois buckskin shoulder-bag decorated with porcupine quills and deer hair, *c.* 1750. In the Linden-Museum für Völkerkunde, Stuttgart, West Germany. Length of pouch 24 cm. (Centre) Karok twined basket, *c.* 1890. Height 15.5 cm. (Right) Woodland Cree birchbark container with scraped-surface motifs, *c.* 1870. Height 19 cm. (Centre and right) In the Denver Art Museum, Colorado.
(Left) By courtesy of the Linden-Museum fur Volkerkunder Stuttgart, Germany, (centre, right) courtesy of the Denver Art Museum, Colorado

North America. The aesthetic products of the prehistoric Indian artist in North America are perhaps the least well known to the non-Indian public. This is partly because these early people left few spectacular architectual ruins as compared with their Latin American cousins. This is not to say that architectural monuments did not exist. Spanish accounts report that great temple mounds were in use in the Southeast at the time of the first European entry, in the mid-16th century. But most of these structures were of perishable wood and have long since disappeared—as have most examples of the great use of colour (in which the ancient folk gloried) and their tremendous range of textiles.

So many Indian materials were perishable that scholars have little by which to judge their arts and must, in effect, evaluate a people by only a small proportion of their achievement.

American Southwest. It is only in the Southwest that any major prehistoric architectural sites are to be found. The monumental stone cliff dwellings that remain are eloquent testimony to the culture that existed there. Progressing from a simple pit house through aboveground homes, these people moved out onto the plateau regions of Arizona and New Mexico and built remarkable multistoried structures, some—such as Pueblo Bonito in New Mexico—sheltering hundreds of families in over 400 rooms.

Pueblo Bonito

These apartment houses were well suited for the demands of their environment; their walls were of stone, clay, and sand mixed as an adobe—although true mixed adobe in block form did not come into use until the coming of the Europeans. The thick stone walls provided excellent insulation, being warm in winter, and cool in summer. Heights reached to seven stories, although most villages were of three or four levels.

Major divisions of these early Southwestern Indians, often called by the generic name *Anasazi*, include the Hohokam, of southern Arizona; the true Anasazi, of northern Arizona–New Mexico; and the Mimbres, of southwestern New Mexico. In addition to these groups—each of which produced a style of its own distinct from all others—were dozens of lesser subgroups that archaeologists have been studying for decades in an effort to assemble the pieces of this giant jigsaw puzzle.

The people living in the pueblos produced some of the most successful Indian art work. They were masters of weaving, painting, and particularly clay. Their weaving techniques long antedated the arrival of Spanish sheep; a native cotton provided ample fibre for intricate weaves coloured with native dyes. Mineral and vegetable pigments provided colourful decorations when applied with a fibre brush to wood, clay, or to white-plastered walls in a fresco technique. Fortunately, abundant kaolin deposits yielded high-quality clay for the creation of excellent pottery forms. Although small stone effigies have been found, sculpture was not a highly developed art form. Pueblo art is essentially linear in design and reveals a preference for applied decoration. The large underground kivas, rooms used for religious purposes, were decorated with murals executed in brilliant mineral-pigment colours.

Pueblo art became a strongly conventionalized art, held to relatively rigid forms. This characteristic was determined, no doubt, by the closely knit communal nature of a culture that depended upon close cooperation for survival. At its best, early Southwestern art is marked by technical competence and fine control of line and form; but it reflected little experimentation, tending more to rework set patterns in many intricate designs. Because of this conservatism, the culture of the contemporary Pueblo Indian shows striking parallels to that of his ancient ancestors.

Conventionalized forms and technical competence

In more recent times, the richest variety of art expressions is to be found in the Southwest. Whereas in other regions the arts flourished and waned, here they are still active forces in the lives of the Indian people. Almost all of the crafts practiced in prehistoric times are still practiced today, along with some newly introduced expressions.

Following early trade routes, merchants brought new ideas to the Pueblos, departing with the finished products they had sought. Thus, new creations continued to develop, and new markets strengthened—as continues to be true today. Yet, because of its essential conservatism, Pueblo art remains a closely related continuum of ancient techniques, affected more by the introduction of new tools than of new ideas.

Along the same trade routes came invading tribes from other regions, particularly the Navajo and Apache, who subsequently settled in the Southwest and in time surpassed their teachers in certain arts that they adopted, improved upon, and made their own—notably, silversmithing and weaving. Whereas Pueblo weavers once dominated the textile field, today the work of the remarkably inventive Navajo weavers is world famous. Silversmithing, another famed Navajo art, is more recent; it was only in 1853 that the first Navajo smith took up the tools of his craft.

As in the prehistoric era, Southwest sculpture has failed to develop as a major art form. The most active sculptural work in the Southwest is reflected in the carved and painted cottonwood kachina dolls of the Hopi and Zuni, which have enjoyed wide popularity. Many variations of these wood carvings are also found in altar and shrine figurines, which are not produced for commercial consumption.

The crafts of basketry and pottery are moderately active. Both face the problems of economic balance: the problem of the time invested as against the economic return of the craft object exacts a serious price. By and large, a craftworker today can enjoy a more stable income from manual labour, or by clerking in a store, than from any creative activity. This is particularly true if one includes the time required to go into the desert, find grasses or pigments, then wash, grind and process them—all preparatory to the tedious hours spent actually forming the products for later sale. It seems certain that, in time, the former will entirely disappear, and only the luxury-scale pottery arts will continue. Even now, very little pottery is made for native use; it is largely intended for the outside market. Although pottery and basketry are produced in much smaller quantity, the work is of better quality than in years past.

Specialization in Southwestern art

Specialization has long been a factor in Southwestern art and has become increasingly so in recent years. Certain tribes produce almost all of the small carved fetishes, or tiny drilled shell and stone beads; the Zuni favour intricately worked silver jewelry with tiny turquoise settings, while the Navajo make use of massive silver castings with heavy turquoise sets. The Navajos also make most of the heavy rugs and textiles, while the Hopi supply lightweight ceremonial kilts, sashes, and similar costume fabrics.

Another art form that may have been brought from the north, but that was more likely adopted from Pueblo culture is sand painting (more accurately termed dry painting). The use of a variety of finely-ground mineral pigments, which are allowed to trickle through the fingers to form a variety of complicated patterns, has become uniquely Navajo. These designs provide a focus for curing ceremonies.

Great Lakes-Midwest. The existence of rich textile art in the prehistoric Middle West is known, but its range and development are lost in hundreds of years of history from which few examples survive. Examples of basketry and wood are similarly rare. Enough of these perishable items have survived to indicate that these people had mastered the arts, but not enough to enable scholars to judge their aesthetic development. What has survived in profusion is stone, worked skillfully and in many ways. Pottery, too, though not of highest quality, and copper ornaments have been found. Of the relatively perishable substances, finely carved and incised shell is common, which, along with bone, indicates the artistic range of these early peoples. The quantity of objects found is impressive. Numerically significant groups, the peoples of the region were active in the production of materials and implements with which to meet the challenge of their environment. The function of all of the recovered examples of stonework is not known, but it is known that much of the archaeological wealth was ceremonial in nature, indicating a highly organized civilization.

"Effigy mounds"

Ritual structures existed, such as the so-called "effigy mounds"—great piles of earth fashioned in various forms to represent animals. The Serpent Mound in Ohio is an example of this custom. Truncated pyramids apparently served as large bases for wooden "temples," now long vanished but still in use when Spanish explorers first entered the region. Cahokia, in Illinois, was the largest such earthwork in existence before it was bulldozed to make way for modern buildings.

Major cultural expressions from this region included the Adena, Hopewell, Oneota, and copper culture peoples; their art was extensive, making great use of sculptured stone pipes, polished ornaments of both stone and copper, and incised shell decorations that may have served both as inlays and clothing accessories.

Great Plains. The later Great Plains region is the area most familiar to the average non-Indian, for this is the world of the "Buffalo Bill" shows, television and movie programs, and fiction. From it came the buckskin and beadwork costume, feathered war bonnets, colourful porcupine quill decoration, and painted shields that personify the American Indian in the minds of most people.

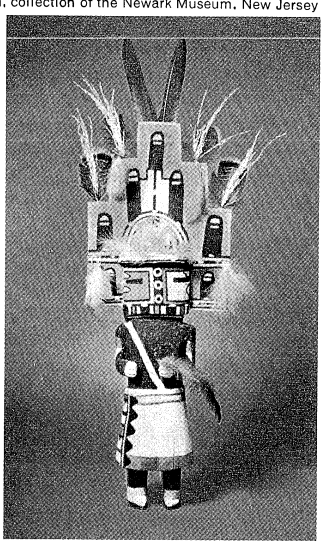
Yet, here was no monolithic culture. The arts of the Great Plains Indian varied from tribe to tribe; some peoples seem to have had superior aesthetic taste, with sensitive and inventive developments in the arts.

Very little woodcarving was produced here in proportion to the other arts; yet a respectable body of wooden bowls, clubs, effigies, figurines, and similar objects indicates that the Plains artist did not ignore this medium. Even less pottery and basketry was produced, for containers were primarily made from buffalo hide.

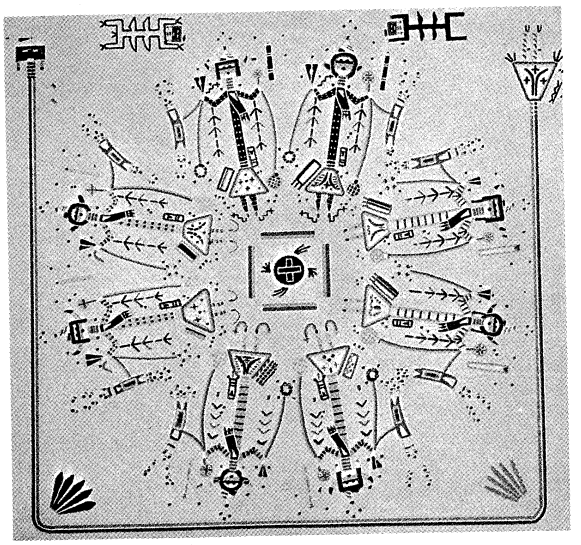
A great deal of Plains art served both decorative and spiritual ends. A given design might appear to be primarily a colourful decoration, yet to the initiated it was also the guardian spirit of the owner.

(Right) By courtesy of the Museum of the American Indian, Heye Foundation, New York; (left) photograph, collection of the Newark Museum, New Jersey





Southwest Indian arts.
(Above) Hopi kachina wooden doll, c. 1925. In the Museum of the American Indian, Heye Foundation, New York City. Height 64 cm. (Left) Classic Navajo blanket, c. 1855–65. In the Newark Museum, New Jersey. 110 x 156 cm,



Navajo sand painting, c. 1940. In the Denver Art Museum, Colorado. 115 X 111 cm.
Courtesy of the Denver Art Museum, Colorado

Colour was originally achieved by mineral pigments or vegetable dyes. In time, these were supplanted by commercial dyes and trade colours. Porcupine quilling—the use of small quills of the American porcupine, *Erethizon dorsatus*, which are flattened, dyed, and then applied to the surface of animal hides or textile materials—is an art produced nowhere else in the world. It became outmoded in favour of glass trade beads, which were not only technically similar in their application to quillwork but did not fade and gave a richness of colour unobtainable in any other way.

The art forms themselves range from realistic to extremely abstract and symbolic. Often they are narrative in content, as with the Winter Counts, those painted records that recounted tribal history by means of annual symbols, and the personal history paintings on hide that recount the exploits of the owner.

Not only did the Plains Indian decorate his home but also his person, with carefully coiffured hair, facial painting, and clothing enhancement. And he devoted the same aesthetic attention to his horse as he did to himself, creating beautifully decorated gear for special occasions. Statically displayed in a museum exhibit, much of this ornamentation loses the grace of motion. When worn as intended, the motion of the wearer and the wafting of the plains breeze gives the feathered war bonnet or the fringed buckskin a lively grace and colour.

Peripheral North America. Four other art-producing regions of North America include the far west, northeast, central south, and southeast.

Prehistoric art of central south and southeast North America

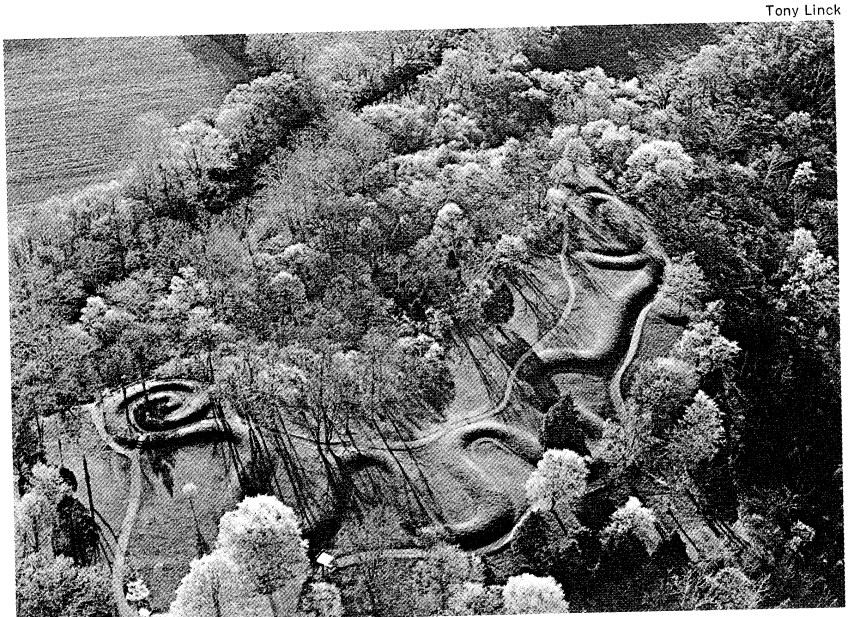
In prehistoric times, the central south and southeast were part of the most artistically exciting region of the North American continent. This land of temples, mounds, and monuments was an amazing world, and one can truly understand the legends that grew up around the riches that were evident when the Spanish arrived and that are still found in archaeological excavations. Testifying to the highly developed civilizations that existed are the beautifully carved shells, incised conches, and intricately decorated clothing ornaments; the carved stone effigies of ancestor figures or deities, which suggest a strong affinity with ancient Mexico; and the many bird and animal pipes in museums throughout the country. Had the Middle Mississippian culture diorite bowl (Museum of the American Indian, New York City) found at Moundville, Alabama, been the only masterpiece to survive, however, no other proof of the artistic brilliance of these peoples would be required.

Wood was used in profusion, although little of it has been preserved in anything resembling its original condition. A quantity of textiles, albeit in fragments, has also survived. Other perishables include decorative freshwater pearls, featherwork, bone, and animal hides.

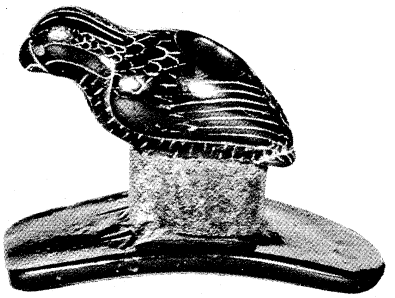
But it is in the claywork that the greatest vitality seems to have been expressed. While much of the clay used was of inferior quality, the results were astonishing. Exuberant forms, delicately traced surface lines, and strong, powerful designs were all executed with a confidence and grace that still attracts contemporary art students. A tremendous assortment of vessel designs was created in the southeast: floral, geometric, clay appliqué, delicate polished water bottles and huge burial jars, as well as many lovely vessels created to hallow a shrine, decorate a temple, or do homage to a god—all providing evidence of the imagination, skill, and sheer love of clay for its own sake that these early potters must have felt. While much of this pottery art is extremely fragile, enough of the remarkably large output has survived to give an excellent idea of the aesthetic heights that were attained.

With the coming of the white man, this creativity was ended or diverted. Tribes were killed off or dispersed, or their social organization was so disrupted that normal pursuits were destroyed or directed into other, non-Indian channels. While the introduction of new and better tools allowed greater technical proficiency, the pride and dignity that had formerly provided such a strong creative

Tony Linck



Adena Serpent Mound near Locust Grove, Ohio, c. 1st century BC. Length 405 m, height 1–2 m.



Hopewell culture pipestone pipe, *c.* 300 BC–AD 400. In the Museum of the Ohio Historical Society, Columbus. Length 7.5 cm.
By courtesy of the Ohio Historical Society, Columbus

impetus was no longer present. The artist had lost his old market—purchasers who understood what they were buying—and instead served a customer more concerned with external appearance than with the function of the object. The result was what is disparagingly known as tourist art—ostentatious elaboration that had little to do with the integrity of the product.

Today, almost all of the aboriginal arts of the Southeast have been lost or are much less actively pursued. The great stone sculpture for which it was so famous has entirely disappeared, although excellent wood sculpture is a continuing art; pottery is quite different from the earlier styles and far less imaginative. The clays from which contemporary Cherokee and Catawba pottery is produced are better grade, but the range of design is markedly limited. The most active art, and probably the most successful, is basketry, in which the present-day artists are in every way equal to, or better than, their tribal predecessors.

The breakup of the great Cherokee tribe, when a major portion was forced to remove from South Carolina and Georgia to Oklahoma in 1838–39, destroyed the unity of the people. Today, very little Cherokee culture as such is to be found in Oklahoma; it has been kept alive largely by the smaller Eastern band still living in North Carolina. It is this group that is responsible for the pottery, wood sculpture, and basketry referred to above.

Basketry of the California Indians

The great art of the California Indians was basketry; no other people in the world has produced such a wide variety of superb basketry. The Pomo, Hupa, Yurok, and Karok peoples of the north developed basketry to its ultimate with weaves so tightly composed as to provide a watertight container, baskets so small that they measure less than one-eighth inch (three milimetres) in diameter, huge grain-storage baskets, and delicately woven "gift" baskets with the feathers of birds interwoven that provided not only opportunity for the weaver to demonstrate her mastery of the art, but also a means whereby she could display her affection for the deceased. Elsewhere the Chumash, Mono-Paviosto, Washo, and Koso proved no less skilled.

The Woodlands tribes of the East fall into two divisions: the south–southeast (discussed above), and the Great Lakes and northeast. The Great Lakes group produced various arts, including woodwork, a style of weaving with rush and hemp, and a strong porcupine quill art, later replaced by beadwork. This style of beadwork was popular around the turn of the 19th century, when large quantities of it became available. The art depended upon a weaving frame, which allowed the manufacture of long strips, useful for necklaces, belts, panels, and headbands. Wood art made effective use of burls (hemispherical outgrowths on a tree), from which bowls and containers were fashioned. Pottery was almost nonexistent.

The people of the Northeast, notably the Iroquois, are famous for their False Face Society masks, quillwork and beadwork, wooden bowls and ladles, and the woven wampum belts, which are important historical documents. Some pottery was produced, but not of significant quantity or quality. Woodlands basketry was common, but it was not of the quality found elsewhere. Primarily a splint-weave type, it was rarely ornamented, and when it was, the ornament consisted of stamped or painted vegetable-dye designs.



Middle Mississippian diorite bowl in the shape of a crested wood duck, from Moundville, Alabama, *c.* AD 1500. In the Museum of the American Indian, Heye Foundation, New York City. Length 25.4 cm.
By courtesy of the Museum of the American Indian, Heye Foundation, New York

By courtesy of the Museum of the American Indian, Heye Foundation, New York



Cheyenne painted buffalo hide, depicting the Battle of the Little Bighorn, *c.* 1878. In the Museum of the American Indian, Heye Foundation, New York City. 116 × 87 cm.

Eskimo. It may seem unlikely that art would occupy very much of the attention of the inhabitants of the bleak Arctic regions; not only is there little raw material to work with but the ever-present need to secure a food supply would seem to leave little time for craftwork. Yet, from this harsh environment came some of the most imaginative and humorous of Indian carvings. During the long winter nights, the Eskimo had ample time to work the ivory that came from the walrus and whale, much as the Yankee farmer passed the time during the winter whittling on a stick of wood as he sat by the stove.

Art styles of the area favoured carving in the round, decoration by incising, and a modest amount of inlay. Since the basic material was often a tusk or a tooth, these objects partially dictated the form, which was embellished after carving by incising or engraving. Black pigment, from charcoal fires, was rubbed into the lines for emphasis. Such prehistoric wood carving as may have existed has almost entirely disappeared, but enough has survived for us to know that it was a rich art form. Ancient ivory carvings have also been excavated, revealing a sophisticated, formal style. The "fossil ivory" from which these carvings were made is highly prized even today and, when found, is invariably turned into beautiful carvings that gain value because of the scarce, richly colourful raw material.

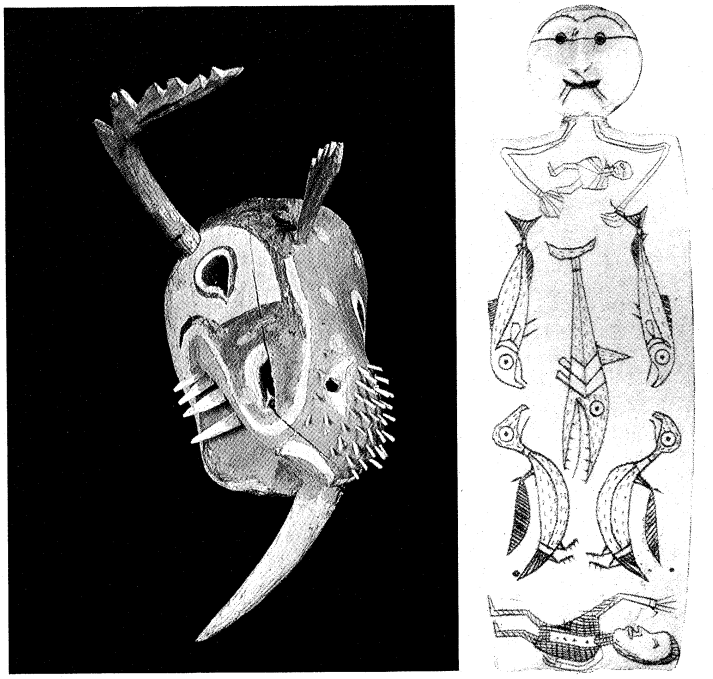
Humour in Eskimo art

A predominant characteristic of Eskimo art is the warm sense of humour that is so prevalent. Sometimes it is expressed in caricature, sometimes in sequential, "cartoon"

form. Its surrealistic expression is probably a reflection of the Eskimo's awareness that, because he lives close to death throughout his life, humour is vital to his psychological health.

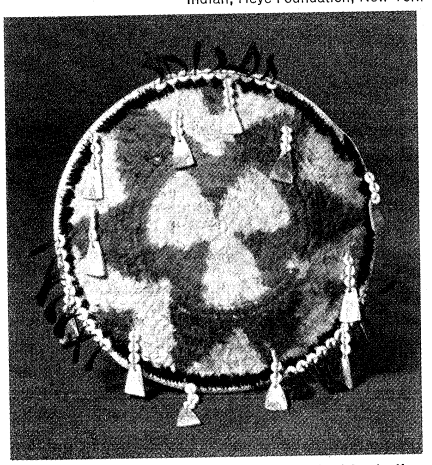
Another significant feature of Eskimo art is the remarkable mechanical skill that was often involved in the creative process. Part of the Eskimo's artistry was his ability to neatly piece together small parts to create a whole—and his ability to fashion the tools needed to carry out the operation, many of which were works of art themselves. This skill is evident in the Eskimo's most famous art form: the fantastic wooden masks used for various dances and social affairs. While many tribes made wooden masks and decorated them with colourful ingenuity, no North American aboriginal people developed the art of imaginative characterization to such an extreme. Surrealism *par excellence*, these masks demonstrate a combination of realistic, imaginative, and supernatural qualities that is uniquely Eskimo.

In more recent times, with its inception around 1950, a stone art form, utilizing deposits of gray and green soapstone, or steatite, found in the vicinity of Hudson Bay, has become familiar to art collectors. Usually given an artificial colouring, these examples of small-scale sculpture are popular examples of genre art. While they reflect the inherent sculptural skills of the Eskimo, they owe their origin and promotion to non-Indian agencies that have worked closely with several of the crafts groups



Eskimo art objects.
(Left) Painted wooden mask of the Kuskokwim Eskimos, 1875; height 31 cm. (Right) Incised walrus ivory shaman's figure, Kinugumiut Eskimos, *c.* 1890; height 9 cm. In the Museum of the American Indian, Heye Foundation, New York City.
By courtesy of the Museum of the American Indian, Heye Foundation, New York

By courtesy of the Museum of the American Indian, Heye Foundation, New York



Pomo feathered gift basket decorated with shell pendants, *c.* 1890. In the Museum of the American Indian, Heye Foundation, New York City. Diameter 22 cm.

in the region. A form of graphic art derived from Japanese printmaking techniques has also become popular in this way.

Northwest Coast. It was in this region, richly endowed with tremendous cedar and spruce forests, that the American Indian sculptor achieved his finest expression. It is probably here that the influence of tools upon the artist is best exemplified; for, with the introduction of steel cutting knives, the Northwest Coast artist was free to demonstrate his latent talent in the aesthetically superb sculpture that is rivalled by no other Indian people in North America.

Tall, straight cedar poles furnished the material for the huge totem poles, the smaller wooden figures, the masks, and the other carved objects so loved by the Northwest Coast Indian. Inlaid with *Haliotis*, or abalone shell, and carefully painted, these products took on a quality so distinctive that they are immediately identifiable.

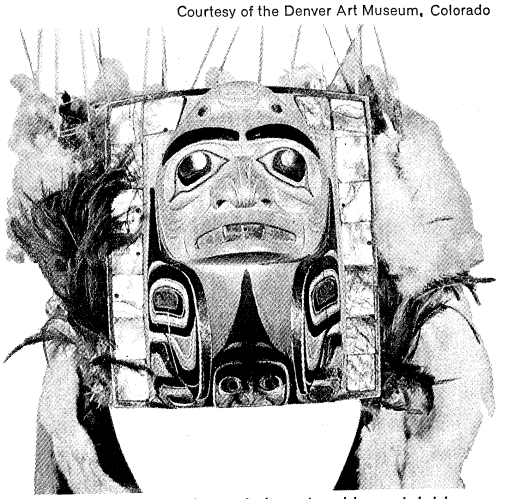
Another remarkable quality of the Northwest Coast artist is his skill and interest in fitting designs into forms. He excels at fitting his designs into a given area, shape, or prescribed form, yet without sacrificing the integrity of the design.

Totem poles

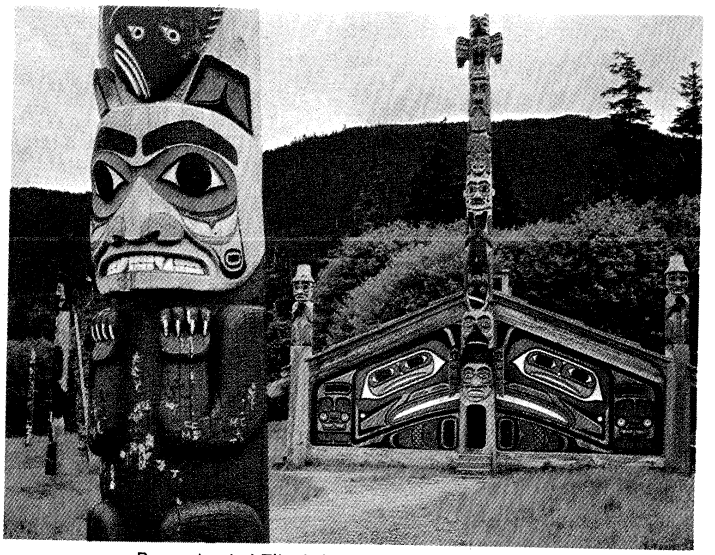
The role of the tall totem poles from this area has not been well understood by non-Indians, and many erroneous accounts have been published as to their purpose and "meaning." In actuality, they were not religious and were never intended to be worshipped. They were historical documents of a sort, recording the social position, wealth, and relative importance of the person who had paid for the pole. Because family lineage, class status, wealth, and other social facts were thus recorded, it was possible to gain an "introduction" to the village chief or house owner by simply examining the tall pole.

The goal of most of this rich art was the exaltation of the individual—more specifically, a wealthy village chief or a great noble, for the society was based on a class system. Part of the insignia of social position was the accumulation of wealth, and *objets d'art* were an important part of that wealth. With the coming of the white man, who coveted the rich furs of the region, the control of the great fishing areas and strategic position of the Northwest Coast tribes enabled them to acquire staggering wealth in an extremely brief space of time. The exis-

Courtesy of the Denver Art Museum, Colorado



Haida painted wood and swan'sdown headdress, inlaid with abalone, *c.* 1870. In the Denver Art Museum, Colorado. Central section 15 × 18 cm.



Reconstructed Tlingit long house with totem poles. In the Totem Bight Totem Pole Park, Ketchikan, Alaska.
Bob and Ira Spring

tence of an Indian purchasing class, with its ever-increasing need for impressive possessions, created a supplier: the professional artist. This was thus one of the very few aboriginal cultures that gave rise to art patrons who hired artists on a commission basis.

The short existence of works of art

More surprising, the works that were commissioned were destined to be destroyed, given away, or otherwise neglected. For the life goal of many of these tribes involved the belief that the greatest value was to give away all of one's possessions. While this may seem paradoxical the logic was simple: the more one gave away, the greater one's prestige.

Combined with the potlatch (a murderous game of swap), this system tended to impoverish groups from time to time, sometimes to the point of enslaving whole villages. The rule was that the wealthy individual hold a feast to which his major rival was invited. The host would give away all of his possessions, and the rival would return the courtesy, giving back more than he had received. To show even greater contempt for material possessions (and one's rival), frequently goods were publicly burned, broken up, or cast into the sea. Slaves were even killed and families sold into slavery. From all of this megalomania came the surviving aesthetic masterpieces that are valued so highly today.

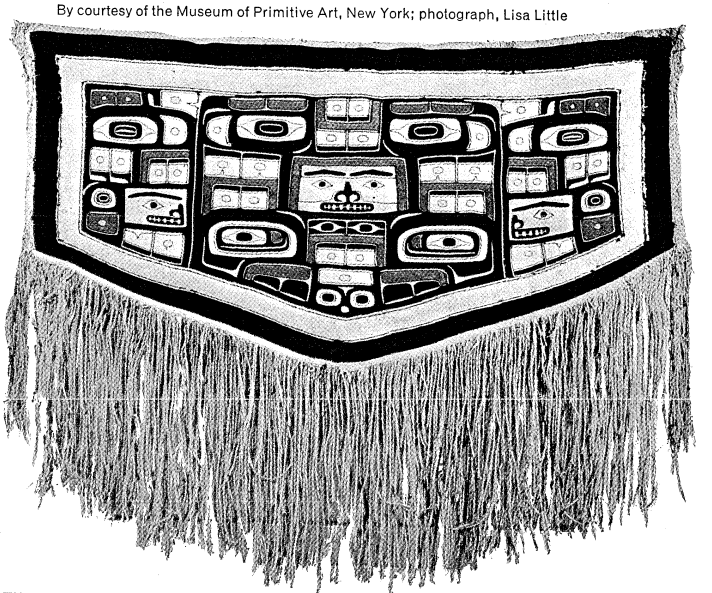
The Northwest Coast tribes were among the first American Indians to master metalcraft. While some copper came from local sources, most came from whaling ships, both as cargo brought in for trade and as peeled from the hulls of wrecked ships. This metal was worked with great skill by Tlingit and Haida artists into fighting knives, masks, overlays for art works, and the great shield-shaped *tinneh* that were so highly prized.

Among the Northwest Coast tribes, the Tlingit people of Alaska seem to have produced the most sensitive and sophisticated sculpture. The Kwakiutl, on the other hand, expressed their feeling for line and form in extremely impressive and powerful painted carvings: the designs are usually outlined in strong colours; there is far less subtlety of form; and the overall feeling is of a potent force at work. Between the two extremes are the Haida carvers, whose work, often less strongly painted than Kwakiutl work, is marked by precision of design, skill in execution, and strength of expression. These are the people who were responsible for the familiar black "slate carvings," which are actually made of argillite, a stone found only on Queen Charlotte Islands, in British Columbia.

The work of the neighbouring Northwest Coast peoples, such as the Niska, Kitksan, and Tsimshian, who lived upriver from the Tlingit, is perhaps slightly less well known, due largely to the smaller population and their more remote interior location. It is, however, of equal aesthetic merit, and can stand comparison on any basis with the art of the rest of the peoples of this region.

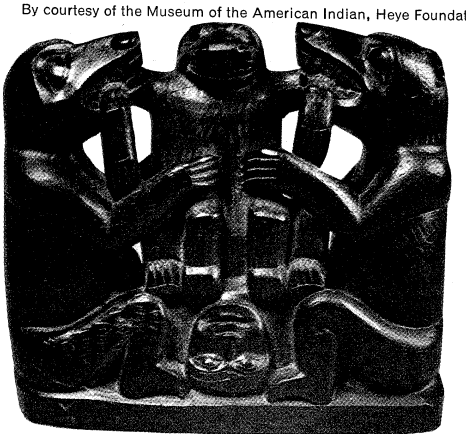
With the coming of the white man, there was a brief period of economic benefit enjoyed by the Northwest Coast people, but this soon disappeared, and the arts rapidly degenerated to curio-shop products. In time, even these provided so little income that all but a few Indian carvers and basket weavers abandoned the arts. Today, again, there are several crafts products that seem destined to disappear, particularly the famed Chilkat blankets. On the other hand, some crafts objects, especially wooden masks that are often carved and painted replicas of older ones, have enjoyed a revival; but, in essence, this is a copying process, largely mechanical and lacking any of the creativity of the original. Argillite carving is experiencing a modest renaissance, but as yet most of the products are very small, ornamental, accessory forms. In general, the exuberance and power of the earlier forms is yet to be fully realized by these gifted, determined, artists of Northwest Coast Indian descent.

By courtesy of the Museum of Primitive Art, New York; photograph, Lisa Little



Tlingit "Chilkat blanket," mountain goat wool and cedar bark, late 19th century, Alaska. In the Museum of Primitive Art, New York City. 167 cm × 134 cm.

By courtesy of the Museum of the American Indian, Heye Foundation, New York



Haida "slate carving" of three bears depicting caesarian birth, argillite *c.* 1890. In the Museum of the American Indian, Heye Foundation, New York City. Height 18 cm.

Mexico and Central America. Although Mexico is geographically part of the North American continent, its cultural world is so different that it is more convenient to consider it as part of Central America. The Indian groups in the region are also somewhat different in that they reflect a strong Spanish influence that was largely absent from most of the North American aboriginal peoples. It is true, however, that a strong intertribal trade existed be-

tween Mexican tribes and those of North America in prehistoric times, and this influence must be borne in mind when considering the arts of both regions.

Olmec art

The earliest identifiable art form of major significance in Mexico is that of the Olmec, the mysterious "rubber people," whose culture was flourishing as early as 1000 BC over an area from Guerrero to Veracruz in Mexico, and into Guatemala, Honduras, and El Salvador. These

George Holton—Photo Researchers



Olmec colossal head, basalt, *c.* 1st century BC. In Parque La Venta, Tabasco Mexico. Height 2.4 m.

Indians carved the tremendous "colossal heads" that the American anthropologist Matthew W. Stirling found at LaVenta; delicate greenstone "baby-face" figures; and figurines with a rounded facial form, thick features, heavy-lidded eyes, and down-turned mouth that are referred to as "were-jaguars" because the image of this mythological or quasi-symbolic, supernatural being is a humanoid type thought by many to combine human aspects with the jaguar concept.

Closely associated with, but somewhat later than, the Olmec culture came the people who inhabited Tlatico, Chupícuaro, and related early sites, which have become well-known for their lovely clay figurines of nude women with fantastic coiffures. At about the time of the slow demise of these civilizations, other peoples had begun to develop their own way of life in western Mexico, notably in Colima, Jalisco, Nayarit, and Michoacán. Far less is known about these areas, for until very recently no professional archaeological work had been undertaken on the scale needed for accurate investigation. The tremendous amount of looting that took place before the major sites came under professional control has destroyed all hopes of unravelling the story of these peoples.

The major architectural construction in ancient America was evolving at this time and reached its apogee about AD 600. The city of Teotihuacán (outside present-day Mexico City), the "Home of the Gods," exercised a tremendous influence from central Mexico into lower Central America; objects inspired by ideas originating at Teotihuacán are still being unearthed throughout the region. The famous masks, so typical of the style, were made *c.* AD 250–750. Their monumental quality formed by the great mass of stone from which the oval eyes, sensual mouth, and broad face are fashioned, provides a powerful sculptural concept. Far too massive to have been worn, it is more likely they were intended as burial offerings, or perhaps facial coverings for wooden effigies.

The Toltecs, Mixtecs, and Zapotecs, widely separated one from the other, have also left their imprint. The former, spreading out from their home area around Tula, eventually travelled as far as the Yucatán Peninsula, leaving evidences of their culture wherever they went. The Zapotec and Mixtec peoples of Puebla and Oaxaca were famed for their unique arts, particularly Mixtec goldwork; these master craftsmen were sought out over great distances for the beautiful jewelry and finely fitted craftworks that are still so highly valued today.

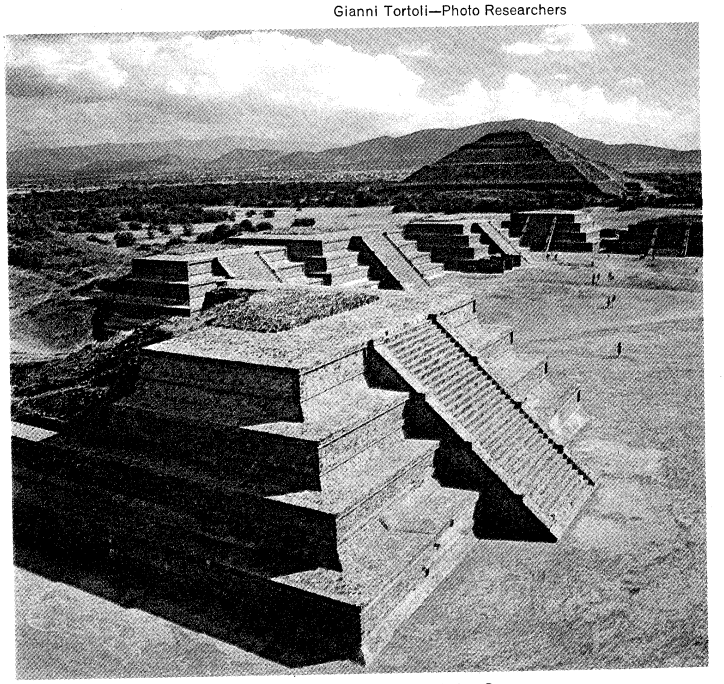
Mixtec goldwork

Along the east coast, in the state of Veracruz, a group of Mayan people called the Huástec had settled by about 250 BC. In time they developed a new cultural expression, which, because they were isolated by Totonac settlers then building up a major centre at El Tajín, remained limited to their own group. Other pre-Totonac folk who were active in Veracruz produced innumerable "smiling-face" figurines and related works that give an impression of an exuberant, happy people. Remarkable among these clayworks are the thousands of small clay whistles that abound in the area. Not only are they valuable as artworks but also as examples of musical instruments popular during that period.

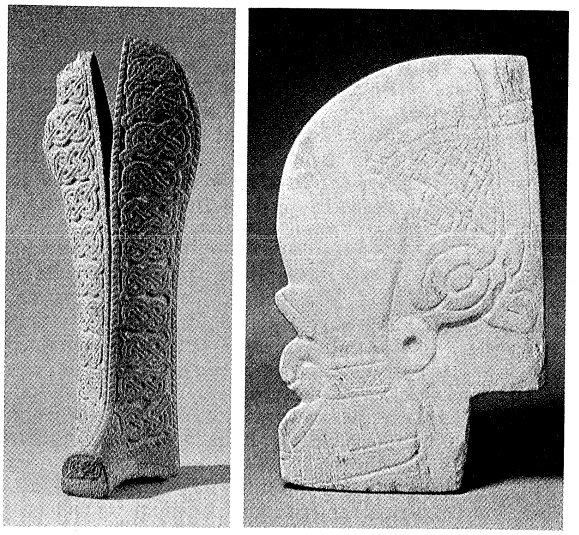
Unique to this region is the use of *chapopote*, a native asphalt commonly applied to clay figurines as a decoration; occasionally, *chapopote* entirely covers the figures, while in other examples, it is used to decorate only the face, mouth, or eyes.

It is also in Veracruz that some puzzling stone carvings have been discovered. Although these objects have been found throughout Central America from central Mexico to El Salvador, their centre seems to have been in the coastal Veracruz area. One of the objects, the *palma*, or palmate stone (shaped like a hand with extended fingers), was first thought to have had some religious significance. The thin-bladed *hacha*, or "axe," so termed because of its resemblance to that implement, was thought to have a ceremonial function of some sort. The *yugo*, or "yoke," was believed to have been used to hold down the head and neck of a sacrificial victim. Today, none of these theories is accepted. Although scholars are still uncertain as to the use of these objects, the best evidence suggests that the *hacha* had an architectural function, perhaps as a wall fitting, or protruding structural decoration. It is possible that some *hachas* were used with the *yugo*; extant sculptures suggest such a use, along with the *palma*, in connection with the ceremonial ball game, *tlachtli*. These sculptures portray an elaborately costumed indi-

Gianni Tortoli—Photo Researchers



Teotihuacán, Valley of Mexico, with the Pyramid of the Sun in the background, *c.* 3rd century BC–8th century AD.



Veracruz stone carvings, about AD 300–900.
(Top left) *palma.* In the Museum of Primitive Art, New York, Height 51 cm. (top right) *hacha.* In the M.H. de Young Memorial Museum, San Francisco. Height 40 cm. (bottom) *yugo.* In the Museum of Primitive Art, New York. Length 41 cm.

By courtesy of (top left, bottom) The Museum of Primitive Art, New York, (top right) The M.H. de Young Memorial Museum, San Francisco; Adriani Bequest to the California Palace of the Legion of Honor; photographs, (top left, bottom) Lisa Little

vidual, wearing what is apparently a *yugo* around his waist. Supported by, and protruding from the latter is a *palma*, indicating that the two may somehow have been fastened together, or were used in combination. It is not believed by scholars that the player actually engaged in the game so encumbered. *Tlachtli* was not unlike today's soccer game; the object was to propel a gutta-percha ball through the air without touching it with the hands; if it went through a small hole in the carved stone disk, or hit that circular goal, the game was won. Tremendous exchanges of personal property resulted from such a victory—indeed, often life itself was forfeit in important contests.

Aztec sculpture

As the Zapotec people of Oacaxa yielded in turn to the more warlike Mixtec, whose centre at Cholula was the site of the largest pyramid of the ancient world (it considerably exceeded the size of the pyramid of Giza in Egypt), so the latter in time became secondary to the Aztec. By 1200, these nomads, who came from the northwest, had established themselves in the central valley, which they called México, whence the name Mexico. The world they built gave rise to a powerful—at times brutal—art form, in which stone was a favourite medium. The rounded, muscular figures that they produced were originally brilliantly painted, much like ancient Greek sculptures. The Aztec turned out an astonishing quantity of these figures, which, standing in rows, served as standard bearers along the avenues leading to various buildings. The other arts, such as pottery and goldwork, seem not to have fared so well during Aztec times. Even the examples that have been found were perhaps less products of Aztec craftsmanship than products of neighbouring peoples that were introduced into the Aztec empire as tribute or as trade objects.

To the east and south, another completely different world appeared under the name of the Maya. Centring in Guatemala and Honduras, where the twin capitals of Quiriguá and Copán are still well-known sites, the Maya spread out to El Salvador, into what is today British Honduras, and into well over one-half of Mexico. The Yucatán Peninsula and neighbouring state of Campeche are areas where a large number of Mayan sites have been found; of these, undoubtedly the most famous Mayan architectural monuments are at Uxmal, Labná, Kabah, and Sayil and the most renowned examples of Mayan fresco painting have been found at Bonampak. Chichén Itzá, the famous archaeological area near Merida, combines both Maya and Toltec influences.

Jaina figurines

Just off the coast of Campeche is the island cemetery of Jaina, from which have come magnificently modelled figurines that are certainly among the finest clay works of antiquity. These sacrificial burial figures, replicas of Mayan personages in ceremonial finery, provide a remarkable insight into the customs, life-styles, and costumes of the classic Mayan people.

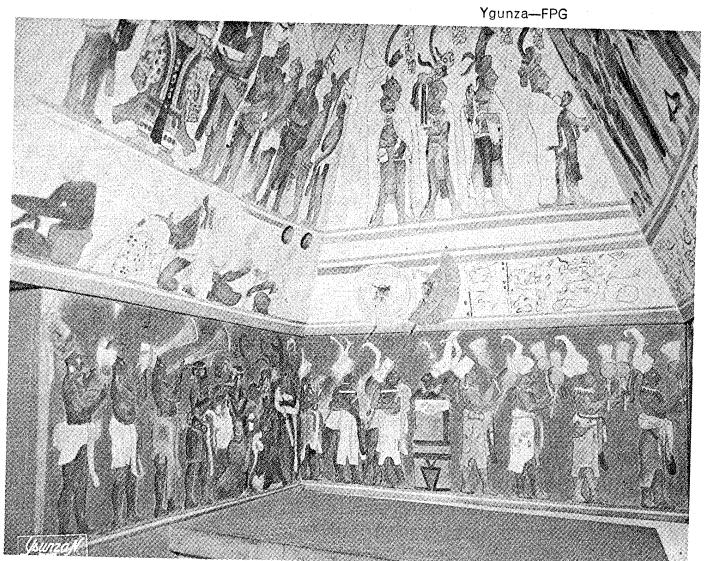
Unlike the brutal force of Aztec sculpture, Mayan art forms are relatively delicate. Yet, although light tracery is characteristic of their sculpture and painting, there are also occasional bursts of sheer power. Perhaps they were the most art-conscious people of the ancient Americas; certainly everything they created seems to have been in terms of aesthetics. They were competent in the use of many raw materials—shell, bone, stone of various kinds, wood, fibres, even feathers became part of their art. Surprisingly, with all of their skills, they seem never to have undertaken much metalwork: gold, silver, and copper objects are exceedingly rare from Mayan sites.

From an early beginning, around AD 250, the Mayan civilization had reached its apogee around 750 and had disappeared by 900. Remnant groups kept the cultural thread as a continuum up to about 1200; but by then, except for their language, they could no longer be regarded as cognate with the earlier Mayan.

Farther south, in Nicaragua, Costa Rica, and Panama, the relationships of the various people are less clear, due largely to the paucity of archaeological work that has been undertaken. Moreover, because Panama was something of a crossroads, both east–west and north–south, the variety of influences found there makes conclusive evaluations of the native art almost impossible. But even though far less is known about these peoples, there is no question about the excellent quality of their various kinds of artwork.

The two greatest artistic achievements of the region seem to have been in jade carving and goldsmithing. From the isthmus area come some of the finest gold cast-

Ygunza—FPG



Mayan fresco from Bonampak, original c. AD 800.
Reconstruction by Antonio Tejda F. In Chiapas, Mexico.



Jadeite ornament, or "axe god" from Costa Rica, *c.* 500–750. In the Dumbarton Oaks Collection, Washington, D.C. 5.1 × 12.5 cm.
By courtesy of Dumbarton Oaks Research Library and Collection, Washington D.C.

Central American jade carving and goldwork

ings known. Although some delicate, finely wrought goldwork is known, most pieces consist of heavy casting, with lost-wax tracery (interlaced patterns of metal cast in a wax mold) in which animal and bird forms predominate. The polished jadework of Costa Rica is famous throughout the Central American area. The beautiful "axe gods," carved from green jade, must have been as sought after in ancient times as they are today.

Pottery was not an unknown art, and the brilliantly painted vessels found here are emphatic proof of this fact. Surprisingly, the art has yet to receive the recognition due it on aesthetic grounds. Some of the designs are remarkably intricate, bold in form, and frequently as sophisticated as anything found in the Western Hemisphere. And here and there are intriguing touches of humour—a quality largely absent from Aztec and Mayan arts.

Little has survived of the architectural expressions from this area. Some large stone sculptures from Penonomé, in Chiriquí province, Panama, suggest that the use of stone in large structures was not unknown; but apparently all of these structures were destroyed, in the years after the Spanish Conquest, by people using the stones for building.

Following the Conquest, the eradication of native culture in Central America was more rigorous than in many areas, and the net result is that south of Guatemala, there is almost no truly indigenous culture to be found. There are some remnants in Costa Rica, but they are few, and their pre-Columbian culture is only marginal. Today, such arts and crafts as are pursued may reflect a continuum of design but they are basically European influenced. The only regions in which considerable prehistoric aesthetic influence survives are Mexico and Guatemala, where the native craftsman has been able to keep his art somewhat alive by recourse to ancient designs and functions.

West Indies. The Caribbean region of the Americas has undoubtedly lost more of its aboriginal character than any other region. The total decimation of the islands shortly after the Conquest and the subsequent populating of the area by Negro slaves made any carry-over of Indian cultural expressions impossible. And so it is today that residents of those islands rarely feel any sense of relationship to the ancestral inhabitants. Certainly it is true that the average non-Indian has no understanding of the wealth of arts that were to be found there in the past.

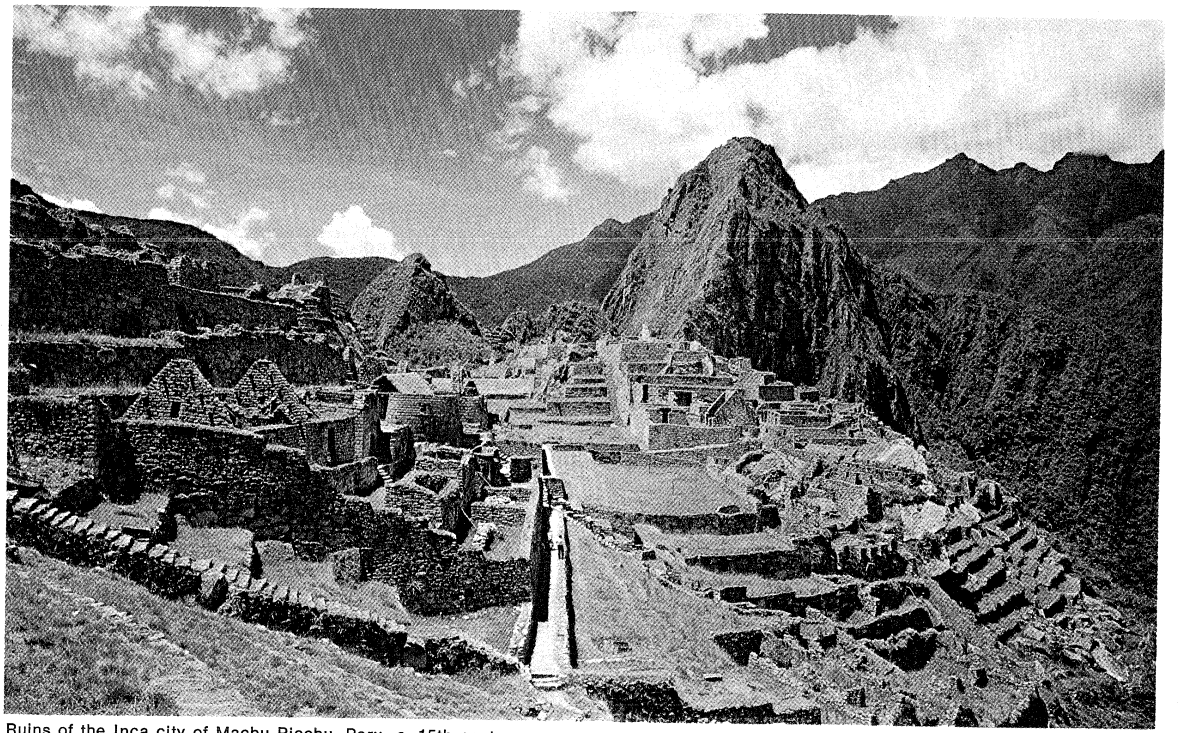
The delicate wood carvings, textiles, featherwork, and related perishable objects that are known from references in Spanish accounts to have existed have largely disappeared. Only a few wood carvings and a small amount of shell and bone carvings have survived. The great strength of pre-Columbian art from the region is in stone; and in this medium, there are remarkably sophisticated, powerful works. Small, tripointed carvings, that were often human or zoomorphic in form, represented the spirits (*zemi*) of the land. The Taino culture is famous for these *zemi* carvings, which are found in many of the islands, notably Puerto Rico and Hispaniola. Carved stone pestles, with human and animal designs, are also common, along with strange "stone collars"—oval carvings that may be related to the *yugos* of Mexico and Guatemala. The most prevalent form, however, is the human head, often a death's-head, which suggests a culture preoccupied with mortality. The peoples of this area were also fascinated by odd shapes in stone. Unusual "comma stones," the meaning of which—if they had any—scholars have been unable to discover, have been found scattered throughout the Antilles. Their number and the care and skill with which they are carved suggests that they had an important role in the culture.

Although the Taino seem to have surpassed the other peoples of the West Indies in aesthetic development, examples of later arts by the Arawak, Carib, and related tribes still surviving in neighbouring South America may provide a link between ancient and modern. Since the Taino were a division of the Arawak, so may modern Arawak weaving indicate something of what must have existed among the prehistoric Taino.

The trans-Caribbean sea route from the islands to the mainland obviously carried influences, as well as materials, back and forth; but far too little is known about these influences to be able to determine which direction was most affected. Little more is known about the West Indies civilization, therefore, than that it produced extremely successful sculpture. The civilization disappeared so rapidly and completely that one can admire but not comprehend it.



Wooden effigy figure, a variant of the *zemi* form, inlaid with shell teeth, *c.* AD 1500. In the Museum of Primitive Art, New York City. Height 69 cm.
By courtesy of the Museum of Primitive Art, New York; photograph, Lisa Little



Ruins of the Inca city of Machu Picchu, Peru, *c.* 15th century.
Shostal

Role of Inca art

South America. The greatest single problem in assessing the Indian art of this region is the unfortunate historical tendency to lump everything together under the heading Inca, as though no other culture had ever attained significance. In fact, when one examines the continent critically, it is evident that the Inca was among the least aesthetically remarkable of the peoples of South America, almost all of whom attained artistic levels only occasionally equalled by the latecomers. It is probably only in the architectural use of stone and in the textile arts that the Inca held his own in artistic comparison.

Weaving was one of the three arts that were South American strengths, the two others being metalwork and pottery. No other peoples in the Western Hemisphere—and less than a handful elsewhere in the world—came close to equalling the Peruvian weaver. One can imagine the astonishment of the early Spanish explorers when they saw this radiant clothing for the first time, even though they very soon passed it over for the gold they coveted.

Metalwork was at its zenith in Peru, Colombia, and Ecuador, each of which developed major cultures whose arts were equal to the demands of the raw material. Tairona gold, in Colombia, rates very high in design and craftsmanship, as does the work of the Quimbaya, whose skill in creating polished metal flasks is remarkable. Notable also is Sinú casting, which could execute works weighing several pounds. In Ecuador, the goldwork found at La Tolita is legendary and shows a skill in casting and overlay that did not seem to exist elsewhere in the region. In Peru, most surviving goldwork was created by the Chimú and Nazca peoples. Yet, that this was a well-advanced art as early as the Chavín era is demonstrated by major discoveries at Chongoyape; indeed, these pieces seem to be the earliest gold products in America, having been created around 900–500 BC.

Perhaps the art that was most widespread and had the greatest variety of form is pottery. In the exciting range of imaginative forms, exuberant vessels are found side by side with sombre, formal works. The use of brilliant colour is common, and the degree of careful modelling makes of many of these pottery containers veritable sculptural masterpieces.

A popular material, stone was used throughout most of South America for massive forms; the small delicately traced stone carvings found in Central America are rarely encountered south of that region. Architecturally, the Inca surpassed all others in his use of intricately cut, giant-size stone blocks, at the sites of Machu Picchu, Sacsahuaman, and Ollantaytambo. Only in San Agustín, in Colombia, was there a similar monumental use of stone. In Manabí, in Ecuador, blocks of stone were skillfully carved into thrones and huge seats. None of this South American stonework, however, reflects the degree of skill common in Mexico.

Antiquity of Colombian art

Colombia. Colombia is among the first South American regions to have enjoyed settlement. A pre-pottery Indian culture is known to have come into the region around 10,000 BC; and Puerto Hormiga excavations reveal that a pottery-making culture existed as early as 3,000 BC. The more definite cultural expressions, however, are not found in quantity until San Agustín, which came into existence with the advent of the Christian Era. Little pottery has been recovered from the region as yet, but stone carvings are very well known. A later culture north of this site, which has yielded a more generous quantity of art objects, is Calima, known for its goldwork, and Quimbaya, whose gold and pottery are both important cultural indicators.

Since the southern and eastern regions of the country are almost unknown archaeologically, conclusive evidence is absent, but at the moment it does not appear that their prehistory was artistically rich. Early pre-pottery sites have been found, notably at El Jobo, in Falcón, that date to about 14,920 BC. Carved stone was used for such objects as small pendants or fetishes; shell and bone are also known to have been used.

It is certain that this was a contact point with many of the Antillean peoples and that travel back and forth between the two regions was a regular custom. Columbus reported such trade—which seems to have been a long-time practice—at the time of his arrival.

Ecuador. It is to Ecuador that one must turn for an examination of early art forms. Straddling the Equator, as the name implies, this region—today the smallest republic in South America—is one of the most intriguing. For decades the region had been ignored by scholars in favour of the more glamorous Peruvian area, but in recent years its tremendous antiquity has begun to be recognized. It now seems that ancient man may have established his first foothold in Ecuador and that the region is also the site of the earliest datable pottery.

From perhaps as early as 15,000 BC until about 3200 BC, when pottery was known to exist at Valdivia, there was a long, steady period of development in the region. And the development was not spotty, for the population increase was also constant.

Although the great cities and some of the major cultural activities found farther south were not found in Ecuador, there was, nevertheless, considerable cultural accomplishment. Weaving was done in quantity, as evidenced by Spanish accounts; and, more spectacularly, goldsmithing was a major expression of the artist's skill. Large pieces, such as crowns and breastplates, and tiny miniatures reflecting the sureness of a master's hand have been found. None of these pieces is unique; they are known in sufficient quantity to prove the existence of a long-time craft. Literally hundreds of thousands of tiny gold beads, each cast individually, have been found in the sands of La Tolita; and others, slightly larger, with granulated surfaces indicating the mastery of a complex casting process, have also been recovered. The technique of inlaying had also been mastered, and the use of emeralds and other gemstones as settings was commonplace. Platinum was worked, as in Colombia; not only was it cast but it also was frequently used in combination with other metals. Copper, too, was worked, both in its "pure" form and combined with tin to make bronze; occasionally, it was gilded to create a pseudo-gold finish. Heavy cast copper axes were stock-in-trade, and many smaller objects were turned out in quantity.

Metalwork in Ecuador

The pottery that emerged from the hands of the clay workers was of high quality, beautifully designed, and well finished. Modelling was powerful, and there were touches of humour. Scholars are not sure to what extent colour was used, for time and soil have removed much of it. Modelled clay effigies discovered in 1966 at Bahía de Manta are not only remarkable for their size and quantity but even more for the astonishing amount of original colour that had been preserved.

It seems evident that this was one of South America's major civilizations and that in time much more will be discovered about it.

Brazil. Like the archaeology of Venezuela, little is actually known of the archaeology of the vast region of Brazil. Only around the Amazon area has very much work been done, and there primarily on the Ilha de Marajó. Size has hampered much of the effort to unravel prehistory, but weather conditions and jungle overgrowth have also combined to resist penetration. What is known, however, is tantalizing to the scholar, for at Lagoa Santa, in Minas Gerais, the bones of a human being have been linked with a mammoth slaughtered for food as early as around 10,000 BC; and pottery vessels have been discovered attesting to a remarkably advanced civilization in the Amazon lowlands perhaps as early as AD 1000. But what lies between these two extremes in time is yet to be discovered.

The most aesthetically exciting object excavated in Brazil is a unique pottery form, found on the Ilha de Marajó and called Marajoara, which incorporates modelling and painting with a low relief carving of the surface. Several successful expeditions have recovered modest amounts of material, but the island, which is regularly flooded by the mouth of the Amazon, has resisted complete excavation. An individual style found on the tiny isle of Maracá, and another from Santarém, suggest the existence in this region of a hodgepodge of aesthetic expressions, some related, some alien. Surprisingly, the strong geometric Marajoara style seems not to have influenced any of the cultures around it.

Marajoara pottery

Peru and highland Bolivia. The great civilizations of Peru and highland Bolivia—with their monolithic stone structures, major political organizations, and elaborate material wealth—have long attracted the attention of the outside world. This was the only area where structures of any real magnitude were built in South America; the ruins of Tiahuanaco, Cuzco, Chan Chan, and similar well-developed urban centres attest to the achievement of highly skilled peoples. Pottery has been found in all styles and types, from relatively crude wares to the most highly painted and polished masterpieces. Whistling vessels are common, and the various forms of musical instruments perhaps exceed those found in other civilizations on the continent. Representations of daily life on the pottery reflect complete, well-rounded civilizations.

Man was definitely active in Peru as early as 10,000 BC, and his pottery making dates to no later than 1200 BC. Slowly, these dates will be pushed back, as scholars unravel more and more of prehistory; for, in view of the advanced stage of some early works, it is certain that they will be found to have been preceded by others.

The great Peruvian ruins around Chavín de Huántar have given the name Chavín to one of the most remarkable civilizations in South America—and one of the earliest, though it is not certain whether it was coeval with the Valdivia of Ecuador (*c.* 3200 BC). Here, remnants, famous throughout the archaeological world, have been found of one of America's earliest cultures. Carved stone objects, fantastic pottery that demonstrates the most advanced skills, evidences of stone construction, and remarkably sophisticated goldwork—all bear witness to a truly magnificent era in ancient history.

Art of the Chavín civilization

By courtesy of the Museum of the American Indian, Heye Foundation, New York



Chavin hammered gold crown, 900–500 BC. In the Museum of the American Indian, Heye Foundation, New York City. 14 x 24 cm.

A recent discovery has brought to light evidences of a civilization at Ayabaca, in Piura, which was probably coeval with the Chavín. Named Vicús after the valley in which it was uncovered and dating between 250 BC–AD 500, this civilization produced pottery that resembles the ware of nearby Ecuador and goldwork not unlike other early forms. The discovery of this wholly new civilization, unknown until just a few years ago, suggests the existence of others.

South of the Chavín region, another high culture developed around the Península Paracas. This civilization produced a famous thin-walled pottery and some of the most extraordinary textiles in existence. Great woven mantles, ponchos, and small tapestries were created between 1000–250 BC.

Paracas textiles

Just as elements from the Chavín civilization filtered south to influence the Paracas people, so they influenced a development in the north, around the Virú, Chicama, and Moche valleys from 250 BC to AD 750. Commonly termed Mochica, these people developed a mature art form that includes some of the finest plastic sculpture in the history of pottery. The range of designs makes these objects remarkable not only as art but also as revelations of the civilization from which they come. The extensive number of objects produced suggests that the civilization was an extremely populous one, in which power and wealth were major goals.



Detail of embroidered Paracas mantle, *c.* 300–100 BC. In the Museo Nacional de Antropologia y Arqueologia, Lima.
By courtesy of the National Archaeological Museum, Lima; photograph, Lee Boltin

Gradually, this civilization gave way to that of invaders, the Chimú, whose capital of Chan Chan was from AD 1000 to 1500 one of the great urban centres of ancient Peru. This huge city, now largely destroyed, once housed 100,000 persons and produced a spectacular array of artistic works: gold jewelry, feather mantles, great textiles, and considerable work in wood and clay. The arid climate has preserved more art from the Chimú region than from many other sections, and Spanish accounts aid scholars in understanding what they find here. Pottery was as skilled as any found elsewhere, although by now something of a paralysis had set in; certainly, many of the designs have a static quality, doubtless due to the extensive use of molds. By now customer demand was so great that the craftsman had to resort to mass production to keep up with his clients.

Chimú and Nazca pottery

In the south, a great pottery talent was at work from *c.* 250 BC to AD 750 around the Nazca Valley. There, perhaps the most technically advanced potters in South America were producing perfectly formed clay vessels, highly fired, brilliantly painted, and often intricately fashioned. Mostly mold-made, they were turned out in great quantities, with the same stiff formality seen in Chimú pottery. The Nazca weavers, however, succeeded in defeating the mass market, for their work was devoted to highest quality, and their skill was such that even though miles of similarly patterned cloth were turned out regularly, the repetition did not destroy its beauty. In fact, the overall patterns so frequently seen provide a harmony that results in a beautiful fabric. There was literally no weaving process unknown to the ancient Peruvian. Nazca goldwork seems not to have been up to the standards achieved by other Peruvian metalwork; by and large, it is a listless product, having only the material to recommend it. Thin, hammered sheet gold was commonly used for Nazca ornamentation.

Closely related to, and extending from, the Nazca work is the work of the Ica civilization (AD 1000–1500). These people produced fine textiles, the designs of which were often reproduced on the pottery of the area. The dry climate has also preserved a wealth of wood carving, much of it in such fine condition that the quality of the art can be clearly seen.

In the central Peruvian area, a group of people emerged, built a modest civilization, and developed it into a world that was in existence when the Spanish arrived. The Chancay people are not known for great artworks; their pottery produced from AD 1000 to 1500 is a simple, black-on-white ware, usually painted in soft colours, simply defined, and frequently crude in appearance. Their one outstanding quality is humour; many Chancay vessels show a lively sense of the absurd, almost providing a Peruvian comic strip. Chancay weaving is excellent, and many thousands of surviving examples attest to this technical pre-eminence.

By courtesy of The American Museum of Natural History, New York



Nazca painted clay double-spouted water jar, 1st–2nd century AD. In the American Museum of Natural History, New York City. Height 21 cm.

By courtesy of the Museum of the American Indian, Heye Foundation, New York



Mochica stirrup-spouted water bottle, *c.* AD 250–500. In the Museum of the American Indian, Heye Foundation, New York City. Height 24 cm.

Farther afield, in Bolivia, another major civilization had been developing: the Tiahuanaco world. Its origin and the whole story of its development are not yet fully understood, but it is known that it came to exert a tremendous influence over a wide area of South America from AD 250 to 750. One of its most characteristic qualities was the use of stone—in walled cities, huge doorways with intricately carved panelling, and great paved roads. Tiahuanaco art is a rather angular expression, with repetitive, rather unoriginal motifs. The pottery from this site is equally uninspiring; although strong in colour, it does not show the variety and technical perfection seen in the wares of the nearby Inca and Nazca. Again, the great art is weaving. It seems that in many cultures the attention devoted to the textile arts far overshadowed that devoted to all of the other arts. And so it is with the Tiahuanaco, who produced ponchos, caps, pouches, and other costume pieces that are instantly recognizable wherever seen and challenge the contemporary weaver

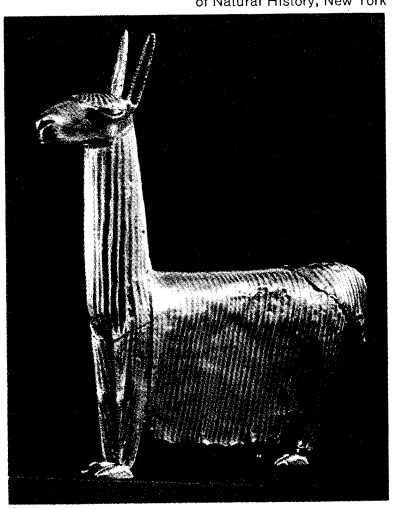
The art of Tiahuanaco

with their variety, fantastically tight weave, and remarkable richness of colour.

The Inca civilization had begun around AD 1200, but the empire itself was not established until 1438, with the accession of Pachacuti Inca Yupanqui, the greatest of the Inca rulers. With the arrival of the Spanish in 1532, the empire was at its height but was suffering from a schism that proved fatal in the face of the European attack. So savage was the Spanish invasion that the empire, numbering some 6,000,000 individuals at its height, was left without a head, and, within a period of thirty years, its population had dwindled to 1,500,000. Part of this legacy is that less art survives from the Inca culture than from many of the far older Peruvian cultures; there are many more Tiahuanaco ponchos known than those from the Inca period, for example. Enough has survived, however, to enable scholars to characterize Inca art forms. The aryballus (globular bottle for liquids) is known the world over, and stonework was common and of excellent quality. Silver and gold were no mystery to the Inca: religious tribute was claimed in the form of worked metal, treated as a gift to the sun god. Indeed, it was this very practice that proved the Incas' undoing, for the Spanish treasure seekers abandoned all other pursuits in their greed for precious metals.

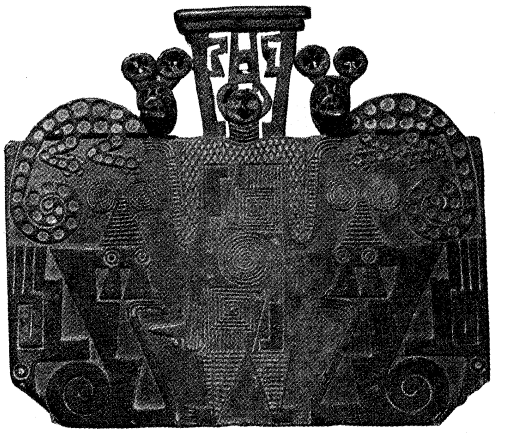
In due course, the Inca civilization declined to a point where it was little more than a shell, as it is today. While much of the ancient weaving is still an active art, there is no pottery of a quality similar to the old, nor is there the wealth of woodwork, stone carving, or other crafts products. As for goldwork, it can safely be said that this art is completely in the hands of Westerners.

By courtesy of The American Museum of Natural History, New York



Inca silver figurine of an alpaca, Inca period (13th–15th century). In the American Museum of Natural History, New York City. Height 19 cm.

Chile and Argentina. Farther to the south, as he continued on his long journey past the Peruvian and Bolivian highlands, man crossed the great Atacama wasteland into Chile and, in so doing, changed his aesthetic expressions. This journey probably took place not much before 500 BC, when pottery-making people became active in the south. However late man came, he quickly set up civilizations. Wandering back and forth over the Andes, he settled both Chile, where he was known as Diaguita, and northern Argentina, where he was known as Chalchaquí. Very soon the peoples of this region developed their own arts, some of which are unique. They produced fine pottery and strong, colourful textiles. Gold was never a major product, although copper became an important metal, due in part to its prevalence. The people cast huge copper disks and plaques and made special burial urns for their children, even reserving cemetery areas in a touching demonstration of affection. The period of Diaguita settlement covered about 1500 years, or from about AD 1 to 1500.



Late Aguada or Early Chalchaquí cast copper plaque from Argentina depicting a man between two felines, *c.* AD 700–1000. In the University Museum of Archaeology and Ethnology, Cambridge. 16 × 14 cm.

By courtesy of the University Museum of Archaeology and Ethnology, Cambridge

With the extension of the Inca Empire into the Chilean and Argentine regions, the ubiquitous aryballus form found its way there, as did other, similar Inca expressions. Indeed, the pottery forms give a clue as to the presence or absence of the Inca overlords.

Western influence on South American Indian cultures

With the arrival of the Europeans all of this changed. Of all of the South American Indian civilizations, only in Ecuador, Peru, Bolivia, and Chile is there anything like a continuum of native arts. And even in these, the Western influence has been so pervasive as to eradicate all but the most dominant aesthetic characteristics. Weaving retains more of its ancestry than perhaps any other art; pottery has taken on much of the Western aura; woodworking is not at all what it was in pre-Columbian times; and very little metalwork is done in what can be called prehistoric design. Although much of the work done today is of highest quality, it reflects more of the tastes and needs of the white man than the Indian.

ARTS OF THE AMERICAN INDIAN PEOPLES IN THE CONTEMPORARY WORLD

In the contemporary art world, American Indian arts occupy a peripheral role. Until the last few decades, the only strong effort to exhibit this art in galleries or museums was made by those few institutions specializing in ethnological, exotic, or art history subjects, together with the rare specialized museum devoted only to Indian materials or to those of the American West. And in these, the usual focus was upon sensational or romantic themes, not upon a carefully balanced understanding of the subject.

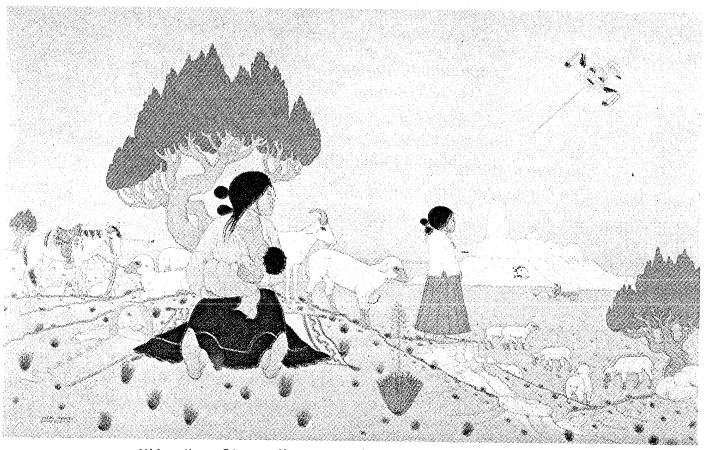
The most active interest in American Indian art seems to have been less in products of still-living cultures than in prehistoric arts; and less in the arts of prehistoric North American peoples than in "pre-Columbian art," which is generally understood to mean the works of the so-called high civilizations—notably the Maya, Aztec, Inca, and Mochica. This is to be regretted, for it not only results in an overemphasis that destroys intellectual balance but it also has relegated to the background some of the more exciting aesthetic accomplishments of the native American. The diorite bowl representing a crested wood duck that has been called by some "the Portland vase of America" is not an isolated instance, for there are other fine sculptures equally deserving of attention.

The reasons for these attitudes are not difficult to discover. With the displacement of the Indian by the Westerner came the rejection of his products, other than as curiosities of a vanquished people. Only in the last few decades has there been a concerted effort to develop an appreciation for art and a concern for the survival of this New World heritage.

More active efforts to preserve Indian art have been made in the United States than elsewhere. The first was in the 1920s, when a group of white artists located in and

around Santa Fe, New Mexico, found excitement in the work of the Indians of the Southwest. Together with the so-called Taos Colony of artists, these influential people succeeded in bringing the values of Indian art to the attention of the outside world through publications, exhibitions, and their own artworks, in which Indian design often figured predominantly. In time, this group saw to the establishment of a School of Indian Art in Santa Fe. Out of this school came many of the most familiar names in Indian art today. Oqwa Pi, Jack Hokeah, Awa Tsireh, Pablita Velarde, Andy Tsinajinnie, Allan Houser, Ben Quintana, Gerónima Cruz Montoya, Eva Mirabal, and Waldo Mootzka are but a very few of the students involved during this exciting period. Following an initial success, the school enjoyed a period of prosperity, but then fell victim to the Depression.

Courtesy of the Denver Art Museum, Colorado



"Herding Sheep," watercolour on paper by Allan Houser, Chiricahua Apache, 1953. In the Denver Art Museum, Colorado. 72 × 46 cm.

Indian Arts and Crafts Board

Another surge of interest came with the enactment of the Indian Reorganization Act of 1934, by means of which the Indian Arts and Crafts Board came into existence. Sparked by John Collier, then commissioner of Indian affairs, this body is one of the few governmental organizations set up specifically to promote, encourage, and revive native arts and crafts. While intended largely as an economic device to increase Indian income, the Board fortunately included members who were knowledgeable about, and sensitive to, the aesthetic and cultural strengths of the Indian. A program of exploration revealed surprising resilience in native crafts, and a core of still-active craftspeople who remembered older techniques was engaged to perpetuate their arts. Out of this program came a renaissance that still continues, even after the Board has become less influential, as the native artist more and more finds himself in his art. What promises to become the major factor influencing Indian art in the remaining decades of the 20th century is the Institute of American Indian Arts in Santa Fe, New Mexico, an outgrowth of the early interest of the Indian Arts and Crafts Board in assisting young Indian artists to secure needed training.

Stimulated by these developments, the interest of art museums and collectors in native art brought home to the general public the existence of a remarkable, if overlooked, art form.

Today, a growing interest in Indian cultural expression is found among North American Indians themselves, as they seek their rightful place in contemporary society. Realizing the values in their heritage, and seeing much of it diminishing, many Indians want to learn what they can of their past and salvage what can be preserved.

Perhaps the greatest positive force to appear in some time are the Indian tribal councils, many of which support the arts in their own areas, not only to augment income but also out of an awareness of the cultural value of those arts. Many of these people, particularly the Navajo, Hopi, Cherokee, and Crow, have set up funds to develop crafts areas, sales centres, and museums to promote the appreciation of their traditional arts and thereby strengthen the fabric of the tribe. Some have set up crafts schools so that the younger people will have access to the necessary training.

Modern painting by North American Indians

With so many years of cultural denial, however, the native artist can no longer draw upon his rich heritage freely but must seek an adjustment with Western-derived motifs that have intruded upon, and in some cases overwhelmed, his world. Among all of the North American tribes, painting is perhaps the one art that has taken a new and positive direction. Many younger Indians of talent are creating remarkable works in acrylic, tempera, oil, and related media, gaining recognition in the fine arts field and showing promise of successful careers in the non-Indian world. Their most serious problem is that of identification. Some prefer to paint in a completely "free" manner, meaning that their work will not necessarily reflect their Indian ancestry. Others remain concerned over the demonstration of that racial strain and seek means whereby they can paint in "Indian style" and yet escape the stereotype so prevalent in the minds of most non-Indians.

In other parts of the Western Hemisphere, organizations similar to those in the United States have recently been established to promote native crafts movements. In general, such movements find their greatest support in areas where nationalism is on the rise and pride in one's background is pre-eminent. The Latin American nation that has taken the most effective steps in this direction has certainly been Mexico; its Museo Nacional de Arqueología (National Museum of Archeology), built in 1964 is perhaps the finest such structure in the world and stands as a monument to the heritage of that country. Rapidly coming up are Guatemala, Colombia, Peru, and Ecuador, each of which has devoted major architectural attention to its native culture. (F.J.D.)

BIBLIOGRAPHY

Literature: (*North America—United States and Canada*): VIRGINIA ARMSTRONG (comp.), *I Have Spoken: American History Through the Voices of the Indians* (1971), a graphic résumé of Indian oratory from the 17th to the 20th century, emphasizing the eloquent speech of the tribes of North America; JOHN BIERHORST (comp.), *In the Trail of the Wind* (1971), a survey of American Indian poetry from many tribes; W.P. CLARK, *The Indian Sign Language, . . .* (1885, reprinted 1959), an early study, but one of the most complete, of this system of visual communication; JESSIE CORNPLANTER, *Legends of the Longhouse* (1938), oral traditions of the Seneca as set down and illustrated by an Iroquois artist-writer; GEORGE W. CRONYN (ed.), *The Path of the Rainbow: An Anthology of Songs and Chants from the Indians of North America*, new ed. (1934; reprinted as *American Indian Poetry*, 1970), a volume of Indian poetry from many tribes; A. GROVE DAY, *The Sky Clears: Poetry of the American Indians* (1951), a compilation of American Indian poetry, including more than 200 poems and lyrics from 40 tribes; WALTER DYK, *The Son of Old Man Hat: A Navaho Autobiography* (1938, reprinted 1967), a classic work recounting a Navajo's life story in his own words; HORATIO HALE (ed.), *The Iroquois Book of Rites* (1883, reprinted 1963), ancient ceremonies as revealed by two aboriginal "manuscripts" discovered at Grand River, with insights into Iroquois social and political life; CHARLES HAYWOOD, *A Bibliography of North American Folklore and Folksong*, 2nd rev. ed., 2 vol. (1961), a remarkable compilation, extremely valuable to anyone interested in American Indian oral literature; ARLENE B. HIRSCHFELDER, *American Indian Authors: A Representative Bibliography* (1970), an extensive listing of oral and written literature authored by persons of Indian descent; MELVILLE JACOBS, *The Content and Style of an Oral Literature: Clackamas Chinook Myths and Tales* (1959), a discussion of a method by which oral literature can be understood in terms of its own content; LOUIS T. JONES, *Aboriginal American Oratory* (1965), a collection of Indian speeches and orations; GARRICK MALLERY, *Picture-Writing of the American Indians* (1893), an exhaustive consideration of the visual forms of communication adopted by many tribes, particularly strong in Plains Indian research; GUY MARY-ROUSSELIERE, *Beyond the High Hills* (1971), a sensitive collection of poems from the Hudson Bay Eskimo people, illustrated by photographs; KAREN PETERSEN, *Howling Wolf: A Cheyenne Warrior's Graphic Interpretation of His People* (1968), a pictorial history of the Cheyenne Indians, based on a collection of drawings by an artist-historian;

PAUL RADIN, *The Road of Life and Death* (1945), a translation of the sacred rituals of the Winnebago, from their origin to their death and reincarnation; LEO W. SIMMONS (ed.), *Sun Chief: The Autiobiography of a Hopi Indian* (1942), a revealing life story depicting the contrast of cultures within the experience of a single individual; STANDS IN TIMBER, JOHN and MARGOT LIBERTY, *Cheyenne Memories* (1967), a systematic collection of tribal history recording early Cheyenne life, by an Indian historian; SARAIN STUMP, *There Is My People Sleeping* (1970), a book of poems by the author, a young Cree-Shoshoni artist; STITH THOMPSON (ed.), *Tales of the North American Indians* (1929, reprinted 1966), one of the earlier and more complete studies of American Indian legend by the dean of folklore research; JUDITH C. ULLOM, *Folklore of the North American Indians* (1969), a collection of titles largely of popular versions but including some excellent analysis of folktales, their distribution, and place in the culture, with special attention given to children's literature; W.C. VANDERWERTH (comp.), *Indian Oratory: Famous Speeches by Noted Indian Chieftains* (1971), a collection of orations by 37 individuals, recorded from 1750 to 1910; JAMES WELCH, *Riding the Earthboy 40* (1971), poems by a Blackfoot youth, reflecting sound knowledge of his people, as related in vibrant terms. (*Mexico, Central America, and South America*): MIGUEL LEON-PORTILLA, *Las literaturas precolombinas de México* (1964; Eng. trans., *Pre-Columbian Literatures of Mexico*, 1969), a selection of myths, hymns, poetry, and prose accounts from Aztec, Maya, Mixtec-Zapotec, and Otomí peoples of Mexico recorded and discussed in depth; ADRIAN RECINOS (trans.), *Popol Vuh: The Sacred Book of the Ancient Quiché Maya* (1950), a complete version of the most important example of Maya literature to survive the conquest; FRANCES GILLMOR, *Flute of the Smoking Mirror: A Portrait of Nezahualcoyotl, Poet-King of the Aztecs* (1949, reprinted 1968), a sensitive portrayal of a major Nahuatl ruler, reflecting the thought processes of the Aztec as expressed through ancient graphic documents; RALPH L. ROYS (trans.), *The Book of Chilam Balam of Chumayel*, new ed. (1967), an account by the prophet of Chumayel village recorded in 1782, rich in Mayan ritual and oral traditions.

Music: PAUL COLLAER (ed.), *Amerika: Eskimo und indianische Bevölkerung*, 2 vol. (1967; Eng. trans., *Music of the Americas: An Illustrated Ethnology of the Eskimos and American Indian Peoples*, 1971), a magnificently illustrated study of the forms and instruments of the aborigines of the New World. (*North America—United States and Canada*): FREDERICK R. BURTON, *American Primitive Music* (1909, reprinted 1969), an important book because of its early exposition of the little known Garden River Reserve area in Ontario and for its transcriptions of Indian hymnody; WALLACE L. CHAFE, *Seneca Thanksgiving Rituals* (1961), a painstaking transcription of texts and melodies by a musically trained linguist; NATALIE CURTIS-BURLIN (ed.), *The Indians' Book* (1907), a pioneering collection of many tribal tunes, written down by ear and lucidly presented; FRANCES DENSMORE, *The American Indians and Their Music* (1926), one of the first major studies of the subject that combined anthropological knowledge with sound musical training to produce a well-organized ethnomusicology; *Chippewa Music*, 2 vol. (1910–13); *Teton Sioux Music* (1918); *Menominee Music* (1932); and *Music of Santo Domingo Pueblo, New Mexico* (1938), particularly important books by Densmore, a pioneer who spent a lifetime salvaging the lore of one tribe after another, transcribing tunes from recordings—early cylinders, then discs, of poor fidelity but of inestimable cultural value; CHARLES HAYWOOD (*op. cit.*), an exhaustive volume, with remarkably detailed consideration of the various areas of American Indian music; GEORGE HERZOG, *Research in Primitive and Folk Music in the United States* (1936), a useful survey by a pioneer musicologist; GERTRUDE P. KURATH, *Iroquois Music and Dance* (1964), transcription and analysis of many Seneca songs, from recordings by William N. Fenton and Martha Huot, with dance diagrams and interpretions from the author's field work; and "Dogrib Choreography and Music," in JUNE HELM MACNEISH and NANCY O. LURIE, "The Dogrib Hand Game," *Bull. Natn. Mus. Can.*, no. 205, pp. 13–28 (1966), transcription, analysis, and choreography; BERNARD S. MASON, *Drums, Tomtoms and Rattles* (1938), a practical book on North American instruments; DAVID P. MCALLESTER, *Peyote Music* (1949), transcription and interpretation of the songs used in this semi-Christian, nativistic cult; and *Enemy Way Music* (1954), one of the few books with musical scores of a Navajo curative rite, with valuable material on native aesthetic concepts; ALAN P. MERRIAM, *Ethnomusicology of the Flathead Indians* (1967), an exhaustive study of the music of a Montana tribe, with musical scores, analyses, and historical notes by an anthropologist; BRUNO NETTL, *Music in Primitive Culture* (1956) and *North American Indian Musical Styles* (1954), representative works by a prolific ethnomusicologist with anthropological as well as musical training, including selected musical scores and bibliographies; HELEN H. ROBERTS, *Musical Areas in Aboriginal North America* (1936), one of the earliest attempts to divide American Indian into its geographical subdivisions; and with DIAMOND JENNESS, *Songs of the Copper Eskimos* (1925); and with MORRIS SWADESH, *Songs of the Nootka Indians of Western Vancouver Island* (1955), two painstaking works with detailed musical scores and background notes by collaborators; FRANK G. SPECK, *Ceremonial Songs of the Creek and Yuchi Indians* (1911) and *Penobscot Man* (1940), two reliable books by an anthropologist, with musical scores but no analyses, by JACOB SAPIR (also contains comments on dance). (*Mexico, Central America, and South America*): GILBERT CHASE, *A Guide to the Music of Latin America*, 2nd ed. rev. (1962), an indispensable bibliography; RAOUL and M. D'HARCOURT, *La Musique des Incas et ses survivances*, 2 vol. (1925), still one of the most exhaustive studies of the ancient Inca and his music, and its present-day place in South America; KARL GUSTAV IZIKOWITZ, *Musical and Other Sound Instruments of the South American Indians* (1935), a remarkable compilation of facts on the forms and functions of instruments; SAMUEL MARTI, *Canto, danza y música precortesianos* (1961), a well-illustrated study of the subject by one of Mexico's foremost ethnomusicologists; and *Instrumentos musicales precortesianos* (1955), a valuable description of native instruments, past and present, with excellent photographs, scales of prehistoric flutes, and notations of native South American and Middle American tunes; CHARLES W. MEAD, *The Musical Instruments of the Inca* (1924), an illustrated examination of various Peruvian examples, largely from the American Museum of Natural History collections; SEGUNDO L. MORENO, *La música de los Incas* (1957), a well-rounded survey, written by an Ecuadorean ethnomusicologist, of the ancient and contemporary Inca musical world; ROBERT STEVENSON, *Music in Aztec and Inca Territory* (1968), a major survey of the most recent attitudes toward ancient music, its instrumentation, and social position in the two areas at the time of the conquest.

Recordings are available in vast numbers. The public has access to the commercial albums of the Library of Congress, Ethnic Folkways, Canyon Records, and to vast archives, such as the National Museum of Canada, the Wenner-Gren Foundation for Anthropological Research, the American Philosophical Society, and Indiana University.

Dance: (*North America—Canada and the United States*): JULIA M. BUTTREE-SETON, *The Rhythm of the Red Man* (1930), choreographies, some music; FREDERICK J. DOCKSTADER, *The Kachina and the White Man: A Study of the Influences of White Culture on the Hopi Kachina Cult* (1954), a survey on the origin of Pueblo masked dances and their development among the Hopi people of Arizona; BESSIE and MAY G. EVANS, *American Indian Dance Steps* (1931), descriptions of steps, six choreographies, and music; WILLIAM N. FENTON and GERTRUDE P. KURATH, *The Iroquois Eagle Dance: An Offshoot of the Calumet Dance* (1935), history, choreographies, music, analysis, photographs, and bibliography; ERNA FERGUSSON, *Dancing Gods* (1931), an evaluation of ceremonial dances of the Indians of the Southwest by a gifted writer long resident in New Mexico; ALICE C. FLETCHER, *Indian Games and Dances with Native Songs* (1915), a combined study of games, songs, and dances, useful to younger readers; CHARLES HAYWOOD (*op. cit.*), bibliography primarily of mythology and music, with a considerable section devoted to works on the dance; GERTRUDE P. KURATH, *Michigan Indian Festivals* (1966), history, choreography, music, photographs, and bibliography; *Iroquois Music and Dance* (1964), diagrams, choreography, music, analysis, and photographs; and *Music and Dance of the Tewa Pueblos* (1970), background, choreography, music, photographs, and bibliography; BERNARD S. MASON, *Dances and Stories of the American Indian* (1944), almost entirely concerned with North American Indian dance steps, forms, and costumes (well illustrated); VIRGINIA M. ROEDIGER, *Ceremonial Costumes of the Pueblo Indians* (1941), a superbly illustrated volume, dealing with all Pueblo tribes and their ritual dress, including dance costumes; CURT SACHS, *Eine Weltgeschichte des Tanzes* (1933; Eng. trans., *World History of the Dance*, 1937, reprinted 1963), includes several sections on various Indian tribal dance performances; FRANK G. SPECK and LEONARD BROOM, *Cherokee Dance and Drama* (1951), a specialized study of Eastern Cherokee dances and related ritual; JOHN L. SQUIRES and ROBERT E. MCLEAN, *American Indian Dances* (1963), a volume intended primarily for hobbyist readers. (*Mexico, Central America, and South America*): AUGUSTE GENIN, *Notes on the Dances, Music, and Songs of the Ancient and Modern Mexicans* (1922), a comparison of pre-Columbian with contemporary folk dances; LISA LEKIS, *Folk Dances of Latin America* (1958), exhaustive bibliography, with his-

torical notes and descriptions; SAMUEL MARTI, *Canto, danza y música precortesianos* (1961), thoroughly illustrated, with equal sections given to an examination of musical instruments, songs, and the forms of the dance in pre-Columbian times; and with GERTRUDE P. KURATH, *Dances of Anáhuac: The Choreography and Music of Precortesian Dances* (1964), a thorough and well-illustrated analysis of the subject; ROBERT STEVENSON, *Music in Aztec and Inca Territory* (1968), includes a large amount of valuable material on the dance in pre-Columbian times, although primarily involved with music, singing, and the vocal arts; FRANCES TOOR, *A Treasury of Mexican Folkways* (1947), a volume on folklore, including the dance, by a long-time resident of Mexico; CARLOS VEGA, *Las Danzas Populares Argentinas* (1952), authoritative history, choreography, and some music.

Visual arts: FERDINAND ANTON and FREDERICK J. DOCKSTADER, *Das Alte Amerika* (1967; Eng. trans., *Pre-Columbian Art and Later Tribal Arts*, 1968), a popular treatment that covers the Western Hemisphere, extremely valuable for the many illustrations, most of which are in colour; FRANZ BOAS, *Primitive Art* (1927; new ed., 1955), an old-time classic with considerable discussion of the place of native arts in European society—especially helpful for the section on Northwest Coast art; G.H.S. BUSHNELL, *Ancient Arts of the Americas* (1965), a general treatment of the Western Hemisphere, with greater attention given to Latin America; MIGUEL COVARRUBIAS, *The Eagle, the Jaguar and the Serpent* (1954), a splendid volume on the arts of the Americas, with many line drawings and colour plates by the author, a major Mexican artist; *Indian Art of Mexico and Central America* (1957), a continuation of the above and one of the best single treatments on the formative centuries of ancient Mexican civilization; HANS D. DISSELHOFF and S. LINNE, *Alt-Amerika* (1960; Eng. trans., *The Art of Ancient America*, 1961), a well-illustrated volume that concentrates primarily upon the prehistoric cultures of Central and South America, and largely ignores North American Indian art; FREDERICK J. DOCKSTADER, *Indian Art in America: The Arts and Crafts of the North American Indian*, 3rd ed. (1966), a survey of the native arts of the United States and Canada; *Indian Art in Middle America: The Arts and Crafts of the Indians of Mexico, Central America and the Caribbean* (1964), a continuation of the above, extending the study to the southern boundary of Panama, with particular emphasis on the lesser known tribes of the region; *Indian Art in South America: Pre-Columbian and Contemporary Arts and Crafts* (1967), the final volume in this series, completing a survey of the entire range of aboriginal art of the New World; ALVIN M. JOSEPHY, JR., *The Indian Heritage of America* (1968), one of the few volumes to discuss the Indian form North to South—a sound introduction to the New World native, although written from the historical point of view; PAL KELEMEN, *Art of the Americas, Ancient and Hispanic* (1969), a well-illustrated study of the earlier periods of the Americas, particularly valuable for its inclusion of the colonial influences on Indian aesthetics; and *Medieval American Art*, 3rd rev. ed., 2 vol. (1969), a rich sourcebook of illustrations that includes the entire hemisphere; GORDON R. WILLEY, *An Introduction to American Archaeology*, 2 vol. (1966–71), a profusely illustrated, comprehensive survey of the prehistory of the New World, by one of its leading archaeologists. (*North America—United States and Canada*): CHARLES A. AMSDEN, *Navaho Weaving: Its Technic and History* (1934, reprinted 1964); MARIUS BARBEAU, *Totem Poles . . .*, 2 vol. (1929–30); ROBERT T. DAVIS, *Native Arts of the Pacific Northwest* (1949); FREDERIC H. DOUGLAS and RENE D'HARNONCOURT, *Indian Art of the United States* (1941, reprinted 1969); DOROTHY DUNN, *American Indian Painting of the Southwest and Plains Areas* (1968); WOLFGANG HABERLAND, *The Art of North America* (1964); ROBERT B. INVERARITY, *Art of the Northwest Coast Indians* (1950); WILLIAM C. ORCHARD, *Beads and Beadwork of the North American Indians*, 2nd rev. ed. (1972), the most complete single study of Western Hemisphere bead artistry; CLARA LEE TANNER, *Southwest Indian Craft Arts* (1968). (*Mexico, Central America, and the West Indies*): MIGUEL COVARRUBIAS, *Indian Art of Mexico and Central America* 1957); ANDRE EMMERICH, *Art Before Columbus* (1963); JESSE W. FEWKES, *A Prehistoric Island Culture Area of America* 1922); PAL KELEMEN, *Medieval American Art*, 3rd rev. ed., 2 vol. (1969); GEORGE KUBLER, *The Art and Architecture of Ancient America: The Mexican, Maya, and Andean Peoples* (1962); SAMUEL K. LOTHROP, *Pottery of Costa Rica and Nicaragua*, 2 vol. (1926); IGNACIO MARQUINA, *Arquitectura Prehispánica* (1951); SYLVANUS G. MORLEY, *The Ancient Maya*, 3rd ed. (1956); TATIANA PROSKOURIAKOFF, *An Album of Maya Architecture* (1946); DONALD ROBERTSON, *Mexican Manuscript Painting of the Early Colonial Period* (1959); MARSHALL H. SAVILLE, *The Goldsmith's Art in Ancient Mexico* (1920); HERBERT J. SPINDEN, *A Study of Maya Art* (1913; rev. as *Maya Art and Civilization*, 1957); J. ERIC THOMPSON, *The Rise and Fall of Maya Civilization* (1954); GEORGE C. VAILLANT, *The Aztecs of Mexico*, rev. ed. (1962); HASSO VON WINNING, *Pre-Columbian Art of Mexico and Central America* (1968); GORDON WILLEY, *An Introduction to American Archaeology*, 2 vol. (1966–71). (*South America*): WENDELL C. BENNETT, *Ancient Arts of the Andes* (1954); JUNIUS B. BIRD and LOUISA BELLINGER, *Paracas Fabrics and Nazca Needlework, 3rd Century B.C.–3rd Century A.D.* (1954); G.H.S. BUSHNELL, *Peru*, rev. ed. (1963); RAOUL D'HARCOURT, *Les Textiles anciens du Pérou, et leurs techniques* (1934; Eng. trans., *Textiles of Ancient Peru and Their Techniques*, 1934); ANDRE EMMERICH, *Sweat of the Sun and Tears of the Moon: Gold and Silver in Pre-Columbian Art* (1965); PAL KELEMEN, *Art of the Americas, Ancient and Hispanic* (1969); SAMUEL K. LOTHROP, *Treasures of Ancient America* (1964); J. ALDEN MASON, *Ancient Civilizations of Peru* (1957); BETTY J. MEGGERS, *Ecuador* (1966); G. REICHEL-DOLMATOFF, *Colombia* (1965); ALAN R. SAWYER, *Ancient Peruvian Ceramics* (1966).

(F.J.D./E.Gu./G.P.K.)

American Sub-Arctic Cultures

Resources and populations

The sub-Arctic, or the physiographic zone called the taiga, is a land of coniferous forest covered with sphagnum moss and traversed by many waterways. Its main natural resources utilized by the Indians include wood, of which spruce, tamarack, and alder have been especially important for manufactures; game animals, such as moose, caribou, and beaver, whose meat is used for food and whose skins are used for blankets and clothing; berries; fish, usually the most abundant and therefore the staple food; and coloured earths converted into paints. Cold winters, when the ground is covered with snow and the waterways are frozen, are reflected in many technological adaptations, including the use of fur garments, toboggans, ice chisels, and snowshoes.

Since the 17th century, traders have been attracted to the region by its wealth in fur-bearing animals. Together with missionaries, they decisively influenced the native Indians' history. Many old beliefs, hunting customs, kinship relations, and so on, however, persisted throughout the fur-trade years in this marginal area and are only now beginning to vanish.

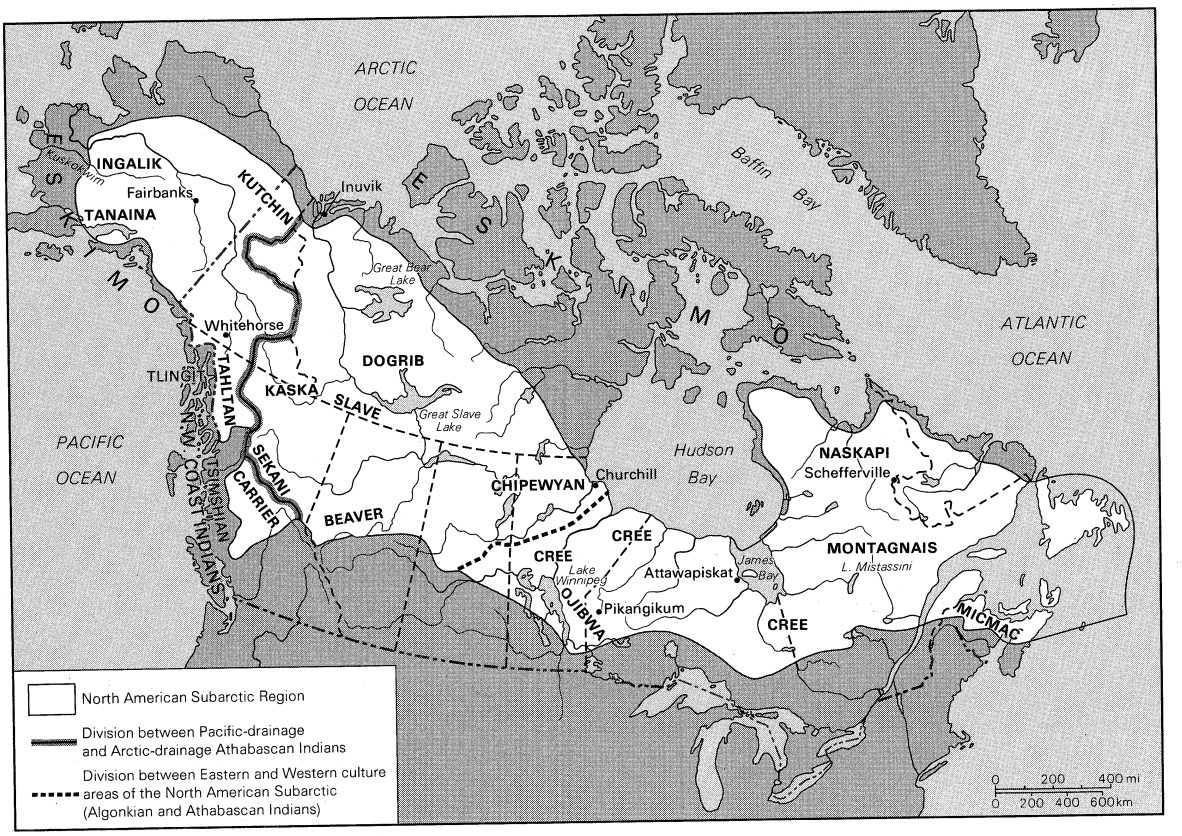
The sub-Arctic contains two relatively distinct culture areas. The eastern sub-Arctic is inhabited by Algonkian-speaking Indians, including the Naskapi of northern Quebec; the far-flung Cree; and an intrusive wedge of Ojibwa, who, after the beginning of the fur trade, displaced the Cree from west central Ontario and eastern Manitoba. The western sub-Arctic belongs largely to Athabascan speakers, who extend from Canada into Alaska. Cultural differences among the Athabascans justify the delineation of two subareas. The first subarea, drained mostly by the northward-flowing Mackenzie River system, is inhabited by Chipewyan, Beaver, Slave, and Kaska Indians. Their culture was generally simpler and less sedentary than that of the second subarea, whose salmon streams drain into the Pacific Ocean. Its groups include the Carrier, part of the Kutchin, the Tanaina, and Ingalik Indians.

Northward the Algonkians and Athabascans border on the Eskimo; to the west the Canadian Athabascans encounter the Tlingit, Tsimshian, and other Northwest Coast tribes, while the Alaskan groups again abut on the Eskimo.

Native population densities, low everywhere, are higher toward the Pacific than in the east. Variation also occurs within culture areas. In the eastern sub-Arctic, the Mistassini Cree in central Quebec have 66 square miles (171 square kilometres) per person at their disposal, whereas in western Ontario, the Penkangikum Ojibwa have less than 12 square miles (31 square kilometres) per person.

TRADITIONAL CULTURE PATTERNS

Ethos. In most of the region east of Alaska, the uncomplex nature of subsistence and manufacturing technology is reflected in the Indians' elemental social structure, based on the family and the small local band, and in their individualistic relationship to the supernatural. Both features of the culture pattern are linked in turn



Distribution of American sub-Arctic cultures.

with characteristic personality features, such as the high value placed on personal autonomy; the generalized fear of the world as dangerous; the emphasis on concrete, current realities rather than on future possibilities; and self-reliance, manifested, for example, in bearing up stoically to deprivation. Psychological homogeneity in the sub-Arctic extends to other personal traits such as reticence, nondemonstrative emotionality, considerable deference to other people, and strong individual control of aggressive impulses. Hostility is not absent, but it is suppressed and only indirectly revealed through such outlets as sorcery, fear of sorcery, or malicious gossip.

Territorial organization. At the time when white men contacted the Indians, social structure tended to be extremely simple. The Indians had no stable, territorial political systems, no special offices charged with the task of coordinating groups, no formal social stratification, and no occupational specialization except those related to age and sex. Temporary leadership emerged when one group raided another or fought with the Eskimo or when there were other tasks requiring cooperation beyond the family.

Bands and villages

The Indians were territorially grouped in unstable regional bands without chiefs. These units took their identity from some geographical feature of the land that members exploited—hence such appellations as "mountain people," or "Marten Lake people." A smaller kind of group east of the Pacific drainage culture area was the local band. It usually consisted of a few related families who travelled together for a time. Several local bands coalesced seasonally at a good fishing lake or near rich hunting grounds for periods that were as intensely sociable as they were abundantly provided with fish or game. In the eastern sub-Arctic, where local bands have been called hunting groups, they averaged about 20 individuals.

The fur-trade period created a new type of territorial group: the trading-post band, named from the settlement in which its members traded. The new bands amalgamated local groups and notably expanded the population in which marriage occurred. In some places the Canadian government concluded treaties with such groups and arranged for the election of chiefs and councillors.

In the Pacific drainage area, sedentary villages took the place of local bands. On the lower Yukon and upper Koskokwim rivers, Ingalik village life centred on the *kashim*, or men's house, where a council of old men met to hear trouble cases and where elaborate seasonal ceremonies were performed.

Authority in bands and villages derived primarily from the combination of a man's eloquence, wisdom based on experience, healing or magical power, generosity, and capacity for hard work. Aided by the latter quality, he would normally be well supplied with food and thereby able to help others.

Family and kinship relations. Within the local band, the simple family of husband, wife, and children, frequently including adopted children, constituted the nucleus of economic activity and emotional security. The intense importance of the family, especially in a young person's life, is revealed in stories about the unhappy lot of cruelly treated orphans and is further suggested by the reported eagerness with which widowers sought to remarry.

Marriage, friendship, and avoidance

The prevailing modes of obtaining a wife required bride service rendered to the wife's father or transfer of a sister of the groom in exchange for a bride. Although households were primarily monogamous, a man occasionally had two wives. This could happen, for example, when by custom a man espoused his dead brother's widow. Cross-cousin marriage (that is, a man's marriage with his mother's brother's daughter or his father's sister's daughter) is known to have been practiced, perhaps by preference, among some groups. Elsewhere it can be inferred from several features in kinship terminology. For example, a man uses the same term for his female cross-cousins and for his wife's sister. The custom declined when it was opposed by missionaries, who condemned it as incest, and when marriageable populations expanded around trading posts and missions.

Kin relations among the Kaska included teasing between grandparents and grandchildren, initiated by the youngsters. Siblings of the opposite sex who had reached puberty conducted themselves circumspectly in each other's presence and even tended to practice polite avoidance, as a father and a grown daughter also did. Parallel

cousins (mother's sisters' children and father's brothers' children) were treated as siblings, both being called by the same kinship term. Ceremonial avoidance also governed the relationship of a man and his mother-in-law, contrasting with the camaraderie linking brothers-in-law, which was one of the warmest of all relationships between grown men. A Kaska man could joke freely with his wife's sister and brother's wife, sexual ribaldry not being barred from their conversation.

Socialization of children. Pregnancy taboos and postnatal observances were intended to serve the mother's own good and to insure the child's well-being. Birth took place in the dwelling, in a special birth hut, or, according to early travelers among the eastern seaboard Micmac, in the woods. One or more knowledgeable women assisted the mother in giving birth and in caring for the delivered child. Swaddled babies were diapered with moss and carried on the mother's back in an ornamented skin bag or a cradleboard.

Education and initiation

Family members and other relatives played the major role in the informal process of childhood education. A child had considerable scope to learn through copying the unself-concious example of models. Thus a Kaska Indian parent might say "Make tea!", and a small girl would try to reconstruct what she had often observed her mother and older sisters doing but what she had never been formally instructed to do. Parents did not neglect disciplining and even chastising a disobedient child for such offenses as stealing and rebelliousness. More important for the formation of personalities suited to a life of food gathering is the fact that parental treatment subtly but firmly encouraged children to become independent and self-reliant.

Several "firsts," including the first tooth, the first game killed by a boy, and the first menstruation (menarche) were ceremonially recognized, sometimes by a small feast or, in the case of menarche, by an elaborate series of taboos. Athabascan Indians paid the greatest ritual attention to menarche, the Kutchin girl going to live for up to a year in a special shelter constructed some distance from the family camp. Here she wore a pointed hood that forced her to look down toward the ground. Other precautions for her own good included a rattle of bone that was supposed to prevent her from hearing anything; a special stick to use if she wanted to scratch her head, and a special cup that should not touch her lips. Subsequent menstruation required only a short period of seclusion.

When a boy approached adolescence, he went alone into the forest to seek a vision. In Kaska Indian terms the vision came from "dreaming of animals in a lonely place" or from hearing "somebody sing," perhaps a moose in the guise of a person. Success in the vision quest supernaturally enhanced the youth's ability, bestowing on him the power of the animal helpers. Dreams also notified an individual of impending events and might advise him on when and where to hunt or how to behave on other occasions in order to achieve success or avoid misfortune.

Social stratification. The maximum of social stratification was reached among the Ingalik, who informally recognized three classes of families. Three-quarters of the village comprised common people. Rich families, so called because they accumulated surplus food thanks to members' industry or superior hunting and fishing abilities, constituted about 5 percent of the community. They took the lead in the community's ceremonial life. The rest of the people did "nothing" and lived off the others; consequently, they enjoyed so little respect that they had a hard time finding spouses.

Seasonal migration

Settlement and housing. In pursuit of a livelihood, families and local bands shifted their location as the seasons changed. In northwest Canada groups scattered in early winter to hunt caribou in the mountains; elsewhere, autumn drew people to shorelines of lakes and bays where large numbers of ducks and geese could be taken for the winter larder. At other times people gathered around fish lakes. Ingalik Indians in late winter quit the village and headed for spring camps, as much for a change as for the good fishing. As dependence on fur trapping became heavier, the Cree, Slave, Kaska, and many other Indians adopted a two-part annual cycle. In winter the family lived on its trapline. Then, with summer, it brought its furs to the trading post where people camped until fall, enjoying a rest from work and abundant sociability. The warm months with their long daylight became a time for dances (the steps and fiddle music being copied from early European types), for marriages, and for visits of the Anglican or Roman Catholic bishop.

Despite much movement, shelters were not always portable. The Ingalik spent winters in houses excavated in the soil, roofed with beams and poles, hung with mats, and provided with an entry. Other groups, such as the Cree and Ojibwa, built conical winter lodges durably roofed with boughs, earth, and snow. On the trail, however, people put up skin or brush shelters, sometimes only a simple lean-to, or camped in the open facing a fire.

Production and technology. Everywhere in the sub-Arctic, fish and game, augmented by berries and a slight quantity of other vegetable products, provided the food supply. Meat was boiled with the aid of hot stones in containers of bark or woven spruce roots. A large, varied set of weapons, traps, and other ingenious appliances played a vital role in providing subsistence. Important devices included the bow and arrow, with stone or bone tips for different kinds of game; lances; the spear-thrower and spear; weirs and basket traps for fish; nets of willow bark and of other substances; snares for such small game as rabbits; deadfalls (traps with logs or other weights that fall on game and kill them); pit traps; and decoys for birds. No less vital for people who depended heavily on mobility for survival were vehicles, such as bark canoes or snow toboggans of hardwood, and travel aids, such as large sinew-netted snowshoes to run down large game, a smaller variety to break trail for the toboggan, and snow goggles to use against the glare of the spring sun. Dogs large or numerous enough to pull toboggans were unavailable until the fur-trade period.

Techniques and skills

In addition to the Indians' technical resourcefulness and skill, magic and divination aided hunting. A noteworthy form of divination used in locating game required heating a large animal's shoulder blade over fire until it cracked. Hunters then went in the direction of the crack. The random element in the method increased the chances that they would go to a fresh, relatively undisturbed piece of ground.

An idea of the extent to which people depended on game and of the labour involved in food getting can be gained from food-consumption figures obtained in the mid-20th century. In the relatively poor country west of James Bay, 400 Cree men, women, and children in the course of a fall, winter, and spring (nine months) consumed about 128,000 pounds (58,000 kilograms) of meat and fish in addition to staples from the store, especially flour, lard, and sugar.

Across the sub-Arctic people preserved meat by drying and pounding it together with fat and berries to make pemmican. The Pacific-drainage Athabascans also preserved salmon by smoking. Other widely distributed, basic technical skills included complicated chemical processes, using animal brains or human urine, to tan caribou and moose skins; these were then sewn into garments with the help of bone needles and animal sinew. Women also plaited rabbit skins into ropes and wove roots to form watertight baskets.

Property. Land and water, the sources of food, were not considered to be either individual or group property, yet nobody would usurp the privilege of a group that was currently exploiting a berry patch, beaver creek, or hunting range. Clothing, the contents of food caches, and other portable goods had individual owners. Legally inalienable family trapping territories came into being with the fur trade and in many places have been registered by the government. Sharing game has always been important economically, while gifts other than food were bestowed primarily ceremoniously.

Religious beliefs. Two important concepts of the Naskapi and other Algonkian groups were Manitou and the "big man." Manitou represents a pervasive power in the world that individuals can learn to use in their own behalf. The term Great Manitou, designating a personal god, probably represents a missionary-inspired adaptation of the Indians' older idea. Among the Naskapi and Rupert House Cree, a person's big man is a spirit intimately associated with him. It confers wisdom, competence, skill, and strength in the food quest as well as in other areas of life, including magic. Maintaining this spirit requires ethically good conduct. Spirit "bosses" who control the supply of caribou, fish, and other creatures are another traditional belief shared by Algonkian and certain Athabascan groups.

Among many sub-Arctic Indians, success in the vision quest augmented personal competence by allying the individual with animal helpers. There was also a widespread belief that if one wanted to continue to hunt game successfully, the dead animals and their remains must be treated with reverence. This required disposing of their bones carefully so that dogs could not chew them. Respect was particularly evident in the use of polite circumlocutions to refer to bears. Quebec Indians undertook several ceremonial observances in bear hunting, including a purifying sweat bath before departing on the hunt and an offer of tobacco to a bear that had been killed. Afterward the people feasted and danced in its honour.

Folklore, magic, and spirit relations

Three of the most popular characters in Algonkian folklore are Wiitiko (Windigo), a terrifying cannibalistic giant apt to be encountered in the forest; Tcikapis, a kindly, powerful young hero and the subject of many myths; and Wiskijan, an amusing, prankful trickster, Wiitiko psychosis refers to a condition in which an Indian would be seized by the obsessive idea that he was turning into a cannibal with a compulsive craving for human flesh.

The shaman, the sole religious functionary, served sub-Arctic people as a part-time specialist curer and diviner. Occasionally he also behaved malignantly. Shamanistic ability came to an individual from dreaming of animals who taught the dreamer to work with their aid; such ability had to be validated through successful performance.

The Ingalik conceived of man as comprising body, soul, and "speech," the latter an element surviving after death but not, like the soul, reincarnated. Hazards to life came from the soul always being menanced by various supernatural figures that were the primary enemies of human survival, and by the souls of powerful evil shamans acting on behalf of these supernatural figures. The animal and berry people supported life. Animal "songs" and amulets created good relations with helpful animal people; and elaborate ceremonies in the men's house, to which animals were invited, protected the food supply.

EVOLUTION OF THE CULTURE TODAY

One view of culture change in the sub-Arctic is that, during the fur-trade and missionary period, the Indians lost vital cultural traits faster than satisfactory replacements were absorbed. An opposite view holds that the fur-trapping era enriched Indian culture. Life came to be organized around the trading post; an emotionally satisfying belief system came to be centred in the Christian Church; and trapping offered new opportunities for prestige. Formal education and modern health care began during fur-trade times, the churches choosing a few localities to institute boarding schools and hospitals that in time received government subsidies.

Since World War II the fur-trade era has been succeeded by a new period characterized by planned social change, incipient urbanization, and intensified modernization. The growing unprofitability of trapping, chronic illness (especially tuberculosis), and high infant-mortality rates aroused the conscience of Canada and introduced the welfare state into the Canadian sub-Arctic. Local schools and health stations replaced distant boarding schools and hospitals. Other government-sponsored programs brought about modern housing, large-scale fur farming, and vocational training. The fishing industry in Alaska, military and other construction, and iron mining in Quebec gave Indians the opportunity to work for wages in place of trapping. A slow trend of migration began to draw native families to established northern cities, such as Fairbanks (Alaska), Whitehorse (Yukon Territory), and Churchill (Manitoba), as well as to new towns, such as Schefferville (Quebec) and Inuvik (Northwest Territories). At the same time expanding jobs in health care, education, mining, defense, and administration stimulated an influx of white families from southern Canada to the north who settled in the northern towns.

Town life for some Indian migrants means a squatter's cabin, welfare, and occasional odd jobs; but for many native men and women it means steady work at open-pit mining, as janitors and hospital aides, in construction and domestic service, or as equipment operators. Indian townsmen restrict hunting and fishing, of which they are still very fond, to holidays and weekends. If the coming of the fur trade and missions widened the intellectual horizons beyond what they were in aboriginal times, then town life with its grade school, high school, library, movies, television, church groups, Boy Scouts, Girl Guides, and other associations increases awareness much more and provides many added chances to enact new, modern roles. Sub-Arctic Indians of all ages are adapting successfully to urban economic and social opportunities, but the ease with which people take to jobs and adjust to new conditions varies in the same town, and some families are experiencing difficulties. Indians are not modernizing as rapidly as Eskimos, who live under similar conditions and sometimes in the same town. Reasons remain to be pinpointed.

Employment and education

The proportions of persons in jobs, in school, and getting into trouble with the law indicate the degree of success or difficulty with which modernization occurs. In 1968 Schefferville, Quebec, for example, was inhabited by 715 Montagnais and Naskapi Indians. Of the males age 15 and over, 36 percent were steadily or fairly steadily employed; this is about the same proportion as Indians and Eskimos in four other northern towns. Of the Schefferville children eligible to be in school, 96 percent were in school, but a quarter of the youth 15 years or over had dropped out. The dropout rate is low compared with Indians and Eskimos in the four other towns. The monthly incidence of arrest for persons 16 years and over was 2 percent, about the same as in three other northern towns. Most arrests are for infractions of the liquor laws. Counting all offenses, young people, especially men, get into trouble proportionately more often than older people do.

Many Indians remain attached to small, isolated trading-post settlements and to a way of life that revolves around the forest, at least for the men. Even among the non-townspeople, one perceives a heightened awareness of native identity, together with growing objectivity about their "traditional" Indian cultural heritage. Their awareness is believed to stem from the knowledge that a different identity and way of life are possible. For the time being, however, they prefer to be cautious about changing the life that they know for one fraught with new challenges.

BIBLIOGRAPHY. Major ethnographic accounts of the pre-fur-trade period include, for the eastern or Algonkian area, FRANK G. SPECK, *Naskapi: The Savage Hunters of the Labrador Peninsula* (1935), an account of beliefs, many of which survive into the present; and JOHN J. HONIGMANN, "The Attawapiskat Swampy Cree: An Ethnographic Reconstruction," *Anthrop. Pap. Univ. Alaska*, 5:23–82 (1956). EDWARD S. ROGERS has written fully on pre- and post-contact culture of the Round Lake (Ontario) Ojibwa; see especially *The Round Lake Ojibwa* (1962). For the western or Athabascan area, JOHN J. HONIGMANN has done an ethnographic reconstruction of *The Kaska Indians* (1954); and CORNELIUS OSGOOD, the leading authority on the Athabascans, reports exhaustively on *Ingalik Material Culture*, *Ingalik Social Culture*, and *Ingalik Mental Culture* (1940, 1958, and 1959). *Man in Northeastern North America*, ed. by FREDERICK JOHNSON (1946), brings together authoritative papers on geography, physical anthropology, linguistics, mythology, psychological characteristics, and culture in general. A study of a northern

town, Inuvik, where Indians live is JOHN J. and IRMA HONIGMANN, *Arctic Townsmen* (1970).

(J.J.Ho.)

Ammunition

In modern military usage "ammunition" is a generic term that includes a great variety of missiles and other devices designed and constructed to inflict damage upon enemy personnel or materiel by action of an explosive, pyrotechnic, or chemical agent. The term includes, but is not limited to, such items as projectiles, propellants, cartridges, bombs, rockets, torpedoes, grenades, explosive mines, and guided missiles. The term ammunition is commonly used in a more restrictive sense, however, to designate only that which is intended for firing in guns. Ammunition for guns is divided broadly into small-arms ammunition—pistols, rifles, shotguns, and machine guns—and artillery ammunition, intended for firing in cannon.

Ammunition size is usually expressed in terms of calibre, which is approximately the diameter of the projectile, given either in inches or in millimetres. A complete round of ammunition includes all of the components necessary for one firing of the gun. The components (see Figure 1) normally include a projectile, a charge of propellant, and a primer that ignites the propellant. Other components such as a cartridge case, fuze, and bursting charge are frequently included. If the components of a complete round are assembled into a single unit, the round is called a cartridge.

Rockets and guided missiles are sometimes classified as artillery ammunition, although they are not ordinarily launched from cannon. A rocket is a self-propelled missile that cannot be controlled after launching. A guided missile is also self-propelled, but it can be controlled during flights, either by internal means or remotely.

Reproduced by permission of the Department of the Army

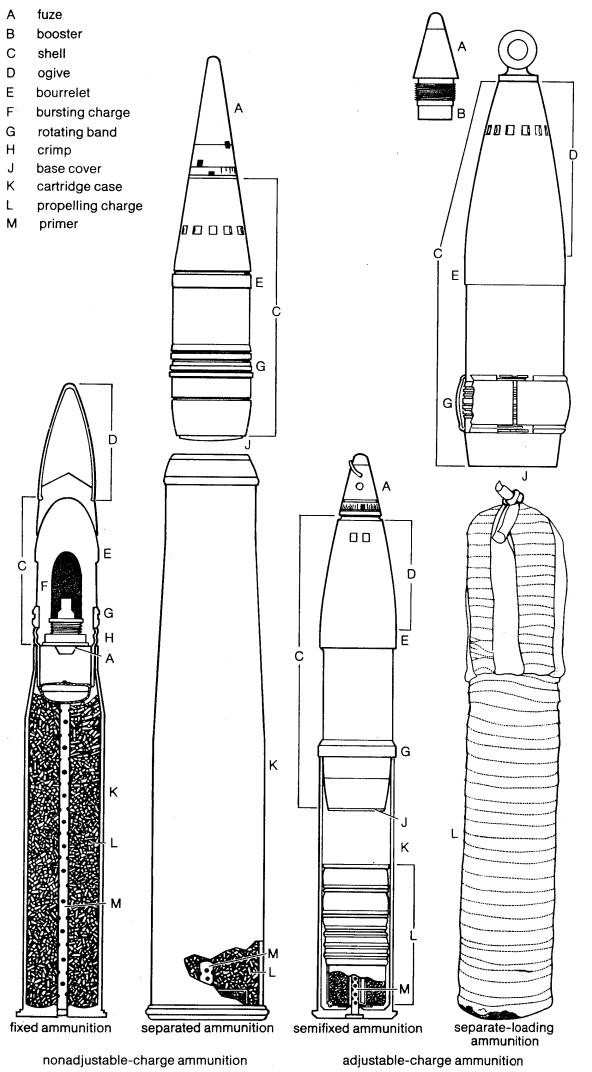


Figure 1: Terms used to identify the principal parts of typical kinds of artillery ammunition.

DEVELOPMENT OF AMMUNITION

Pre-firearms missile ammunition. The use of missile ammunition is undoubtedly as old as mankind, since even primates of lower order are known to throw objects at their adversaries. The first purposefully fashioned missiles were probably developed from thrusting and striking weapons. Thus, the throwing spear and the javelin evolved from the primitive thrusting spear, while the boomerang, tomahawk, and throwing hammer developed from the primitive club.

The bow and the sling both antedated the invention of writing, and therefore the invention of neither is recorded in history. Archaeological evidence indicates, however, that the bow and arrow existed before 10,000 BC. Simpler than the bow and arrow, the sling is probably at least as old.

Primitive arrows were thin sticks or reeds, fletched with small tufts of feathers and tipped with bits of stone. Fire arrows, among the first forms of incendiary missile ammunition, are also very ancient. Arrows were improved for more than 100 centuries, reaching a high point during the 13th to 16th centuries, in England. Typically three feet (90 centimetres) or more in length, English arrows could be shot from the English longbow with great rapidity, accuracy, and lethal effect to a range of several hundred yards or metres.

English arrows

The ammunition used in the sling was most often a smooth stone, of suitable size and shape. The ancient Phoenicians, however, renowned for the prowess of their slingmen, molded pellets of lead for ammunition.

The ammunition used in crossbows was an arrow-like missile, called a bolt or quarrel. Short and heavy, it was often fitted with a barbed metal head, designed to increase the severity of wounds produced upon penetration. Because of this feature in its ammunition designs, the crossbow was often looked upon as a barbarous and contemptible weapon.

Weapons of the individual warrior or huntsman, the sling, bow, and crossbow, were the precursors of modern small arms. The ancients also invented missile weapons operated by a team or crew of men; these were the precursors of modern artillery. According to the Bible, it was said of King Uzziah of Judah (783–742 BC) that "In Jerusalem he made engines, invented by skilful men, to be on the towers and the corners, to shoot arrows and great stones."

War engines were of various types. The Roman ballista could shoot ten- or 12-foot- (3–3.7-metre-) long javelins or a whole sheaf of arrows at once. The catapult used a great mechanical throwing arm, powered by heavy falling weights or by torsion of a bundle of elastic skeins, to hurl large stones or large projectiles molded from clay containing small stones. These molded projectiles shattered upon impact, sending the smaller stones ricocheting among the enemy. Some catapults were large enough to hurl a dead horse.

Origins and early use of gunpowder. Gunpowder is a mixture containing saltpetre (potassium nitrate), charcoal, and sulfur. The term now commonly applied to that mixture is black powder, in contradistinction to smokeless powder, a gun propellant invented during the 19th century.

Black powder and smokeless powder

The origin of gunpowder is subject to debate. Though some authorities believe it an ancient Chinese invention, most now assert that, although the Chinese invented various pyrotechnic mixtures similar to gunpowder, true gunpowder was a European invention of the 13th century. A manuscript written by the English philosopher and scientist Roger Bacon in 1248 describes the explosive properties of saltpetre, charcoal, and sulfur, but Bacon does not suggest use of the mixture as a propellant. Some 14th-century documents attributed to Berthold Schwarz, a German monk, describe gunpowder and its use in fire-

arms. Accounts of Schwarz's life and work are not established by irrefutable records, however, and some authorities doubt their validity. Authentic town records of Ghent (in modern Belgium) indicate that firearms existed there in 1313 or 1314. There also exists credible evidence that gunpowder and firearms were used in southern Europe by the early 14th century, and perhaps considerably earlier. The exact date of the invention of gunpowder probably will never be known.

Early firearms projectiles. An illustrated manuscript executed in 1326 contains one of the earliest drawings of a gun and ammunition. The manuscript depicts a primitive cannon in the act of being fired; the projectile emerging from the muzzle resembles a large crossbow quarrel. Balls of stone, lead, bronze, or iron were also used during the 14th century. The smaller firearms usually fired metal balls, but the larger guns fired balls of stone; these were lighter than solid metal balls and therefore better suited to the uncertain strength of the primitive cannon. Large mortars were sometimes loaded with numerous medium-sized stones, which were fired on high trajectories to fall among the enemy. Wadding or softwood plugs were sometimes employed to reduce the escape of propellant gases around the loosely fitting projectiles in the gun barrel. For incendiary purposes, iron balls were heated red-hot in a fire before loading. In that case, moist clay was sometimes packed atop the wadding to prevent premature firing of the powder charge by the red-hot ball.

Explosive projectiles were in use by the 16th century, perhaps even earlier. First used only in mortars, these were hollow cast-iron balls, filled with gunpowder, called bombs. A crude fuze was employed, consisting of a short tube, filled with a slow-burning powder, driven into a hole through the wall of the bomb.

Other early artillery projectiles, developed for special purposes, included the carcass, canister, grapeshot, chain shot, and bar shot. The carcass was a thin-walled shell containing incendiary material. Rounds of canister and grapeshot consisted of numerous small missiles, usually iron or lead balls, held together in various ways for simultaneous loading into the gun but designed to separate upon leaving the muzzle. Because they dispersed widely upon leaving the gun, the projectiles were especially effective at short range against massed troops. Bar shot and chain shot consisted of two heavy projectiles joined by a bar or a chain. Whirling in their trajectories, they were especially effective in cutting the spars and rigging of sailing vessels.

Grape shot

Projectiles for small arms did not change significantly in form or material from their earliest use until the 19th century. Lead balls, fired either singly or several at a time (as in a shotgun), were almost universally the ammunition of small arms for more than 400 years.

Developments during the 18th and 19th centuries. Advances in the physical sciences and in technology during the 18th and 19th centuries enabled the ordnance (military supplies) engineer to put into practice many ideas that had been conceived earlier but that had remained impractical. Weapon and ammunition developments were closely interrelated, each affecting the other. The breech-loading gun and the rifled bore were the most important weapon developments, while percussion ignition, projectiles for rifled guns, and smokeless powder were the most significant innovations in ammunition.

Ammunition for rifled guns. Rifling is the term applied to a system of helical grooves cut into the interior surface of a gun bore. The raised portions of the rifled bore that remain between the grooves are called lands. The function of rifling is to impart spin to the projectile in its passage along the bore. Though the date of the original invention of rifling cannot be determined precisely, it was in limited use at least as early as the 15th century. A rational explanation of the benefits of rifling was not put forth until 1747, however, when the English mathematician and engineer Benjamin Robins published a treatise on the subject. Systematic development of rifled guns began in the mid-18th century.

Rifling improves the accuracy of a round ball by imparting uniform spin. A greater benefit, however, is that rifling can impart the kind of stability manifest in a spinning top or a gyroscope to elongated projectiles, thus enabling such projectiles to maintain a point-forward attitude in flight. Elongated projectiles are aerodynamically much superior to round balls of the same weight and thus sustain their velocities better in flight. The introduction of rifling made possible the change from round lead balls to the now familiar bullet form of projectile, which is called cylindroconoidal, or, sometimes, cylindro-ogival.

Experiments with cylindroconoidal bullet forms began about 1825. Difficulties were experienced at first because the bullet was required to fit tightly in the barrel, and it was excessively difficult to load a tight-fitting bullet in a muzzle-loading arm. The most successful solution to this problem in small arms was a soft lead bullet having a cavity in the base. The bullet was made small enough for convenient loading; the gas pressure upon firing expanded the hollow base to fit tightly in the grooves and lands of the barrel. The best known projectile of the hollow-base type was the Minié bullet. Cylindroconoidal bullets were widely used in small arms by the mid-19th century.

The solution represented by the soft-lead Minié bullet in small arms could not be practically applied to projectiles for muzzle-loading artillery. Though many ingenious attempts were made to solve the problems of using cylindroconoidal projectiles in muzzle-loading cannon, none was entirely satisfactory. Artillerymen continued to fire spherical projectiles from smoothbore cannon for about 50 years after the practice was abandoned in small arms, or almost until the end of the 19th century.

Percussion ignition. The development of practical percussion-ignition is generally credited to Alexander Forsyth, a Scottish clergyman, whose avocational interests included both chemistry and firearms. Forsyth's priority in this invention is not unchallenged, because other men were achieving some degree of success, independently, at about the same time. Forsyth's percussion gunlock, patented in 1807, employed a small charge of chemical explosive (mercury fulminate or potassium chlorate) for ignition, which detonated upon being struck a sharp blow by the hammer of the gun. Such materials, originally called detonating powders, were not new in 1807. Forsyth's contribution was in the application of the material to guns.

For convenience in handling, and for protection from the elements, detonating powder was soon encapsulated in paper, foil, or thin metal for use in gunlocks. The percussion system was much more reliable than the flintlock. Trials were conducted in England during 1834, in which 6,000 rounds were fired in flintlock muskets and 6,000 in percussion guns. The weapons were loaded and fired at a rate of about three rounds per minute. The average frequency of misfires in the flintlocks was one per six or seven rounds, whereas the average frequency for the percussion arms was only one misfire per 166 rounds.

Ammunition for breech-loading guns. The development of strong and practical breech-loading cannon immediately solved the problems that had formerly thwarted efforts to use cylindroconoidal projectiles in artillery. A projectile soon emerged that had a band of soft metal encircling it at the rear, slightly larger in diameter than the body of the projectile. The rifling lands were removed from the gun bore adjacent to the breech to accommodate the banded rear part of the projectile and the powder charge seated behind it. The band on the projectile is called a rotating band or, sometimes, a driving band. When the gun is fired the band is engraved by the rifling lands, imparting spin to the projectile and sealing the bore against escape of the propellant gases. This design has not been superseded; it is still used in modern artillery ammunition.

In the case of small arms, the development of practical breech-loading weapons and self-contained cartridges proceeded together. The self-contained cartridge consists of a tube (the cartridge case) having the bullet affixed at the front end and the cap or primer at the rear, with the propellant contained in the tube between. Thus, all com-

The self-contained cartridge

ponents for a complete round are contained in one convenient assembly for handling and loading into the gun. The cartridge case also solved the problem of breech sealing (called obturation), which had been the most formidable obstacle to the development of practical breech-loading small arms.

Muzzle-loading rifles were obsolescent for military use by the end of the American Civil War in 1865. A great variety of breech-loading sporting arms and ammunition also appeared, especially in the United States, during the period 1865 to 1900.

Smokeless powder. The first smokeless powder successfully used in guns was invented in 1864. Moderately successful in sporting ammunition for shotguns, it burned too rapidly for rifles or cannon. The first successful smokeless powders for military use appeared between 1880 and 1890. There were two types, called single-base and double-base. Both were produced from guncotton, a form of nitrocellulose first described in 1846. Guncotton was not itself suitable as a propellant because it burned much too rapidly. Single-base powder, consisting almost entirely of nitrocellulose, was first prepared in France about 1884. Double-base powder, containing both nitrocellulose and nitroglycerine, was invented about 1888. Both types of powder were prepared by plasticizing guncotton with suitable solvents, rolling it into thin sheets, and cutting the sheets into small squares called granules or grains, which were then dried. The thickness and size of the grains controlled the rate of burning. Other types of double-base and single-base propellants soon appeared.

Smokeless powder has several important advantages over black powder. It produces about three times as much energy as black powder for the same weight of charge. The products of combustion are substantially all gaseous, whereas the combustion of black powder produces approximately 40 percent gaseous and 60 percent solid products. It was the solid products that produced much of the smoke upon firing black powder, and the new powder was pronounced "smokeless" by comparison, although it is not literally so. Elimination of solid combustion products also greatly reduced fouling of the bore, a serious nuisance for hundreds of years. Finally, smokeless powder was not destroyed by exposure to moisture, as black powder often was. By 1900 smokeless powder was in general military use.

MODERN AMMUNITION

Modern propellants and explosives. These include gun propellants, rocket and missile propellants, and high explosives (see also EXPLOSIVES).

Gun propellants. Most modern gun propellants are of single-base or double-base types. A few contain other energy-producing substances, such as nitroguanidine, in addition to nitrocellulose and nitroglycerine. Modern improvements in smokeless powder have included reductions in flash and smoke, improved stability in storage, reduction in erosion of gun bores, and improvements in control of the rate of burning.

Improvements in smokeless powder

Reductions in flash, smoke, and bore erosion are accomplished by careful selection and control of the ingredients that are added to the basic energy-producing composition of the propellant. Improved storage stability is achieved by the refinement of manufacturing processes, so as to remove impurities that accelerate deterioration, and by chemical additives called stabilizers.

Control of the burning rate is achieved by the composition, size, and geometric configuration of the propellant grains, and sometimes by surface treatment or coating of the grains. Generally, the goal is to produce a propellant that is slowly converted to gas in the initial stages of burning, and more rapidly converted as burning progresses. This characteristic is called progressive burning, and its antithesis is called degressive burning; propellants that are neither progressive burning nor degressive burning are described as neutral burning.

Propellant burns at the exposed surfaces. Therefore, geometric shapes that increase in surface area as they are consumed are progressive burning. A cylindrical grain that is perforated by several holes parallel to its axis is progressive burning until it is nearly consumed because the area inside each hole is increasing. A tubular grain is practically neutral burning, because the increase in inside area just offsets the simultaneous decrease in outside area as the grain is consumed. An unperforated cylinder or a sphere is geometrically degressive burning. Grains having degressive geometric configuration, however, can be made to burn progressively by impregnating the exposed surface with a substance called a deterrent coating, or moderant, which retards the burning rate of the basic material. Deterrent coatings are commonly used on modern small-arms propellants.

Rocket and missile propellants. Modern rocket and missile propellants are solid or liquid. Solid propellants are characterized, according to their composition, as double-base, cast perchlorate, and composite. Double-base rocket propellants are similar in composition to double-base gun propellants. Cast perchlorate propellants use mixtures of various perchlorates as oxidizers, with materials such as asphalt oil or polysulfide rubber as the fuels that are oxidized. Composite propellants usually contain ammonium picrate, potassium nitrate, or sodium nitrate, and a plastic binder.

Liquid propellants may be classified as monopropellants, bipropellants, or multipropellants. A monopropellant is a single liquid, containing both an oxidizer and a fuel, which is ignited by some external means in a manner similar to the ignition of solid propellants. Bipropellants consist of a liquid oxidizing agent and a liquid fuel that are kept separate and mixed only at the time of combustion. The ignition of bipropellants or multipropellants may occur spontaneously upon mixing, a phenomenon called hypergolic ignition. Liquid monopropellants are generally rather sensitive and unstable; for this reason bipropellants or multipropellants are usually preferred. Nitromethane is an example of a liquid monopropellant. One simple and effective bipropellant consists of liquid oxygen and liquid hydrogen, but there are literally thousands of other feasible combinations for bipropellants and multipropellants. German experimenters during World War II tested several thousand different combinations in developing propellants for their long-range rockets.

Liquid oxygen and hydrogen propellants

High explosives. High explosives differ from propellants (classified as low explosives) in respect to the rapidity with which they decompose. Whereas propellants are decomposed by burning, high explosives are decomposed by a phenomenon called detonation, which proceeds roughly 100,000 times as rapidly as the burning of a gun propellant (see EXPLOSIVES).

High explosives are classified, according to their sensitivity, as primary or secondary. Primary high explosives are relatively sensitive to detonation by shock, heat, or friction, but are generally inferior in strength and brisance (shattering effect) to secondary explosives. Because their detonation is easily initiated, primary explosives are used in small quantities to detonate the more powerful but less sensitive secondary explosives. Such an arrangement, in which a series of progressively less sensitive explosives are detonated in sequence, is called an explosive train. Some commonly used primary explosives are mercury fulminate, lead azide, and lead styphnate. A few examples of secondary explosives are TNT (trinitrotoluene), tetryl, and RDX (cyclo-trimethylene trinitramine, or cyclonite), but these and other substances are combined in various ways to make a great variety of secondary explosives that are used in modern ammunition.

Projectiles. Modern projectiles are divided into two general classes, according to whether they are intended to accomplish their purposes by kinetic or chemical energy. Kinetic-energy projectiles produce their effects on the target solely by virtue of their mass and striking velocity. Chemical-energy projectiles achieve their effects primarily by the explosion of a bursting charge, which is a chemical reaction.

Projectiles are also classified as spin-stabilized or fin-stabilized. Spin-stabilized projectiles remain point-forward in flight by virtue of gyroscopic action and must be

Stabilization of projectiles

fired from rifled guns. Fin-stabilized projectiles have arrow-like stability and may be fired from smooth-bore guns. Fin-stabilized projectiles can be designed to encounter less air resistance and thus retain more of their velocity in flight, but spin-stabilized projectiles usually fly more accurately along the desired trajectory.

Armour-piercing projectiles. Kinetic-energy armour-piercing projectiles depend upon a penetrator of hardened steel or of a harder material such as tungsten carbide. If the penetrator is the solid body of the projectile, it is called a full-calibre penetrator. When the penetrator takes the form of a hard core in the projectile, it is called a subcalibre penetrator.

Modern chemical-energy armour-piercing projectiles usually depend upon a shaped charge, a high explosive formed around a thin-walled metal cone, usually of copper. The open end of the cone faces the target and is covered by a hollow ballistic cap, or windshield. The cap enhances the flight characteristics of the projectile and also provides a predetermined optimum space, or stand-off distance, between the target and the explosive charge at the instant of detonation. The fuze and detonator are situated at the rear of the charge, behind the apex of the cone. Upon impact of the shell, the detonation begins at the rear of the charge, and the conical liner begins to collapse. The progressive collapse of the cone, starting at the apex, is accompanied by formation of a jet, along its central axis, moving at very high velocity. It is this high-velocity jet that achieves the remarkable penetration obtained with a shaped charge.

General-purpose and anti-personnel projectiles. The effectiveness of a general-purpose high-explosive shell depends in part upon the number, size, and velocities of fragments or submissiles that are produced upon detonation of the bursting charge. A small number of low-velocity submissiles is generally less effective than a large number of high-velocity submissiles. In an ordinary high-explosive shell, the fragments of the shell body are the submissiles. The number, size, and velocities of fragments are controlled by proper selection of the explosive used, by the metallurgical properties of the shell body, and by the mechanical design of the shell.

Another approach to improving the effectiveness of submissiles is to preform them and arrange for their projection and dispersal by the bursting charge. This approach was taken during the 19th century by a British officer, Henry Shrapnel, who encased musket balls inside the shell, from which they were propelled by a charge of black powder. A great improvement was the use of small steel darts, called fléchettes, expelled at high velocities by the bursting charge; a modern 90-millimetre (three-and-one-half-inch) projectile of this type contains more than 4,000 fléchettes.

Fléchettes

Fléchettes have also been used to increase the effectiveness of canister projectiles, which do not contain a fuze or a bursting charge but are designed to burst open immediately upon leaving the muzzle of the gun, thus releasing the submissiles. A modern canister for a 152-millimetre (six-inch) gun used by the United States contains approximately 10,000 fléchettes and has an effective range of 400 metres (1,300 feet). Though the use of fléchettes in small-arms ammunition has been explored in several countries, no such ammunition is yet in common use.

Nuclear artillery projectiles. The modern development of nuclear artillery ammunition is of profound potential importance, because the effects of nuclear bursting charges are enormously greater than those of chemical-energy explosives (see also NUCLEAR WEAPONS; GUNNERY).

The first nuclear bombs developed were very large; minimum size was fundamentally limited by the physical processes of nuclear fission, upon which their functioning depended. Fundamental improvements were required to develop nuclear munitions small enough to be fired from guns. In 1953 development of the first artillery weapon having a nuclear capability was announced. Calibre was 280 millimetres (11 inches), and the gun weighed approximately 85 tons (77,000 kilograms), not including the two motorized transporters that were required to move it. Subsequently, nuclear artillery of smaller calibres was developed; the U.S. Army reported an eight-inch (200-millimetre) projectile in 1956 and a six-inch (155-millimetre) projectile in 1963. Details of such nuclear projectiles have not yet been publicly disclosed by any nation.

Fuzes. Fuzes must be safe against inadvertent firing and reliable in action when initiated. The two basic conditions of a fuze are unarmed and armed. Unarmed fuzes are substantially incapable of functioning and are therefore safe to transport, handle, and affix to the munition for which they are intended. Armed fuzes are prepared to function, upon receiving the proper stimulus. Arming is accomplished in various ways, appropriate to the type of munition involved. Bomb fuzes may be armed by aerodynamic forces or by changes in barometric pressure. Fuzes for gun-launched projectiles may be armed by forces generated during firing, such as by linear acceleration (called setback), or spin. Fuzes for emplaced munitions, such as mines, may be manually armed. Combinations of these techniques are often used.

Missile fuzes are required to be safe, not only before launching but until they are sufficiently far along the trajectory so that a premature functioning would not endanger the launching weapon or its crew. The interval between launching and arming is called arming delay.

Fuzes can be classified according to the means by which stimulus for functioning is effected. The categories on this basis are impact, time, proximity, command, and combination fuzes. Impact fuzes, initiated by contact of the munition with a target, are sometimes called contact fuzes. They are categorized as superquick, nondelay, or delay, according to the interval of time between initiation and functioning. Time fuzes initiate functioning at some predetermined interval after the munition has been fired, dropped, or emplaced. The time interval is adjustable and may be controlled by a pyrotechnic train, mechanical clockwork, electrical circuitry, or other means. Proximity fuzes, also called influence, or VT, fuzes, function when an internal sensing mechanism detects that the target is close enough for effective functioning. The sensing element is generally a small radio unit that transmits signals and receives reflections from the target. Command fuzes, designed to receive a radio or wire-line signal sent from some remote point, are useful for emplaced munitions, such as mines or demolition charges, and for some types of guided missiles. Combination fuzes employ the functioning principles of two or more other types. Time fuzes, for example, may be designed so as to function upon impact if they strike a target before the set time has elapsed. Antiaircraft shells may be fuzed to burst aloft (self-destruct) in the event they miss the target, so as to avoid the hazards of a burst upon ground impact in friendly territory.

Complete rounds. These include artillery ammunition and military and nonmilitary small-arms ammunition.

Artillery ammunition. In separate-loading artillery ammunition, (Figure 1) a complete round consists of three components: the fuzed projectile, the propellant charge (in several combustible cloth bags), and the primer. In fixed rounds, all the components are securely joined together by a cartridge case. In semifixed ammunition, the projectile is not securely joined to the cartridge case, an arrangement that allows for field adjustment of the propellant charge. After the charge has been adjusted, the projectile is loosely inserted into the cartridge case and loaded in the weapon in the same manner as a fixed round. Separated ammunition consists of a primed cartridge case containing the propellant charge and sealed by a closing plug, and a projectile separately handled and separately loaded into the gun. Though brass was almost invariably used for cartridge cases before World War II, it has since been largely superseded by steel. A recent innovation is a combustible cartridge case, entirely consumed upon firing.

Complete rounds are further classified according to the type of projectile employed, such as high-explosive, or armour-piercing. There are also special-purpose types, including blank rounds and dummy rounds, used for sa-

luting, ceremonies, manoeuvres, drills, instruction, or displays.

Recoilless ammunition

A special type of complete round is made for recoilless guns. The cartridge cases for recoilless ammunition are so designed, often by a pattern of perforations in the cartridge wall, that propellant gas rapidly escapes from them on firing. The gun chamber containing the case is also vented, allowing gas to escape into a peripheral space around the case, from whence it is directed rearward, where it finally escapes through a venturi nozzle (containing a constricted part to control the rate of gas flow) in the breech. The rearward escape of gas through the venturi is carefully regulated to ensure that the resulting forces of reaction are equal to the recoil forces caused by propulsion of the projectile and the gases expelled at the muzzle. The gun is thus acted upon simultaneously by two opposite impulses of equal magnitude, the sum of which is zero, and the gun is not disturbed by the forces of firing.

Military small-arms ammunition. Small-arms ammunition is always of the fixed type; complete rounds (see Figure 2) are usually called cartridges. In calibres larger than 0.60 inch (15 millimetres), the projectiles are similar to artillery projectiles, and the nomenclature of the ammunition is generally the same. In calibres of 0.60 inch or less, small-arms projectiles are usually called bullets.

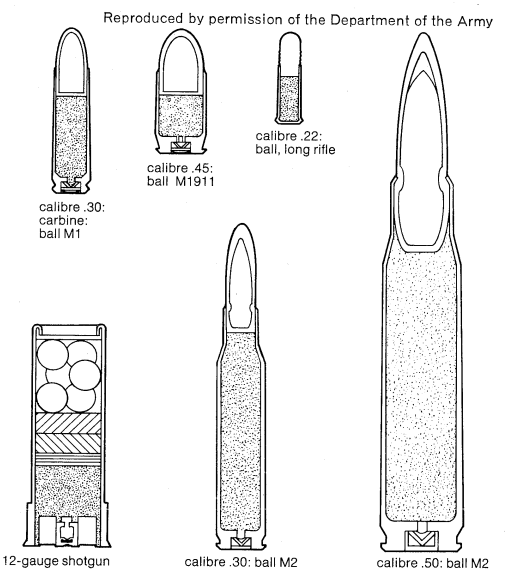


Figure 2: Types of small-arms cartridges.

Cartridge cases for small-arms ammunition are most commonly made of brass, although steel is also widely used. Though lighter materials, such as aluminum and plastics, have received consideration, they have not yet found significant use in military ammunition. Caseless cartridges, employing a type of molded propellant charge, have also been investigated experimentally.

The most common type of military small-arms ammunition is called general-purpose in England, and ball in the United States. The term ball is historical nomenclature, having originated when the lead ball was the common form of small-arms projectile. The bullet of modern ball ammunition usually consists of a lead-alloy core, encased in an envelope or jacket of harder metal. Jackets are usually made of a copper alloy, or of mild steel coated with a copper alloy. Cores of ball ammunition are occasionally made of mild steel instead of lead alloy.

Special-purpose ammunition for small arms includes armour-piercing, tracer, incendiary, and various dual-purpose or multi-purpose types that combine different characteristics, such as armour-piercing-incendiary-tracer (API-T). Other special-purpose ammunition includes blank, dummy, grenade, and spotting cartridges.

Armour-piercing bullets have cores of hardened steel or are made of harder materials, such as tungsten carbide. Tracer bullets have a column of pyrotechnic composition in the base that is ignited by the flame of the propellant; this provides a visible pyrotechnic display during the bullet's flight. Incendiary bullets, intended to ignite flammable materials such as gasoline, contain a charge of chemical incendiary agent. Spotting bullets contain chemical agents that produce a visible flash, or a puff of smoke, upon impact. They are used principally to adjust the fire of larger guns so arranged that the trajectory coincides approximately with that of the spotting bullet. Blank cartridges contain no projectile, and are used for manoeuvres and training. Grenade cartridges contain no projectile, but contain a special propellant charge for launching rifle grenades, which are affixed separately to the muzzle of the rifle in preparation for firing. Dummy cartridges contain no explosive components and are used for instruction, display, and similar purposes.

Incendiary bullets

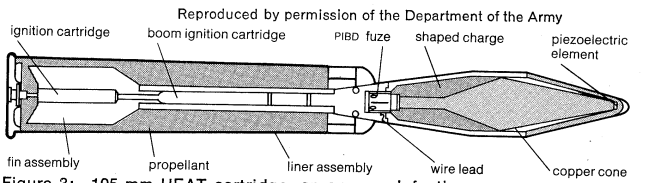


Figure 3: 105-mm HEAT cartridge, an armour-defeating round.

A significant 20th-century improvement in small-arms ammunition was the development of primer compositions to replace the old mixtures containing potassium chlorate, of which a decomposition product was potassium chloride, a corrosive substance.

Nonmilitary small-arms ammunition. Most nonmilitary small-arms ammunition is used for hunting and target shooting. Some is used by civilian law-enforcement agencies. The three technical classifications of commercial ammunition are rimfire, centre-fire metallic-case, and shotshells.

Rimfire ammunition does not require a separate primer but contains primer mixture within a folded rim or flange that is formed integrally with the cartridge case. The virtue of this design is its low cost. Thin-walled, rimfire cases can withstand only moderate chamber pressures and are now used only in inexpensive types of ammunition, such as the common .22-calibre.

Centre-fire metallic-case ammunition includes practically all types used in rifles and handguns, excepting .22-calibre rimfire. The primer constitutes a separate component, as in military ammunition. Cases are brass and are classified as rimmed, semirimmed, rimless, or belted, according to the details of their shape. The bullets may be of lead alloy, if the velocity and pressure are low. If higher velocity and pressure are required, jacketed bullets must be used. The details of bullet design vary widely, according to their intended purpose and the respective practices of various manufacturers.

Shotshells

Shotshells contain numerous round lead pellets, called shot, instead of a single bullet. Shotshell cartridge cases were made for many years by assembling coiled paper tubes to cuplike metal heads. In recent years, plastic materials have largely replaced paper in shotshells. The size of shotshells is indicated by gauge instead of by calibre. A 12-gauge shotgun is one having a bore diameter corresponding to the diameter of a lead ball that weighs one-twelfth of a pound (37.8 grams). The gauges that have been generally standardized in all countries are 12, 16, 20, and 28. There is also a .410-bore size, which, in exception to common practice, is designated by calibre, as in rifle ammunition. The sizes of shot used for small-game hunting vary from about 0.08 to 0.15 inch (two to four millimetres) in diameter, and are identified by arbitrary number systems that vary from country to country. Buckshot are larger; sizes up to about 0.34 inch (about 8.5 millimetres) in diameter are frequently used for hunting medium-sized animals such as deer. Some shotgun ammunition is intended for the competitive shooting sports of trap and skeet, in which the object is

to break frangible flying targets hurled from a mechanical throwing device called a trap. The size of shot and weight of shot charge for competitive shooting are stipulated by the rules of the sport, according to the gauge of the gun being used (see GUNS, SPORTING AND TARGET).

Rockets and missiles. The use of self-propelled missiles is ancient, rockets having been invented before closed-breech guns. During the 18th and early 19th centuries, rockets were often used in warfare. During the 19th and early 20th centuries, however, the development of guns surpassed that of rockets, which then passed into disuse until World War II.

During World War II, scientists and engineers again developed rockets into weapons of great tactical and strategic importance. The American bazooka of World War II was an effective portable rocket launcher for use against tanks. The towed multishot rocket launcher of Germany was light and highly mobile; it could produce a barrage comparable to that of several field guns. The V-2 rockets launched by Germany against England were of some immediate strategic importance but even more importantly were the precursors of modern missile systems (see also WEAPONS AND DELIVERY SYSTEMS; ROCKETS AND MISSILE SYSTEMS).

At short and medium ranges, rockets have an advantage over conventional artillery ammunition because they can be fired from launchers which are much cheaper, lighter, and more mobile than guns. At very long ranges, self-propelled missiles have capabilities that cannot be remotely approached by gun-launched projectiles having ballistic trajectories.

BIBLIOGRAPHY. Extensive historical material, as well as technical information of general interest, may be found in W.H.B. and J.E. SMITH, *Small Arms of the World*, 7th ed. (1962); and JAC WELLER, *Weapons and Tactics: Hastings to Berlin* (1966). Two classic works particularly recommended, are the *Textbook of Small Arms* (HMSO, 1929); and W.W. GREENER, *The Gun and Its Development*, 9th ed. (1910). For general information on small arms, see C.S. CUMMINGS, *Everyday Ballistics* (1950); and E.D. LOWRY, *Interior Ballistics: How a Gun Converts Chemical Energy into Projectile Motion* (1968). T.J. HAYES, *Elements of Ordnance* (1938), contains professional-level material on small arms, artillery, and weapons.

(W.C.Da.)

Amphibia

The Amphibia constitute a class of vertebrate animals in the evolutionary series between the fishes and the reptiles. Living amphibians include the frogs and toads, the salamanders, the caecilians, and the sirens. The amphibians are of primary interest to students of evolution because they were the first vertebrates to make their way from an aquatic environment to land and because the group includes the ancestors of all the other terrestrial vertebrates—the reptiles, birds, and mammals.

General features. *Economic importance.* The amphibians are of little direct economic importance to man. Toads of the genus *Bufo* have long been noted as avid devourers of insect pests. Large species have been introduced into Florida, Jamaica, and Hawaii to help control pests in the cane fields. There is some demand for frog legs as table delicacies.

Amphibians are of great importance to biologists. Most students have their first introduction to vertebrate anatomy through the dissection of a frog. Salamanders have contributed to knowledge of embryology and of tissue regeneration.

Size range and diversity of structure. In size the living amphibians range from the tiny frog *Sminthillus limbatus* of Cuba, less than 12 millimetres (about one-half inch) in length, to the giant salamander *Andrias* of eastern Asia, which reaches a total length of 160 centimetres (about 63 inches). There are fewer species of amphibians than of any other terrestrial vertebrate class; they show a great diversity of structure, however, ranging from the elongated, legless, wormlike caecilians to the short-bodied, tailless, long-legged frogs.

Distribution and abundance. The amphibians are concentrated in the tropics; some families and many genera

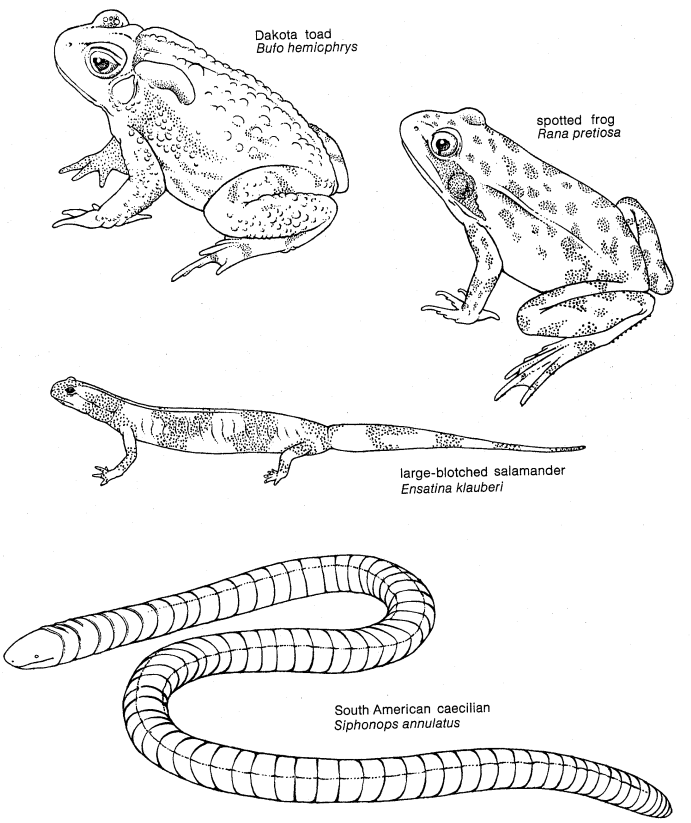


Figure 1: Representative amphibians.
Drawing by S. Jones

are confined to these regions. As a group, though, the amphibians have a worldwide distribution. The salamanders are primarily animals of the North Temperate Zone, although the range of one family extends into Central America and northern South America, where it has evolved a large number of species. Frogs of the family Hylidae range as far north as the swamplands of northwestern Canada; and the family of true frogs, Ranidae, is found on all major landmasses except Antarctica, New Zealand, and Greenland. The caecilians occur throughout the tropics and the sirens only in southeastern United States and northeastern Mexico.

Members of this class, for the most part, are found as isolated individuals; but many of the frogs, particularly in the temperate zones, assemble in large breeding choruses during the spring and summer.

Natural history. *Life cycle.* The amphibians are so named because most members have aquatic larval stages that metamorphose (transform) into terrestrial adults—*amphi* is the ancient Greek word for "both" and *bios* means "life." A few extant species remain aquatic as adults. Yet many species of salamanders and frogs and all egg-laying caecilians deposit their eggs away from open water, and some species lack the aquatic larval stage. Because no amphibian has a shelled egg, eggs must be deposited in a moist place—soil, rotting wood, a foam nest constructed by the parent, or the skin of the mother's back.

Behaviour. Because their life history involves both land and water, most amphibians have behaviour patterns built around a seasonal movement between breeding sites and other places.

For the caecilians and sirens the breeding site is probably essentially the same as the nonbreeding site. Although this is also true for some salamanders, others must undertake annual journeys, perhaps from a relatively dry hillside to some pond or stream in a valley. Evidence indicates that they are guided on such journeys by chemical stimuli as well as by physical characteristics of the land (*e.g.*, slope). In frogs and toads that have voices, the calls of the males that have located a suitable

breeding site attract other males and females, often from one-fourth to one-half mile away. It has been demonstrated that some amphibians are able to use celestial bodies for orientation.

Fertilization and egg laying

Reproduction. Fertilization is internal in the caecilians and most salamanders. In caecilians, sperm are introduced into the female's body by means of a protrusible portion of the male's cloaca (*i.e.*, the lower end of the digestive tract), which is modified to form a copulatory organ. The male salamander deposits his sperm in a gelatinous structure (spermatophore) consisting of a stalk on which rests a little cap containing the sperm. The female manoeuvres the cap into her cloaca, where the gelatinous material dissolves, and the sperm are freed to fertilize the eggs as they descend through her reproductive tubes. It is not yet known whether fertilization is internal or external in the sirens.

With few exceptions, the frogs and toads have external fertilization. The male clasps the female and releases sperm over the eggs as the female extrudes them.

Most caecilians deposit their eggs on land; the aquatic forms bear living young. Sirens and many salamanders lay eggs in water. Other salamanders lay them on land, in burrows, in rotten logs, or under debris where they will remain moist. Frogs have various means of caring for the eggs after deposition. In the North Temperate regions most species abandon their eggs in a pool or stream; many tropical species have developed other, sometimes bizarre, techniques of egg care. Some lay eggs in a hole on land; after the eggs hatch into tadpoles, the male transports them on his back to water. In other species the mother has a special pouch on her back in which she carries the eggs until they hatch. The male of Darwin's frog (*Rhinoderma darwinii*) picks up the eggs with his tongue and they hatch in his vocal pouches, where the young develop.

Ecology. The primary element in the ecology of amphibians is their need for a moist environment. With their soft, scaleless, glandular skins, they are very susceptible to drying out. Although a few species spend their entire lives in water, most live at least part of the time on land, either under moist debris or not far from water. A few, such as the toads, have become quite terrestrial, but even they are not entirely independent of moisture. The familiar garden toads, often seen at night, live where there is moisture near the surface; during the day, when water evaporates rapidly, they tend to remain in their burrows.

Form and function. *Distinguishing features.* The four major groups of living amphibians differ markedly from one another in structure. Caecilians are slender, wormlike animals without limbs or limb girdles, with practically no tail, and with a simple intestine. The minute eyes are buried in the smooth skin, and there are no eyelids. The largest grow to about 135 centimetres (about 53 inches). Because caecilians use their heads to burrow in the soil of the forest floor and in the mud along riverbanks, the skull is compact. The cloaca of the male is modified into a protrusible copulatory organ. The body is segmented into circular folds, called annuli, separated by shallow grooves. The more primitive species have tiny scales imbedded in the skin. Some aquatic forms have a dorsal fin fold. All caecilians have tentacles, small sensory structures that lie buried in pits near each eye and are protrusible in many species. Several series of teeth are arranged on different bones around the jaws.

The sirens, or trachystomes, which lack hindlimbs and pelvic girdles, are elongated, aquatic animals, with an anterior pectoral girdle to which the upper bones of the front legs are attached. The tiny front legs have three or four toes on each foot. Gills are present. Sirens also breathe by gulping the surface air. The eyes are lidless and are buried in the smooth glandular skin, which is lacking in scales and annuli, although so-called costal folds are present along each side of the body. A horny claw caps each toe, and a horny sheath covers part of each jaw. Except for the tiny splenial bones, the bones of the jaws lack teeth; but teeth are present on the bones forming the roof of the mouth. The well-developed tail has dorsal and ventral fin folds. Sinuous body motions and the tail fins are used for locomotion.

The salamanders, which have forelegs and hindlegs, retain a tail throughout life. A distinct neck is present between the head and body. The smooth skin lacks scales, and the eyes of terrestrial forms have lids. Teeth are present on the jaws and roof of the mouth. Although fertilization is internal in most species, the male lacks a copulatory organ, sperm transfer occuring via spermatophores (sperm capsules). Some species never metamorphose but remain in the water as permanent larvae, retaining gills throughout life.

Leg bones of frogs and toads

The largest group of amphibians are the familiar frogs and toads. The one characteristic that clearly distinguishes this group from others is the presence in the hindleg of a segment formed by two elongated tarsal bones, the astragalus and calcaneum. In salamanders these bones are small, merely forming part of the ankle. The elongated hindleg of frogs and toads is an adaptation for hopping or swimming. Adults lack a tail; and teeth, although usually found on the upper jaw, are (except in one tropical species) absent from the lower jaw. The tympanic membranes, used for hearing, are on the sides of the head, behind the eyes. Males of species with well developed voices have resonating pouches, called vocal sacs, in the throat region. The glandular skin, which lacks scales, is usually smooth and soft; in some terrestrial forms, however, it is rather dry and rough.

Physiology and biochemistry. The amphibians have not evolved an impermeable skin. They have, therefore, special problems involving internal water balance and the removal of nitrogenous wastes—by-products of metabolism. They take in water through the skin and excrete it, along with the nitrogenous wastes, through the kidneys.

Most larval amphibians and permanently aquatic adult forms live in fresh water. Water, which flows constantly into the body, removes nitrogenous wastes in the form of ammonia. The kidneys produce copious amounts of very dilute urine, thereby eliminating both excess water and waste products. Terrestrial forms, lacking an unlimited supply of water, cannot excrete such dilute urine; instead, they convert their nitrogenous wastes into urea, which, because it is less toxic than ammonia, can be excreted in a much more concentrated form. As a result, less water is lost through excretion.

A physiological mechanism also controls urine production in relation to temperature. As temperature rises, water flows more rapidly through the skin, and urine flow increases. In experiments with the green frog (*Rana clamitans*) it has been found that no urine was formed at 2.5° C (36.5° F), although some water moved into the body.

Evolution. Fossil remains of the earliest known amphibians have been found in freshwater beds of the Upper Devonian (about 370,000,000 years old). The amphibians evolved from the lobe-finned fishes (Crossopterygii), which were abundant and diversified during the Devonian. These fishes had muscular fins supported by bony elements comparable to the bones of amphibian limbs; functional lungs; and, in certain species, passages leading from the nasal sacs to openings (internal nares) in the roof of the mouth, an arrangement characteristic of all terrestrial vertebrates. The crossopterygian skull was divided transversely into front and back parts that were slightly movable on each other. Early crossopterygians had teeth with an infolding of the enamel surface. Similar teeth were also present in the primitive amphibians.

The amphibian ancestor was able to take air into the mouth and lungs through the nasal passages and internal nostrils when the mouth was closed and only the external nostrils were above water.

Evolution from water to land

The question of why the ancestral amphibian went from an aquatic environment to a terrestrial one has been much debated. One suggestion is that in times of drought these animals, with their muscular fins and well-developed lungs, would have been able to heave themselves out of the drying pools and travel overland in search of

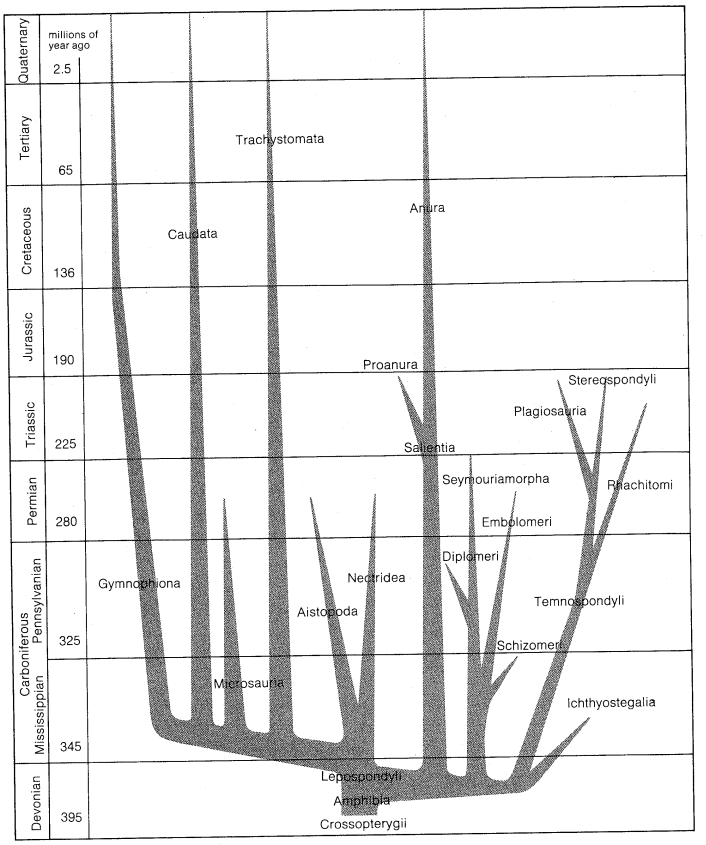


Figure 2: Dendrogram of Amphibia. The fossil record of many forms is incomplete, particularly during the Carboniferous Period.
From J.Z. Young, *The Life of Vertebrates* 2nd. ed. (1962); the Clarendon Press, Oxford

other bodies of water. It has also been proposed that, since the crossopterygians were carnivorous, they probably preyed on the young of various species, including their own. Small individuals may have remained in shallow water inaccessible to larger fish or they may have fled to the land in order to escape a predator. Perhaps they later spent more time on shore, especially at night when the humidity was high, returning to water to feed during the day; gradually, they could have supplemented their diet with primitive terrestrial arthropods, and so become less dependent on water except for breeding purposes.

The earliest amphibians of the Upper Devonian are the ichthyostegalians. The best known member of the group, *Ichthyostega* from freshwater beds of Greenland, was a fairly large animal with a skull 150 millimetres (about six inches) or more in length. Instead of fins it had short, stubby limbs with five toes; its tail was something like that of a modern lungfish, with a caudal, or tail, fin supported by fin rays. Small, bony scales were present on the belly and tail. The vertebrae were like those of crossopterygians, the skull more amphibian-like. *Ichthyostega* probably spent most of its time in water, but the well-developed limbs indicate that it must have been capable of walking on land. In many ways the ichthyostegalians were intermediate between the crossopterygians and the later amphibians, but they had some specializations that suggest they represent a sideline rather than the main stock from which the later amphibians evolved. One of the bones of the skull, the intertemporal, which is present in many fossil amphibians, had been lost by the ichthyostegalians.

Fossil remains of amphibians are rare in deposits of Mississippian age (325,000,000 to 345,000,000 years ago), but during the Pennsylvanian (280,000,000 to 325,000,000 years old), amphibians were abundant and diversified. They were divided into three major stocks: temnospondyls, anthracosaurs, and lepospondyls. The earliest temnospondyls were probably aquatic, but some later members of the group were well adapted to life on land. They were a widespread and flourishing stock, showing considerable morphological diversity. Almost all the large amphibians of the Permian (225,000,000 to 280,000,000 years ago) belonged to this group. *Eryops* of the Lower Permian (about 250,000,000 to 280,000,000 years ago) was about 150 centimetres (about 60 inches) long. As the reptiles spread over the land during the Late Paleozoic (about 225,000,000 to 325,000,000 years ago), the terrestrial amphibians declined. The temnospondyls of the Triassic Period (about 190,000,000 to 225,000,000 years ago) were aquatic animals with reduced limbs. Some were very large. *Mastodonsaurus*, with a skull length of over 90 centimetres (about three feet), was the largest known amphibian. The temnospondyls were extinct by the end of the Triassic.

The centrum (base) of the vertebra in primitive anthracosaurs was formed of a crescent-shaped intercentrum and two pleurocentra shaped like half rings. In more advanced forms the intercentrum was reduced and the pleurocentra became enlarged, joining to form a complete ring around the notochord (*i.e.*, the skeletal rod passing through the centre of the vertebral column). The anthracosaurs, though less numerous and diversified than the temnospondyls, also included both aquatic and terrestrial forms. The fish-eating *Pteroplax* of the Pennsylvanian reached an estimated length of more than 450 centimetres (nearly 15 feet). The anthracosaurs are of interest because the reptiles evolved from them. They became extinct during the Permian.

Lepospondyls of the Paleozoic Era were of modest size; many were eel-like or snakelike in appearance and had lost their limbs. They were aquatic or semi-aquatic; a few may have been fossorial—*i.e.*, burrowing forms. They were present in the Lower Mississippian, flourished during the Pennsylvanian, and disappeared in the Lower Permian.

From A.S. Romer, *Vertebrate Paleontology*, Copyright 1966 by the University of Chicago Press, all rights reserved

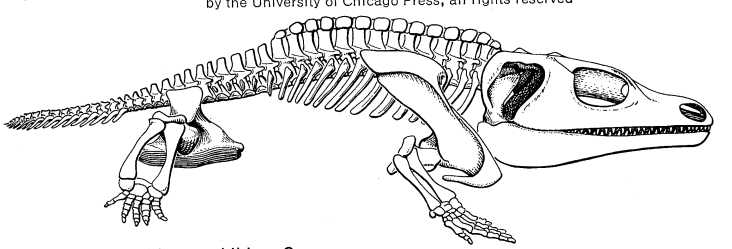


Figure 3: A primitive amphibian, *Cacops*.

Origin of modern orders

The modern orders of amphibians cannot be positively traced to any of the Paleozoic groups. No fossil caecilians are known. The trachystomes have been found in the Cretaceous and the salamanders in the Jurassic. Both groups seem to have evolved in the Northern Hemisphere; in fact, only recently have members of one family of salamanders (Plethodontidae) invaded the Southern Hemisphere in South America. The earliest remains of frogs, on the other hand, have all been found in southern landmasses. It has been suggested that certain footprints in the Permian of South Africa may have been made by a form ancestral to the frogs. *Triadobatrachus* of the Lower Triassic (about 210,000,000 to 225,000,000 years ago) of Madagascar had a skull like that of a frog, elongated hindlegs, and a short tail. Frogs were undoubtedly present in the Lower Jurassic (about 160,000,000 to 190,000,000 years ago) in the southern part of South America.

Classification. *Distinguishing taxonomic features.* The method of formation of the vertebral centrum (*i.e.*, the heavy ventral, or bottom, part that connects with the adjacent vertebra) is important in the classification of the amphibians, particularly of fossil forms. The recent orders are distinguished by general body structure and by the presence or absence of limbs and tail. The structure of the sound-transmitting mechanism of the ear is also an important clue to identification. Adult frogs have an otic notch on either side of the skull that houses the tympanic membrane, or eardrum. Salamanders, caecilians, and sirens lack the otic notch, eardrum, and middle-ear cavity.

Sound transmission in salamanders is by way of the forelimb and shoulder blade to a muscle (opercularis) that passes from the shoulder blade to the operculum, a bone in the ear. Some fossil amphibians have the otic notch; presumably, many also possessed an eardrum and middle-ear cavity. Others lack the otic notch.

Annotated classification. The classification of the amphibians has not yet been satisfactorily established. The lack of critical fossil material prevents the establishment of clear-cut relationships between the major groups. The arrangement below is based on that of Romer (1966) and Goin and Goin (1971). Groups indicated by a dagger (†) are extinct and known only from fossils.

CLASS AMPHIBIA
Vertebrate animals without hair, feathers, or horny scales of epidermal origin (epidermis is the outer cell layer of the integument); small bony scales of dermal origin (dermis is the inner cell layer of the integument) occasionally present; skin soft and glandular; ectothermic (*i.e.*, dependent on external sources for body heat); fertilization external or internal; eggs usually laid in water; typically an aquatic larval stage; about 2,400 living species.

†Superorder Ichthyostegalia
Fossils only; Upper Devonian to Lower Mississippian; Europe, North America, Greenland; with limbs and fishlike tail; otic notch present; size to about 90 cm; probably largely aquatic.

†Superorder Temnospondyli
Fossils only; centrum primitively formed of wedge-shaped blocks; otic notch present; no more than 4 toes on front feet.

Order Rhachitomi
Upper Mississippian through Lower Triassic; worldwide distribution; intercentrum (crescent-shaped structure of centrum) not forming a complete ring, pleurocentra (semicircular structures of centrum) bony; size to 150 cm; aquatic and terrestrial.

Order Stereospondyli
Middle Permian to Upper Triassic; worldwide distribution; intercentra sometimes forming complete disks, pleurocentra reduced or absent; includes largest known amphibian, *Mastodonsaurus;* all forms aquatic.

Order Plagiosauria
Upper Permian to Upper Triassic; Europe and Africa; centrum closed cylinder; body broad and flat; scales present; size to about 90 cm; aquatic.

Superorder Lepospondyli
Centrum a bony cylinder around notochord; body more or less elongated; otic notch usually absent; limbs frequently reduced or absent.

†Order Aistopoda
Fossils only; Lower Mississippian to Lower Permian; North America and Europe; elongated body form, limbs absent; ribs attached to vertebrae by one head; length about 75 cm; possibly fossorial (burrowing).

†Order Nectridea
Fossils only; Lower Mississippian to Lower Permian; Europe and North America; body elongated or broad and flat; limbs reduced or absent; caudal vertebrae with expanded fan-shaped neural (dorsal) spines and hemal (ventral) spines; otic notch sometimes present, though small; some with cranium extended into posterior "horns"; length to about 60 cm (about 2 feet); aquatic, some possibly fossorial.

Order Trachystomata (sirens)
Lower Cretaceous to Recent; North America; eel-like forms, hindlimbs absent, forelimbs minute; expanded spines on caudal vertebrae; ribs 2-headed; length to 90 cm; 3 living species, all aquatic.

†Order Microsauria
All fossil; Lower Mississippian to Lower Permian; Europe and North America; body somewhat elongated, limbs rather small; caudal vertebrae without expanded spines; ribs 2-headed; small forms, probably most were semi-aquatic.

Order Gymnophiona (caecilians)
Recent; worldwide in tropics; wormlike burrowing forms without limbs or limb girdles; eyes reduced; fertilization internal, with copulation; length to about 135 cm (about 53 inches) but most much smaller; burrowing or aquatic; about 160 living species.

Order Caudata (salamanders)
Jurassic to Recent; worldwide in Northern Hemisphere, 1 family widely distributed in northern South America; vertebral centrum develops by direct development of bone tissue around notochord; ribs 2-headed; fertilization external or internal by means of spermatophore; size to 160 cm (about 63 inches) but most much smaller; terrestrial or aquatic; about 300 living species (see URODELA).

Superorder Salientia
(frogs and their possible Triassic predecessor)
Triassic to Recent; worldwide.

Order Proanura
One fossil form from Triassic of Madagascar; skull like that of a frog; tail short; hindlegs elongated.

Order Anura (frogs and toads)
Jurassic to Recent; worldwide; hindlegs elongated, body short and stout, tail absent; centrum preformed in cartilage; otic notch present; fertilization usually external, no spermatophore; body length to 30 cm (about one foot); terrestrial or aquatic; about 1,900 living species (see ANURA).

†Superorder Anthracosauria
Fossils only; elements of centrum primitively crescent-shaped, pleurocentra later fuse to form main body of centrum, intercentrum reduced; otic notch present; 5 toes on front feet.

Order Schizomeri
Upper Mississippian, Europe; pleurocentra paired half rings, intercentrum large and crescent-shaped; trunk elongated, limbs reduced; length to about 120 cm (about 4 feet); probably aquatic.

Order Embolomeri
Upper Mississippian to Lower Permian; Europe and North America; intercentrum and pleurocentrum complete disks, body elongated, tail powerful, limbs reduced; length to over 450 cm (about 15 feet); probably aquatic eaters of fish.

Order Seymouriamorpha
Upper Pennsylvanian to Upper Permian; North America, Europe, Australia; intercentrum reduced and crescent-shaped, pleurocentrum ring-shaped; limbs well developed, body stout; otic notch very large; length to 300 cm (about 10 feet); terrestrial.

Order Diplomeri
Middle Pennsylvanian; Europe and North America; intercentrum large and crescentic, pleurocentrum disk-shaped; limbs well developed; otic notch present, skull otherwise like that of a primitive reptile; presumably represents stock from which reptiles evolved; body length (without tail) about 10 cm (about 2.5 inches); terrestrial.

Critical appraisal. There is no agreement among herpetologists that the amphibians evolved from a single stock. It has been suggested that the fossil lepospondyls and the salamanders are derived from one group of crossopterygians (porolepiforms) and that the ichthyostegids, temnospondyls, anthracosaurs, and frogs are derived from another (osteolepiforms). For alternate classification of Caudata (salamanders) see URODELA.

A.S. Romer and others divide the amphibians into three subclasses: Labyrinthodontia for the ichthyostegalians, temnospondyls, and anthracosaurs; Lepospondyli for the fossil orders of lepospondyls; and Lissamphibia for the modern amphibians. The combination of ichthyostegalians, temnospondyls, and anthracosaurs into one subclass is based in large part on the assumption that the three-part centrum found in primitive members of the last two groups is derived from the crossopterygian-type centrum of *Ichthyostega*. It has been suggested, however, that this assumption is false and that the divided centrum of the temnospondyls and anthracosaurs represents a new, tetrapod structure.

Whether the modern amphibians should be placed in one suborder is another much-debated question, which will probably not be answered unless and until sufficient fossil material is found to allow the modern orders to be traced to their Paleozoic ancestors.

BIBLIOGRAPHY. FERNAND ANGEL, *Vie et moeurs des amphibiens* (1947), a popular account in French; DORIS M. COCHRAN, *Living Amphibians of the World* (1961), a lively and beautifully illustrated popular account; E.T.B. FRANCIS, *The Anatomy of the Salamander* (1934), the only complete account of salamander anatomy; HANS GADOW, *Amphibia and Reptiles* (1901, reprinted 1958), an English classic, old but still worth reading; E. GAUPP, *Ecker and Wiedersheims' Anatomy of the Frog* (1904), the standard work on this topic; COLEMAN J. and OLIVE B. GOIN, *Introduction to Herpetology*, 2nd ed. (1971), a college-level textbook; ROBERT MERTENS, *La Vie des amphibiens et reptiles* (1959; Eng. trans., *World of Amphibians and Reptiles*, 1960), a beautifully written and

illustrated popular book; JOHN A. MOORE (ed.), *Physiology of the Amphibia* (1964), technical but with much useful information; G.K. NOBLE, *Biology of the Amphibia* (1931, reprinted 1954), the standard English work on amphibians, already a classic; JAMES A. PETERS, *Dictionary of Herpetology* (1964), a useful guide, especially for non-specialists, to an understanding of the technical terms used by herpetologists; A.S. ROMER, *Vertebrate Paleontology*, 3rd ed. (1966), a standard text; EDWARD H. TAYLOR, *The Caecilians of the World* (1968), the only comprehensive account of the order; J.Z. YOUNG, *The Life of Vertebrates*, 2nd ed. (1962), an evolutionary approach that contains much more than the usual anatomical accounts.

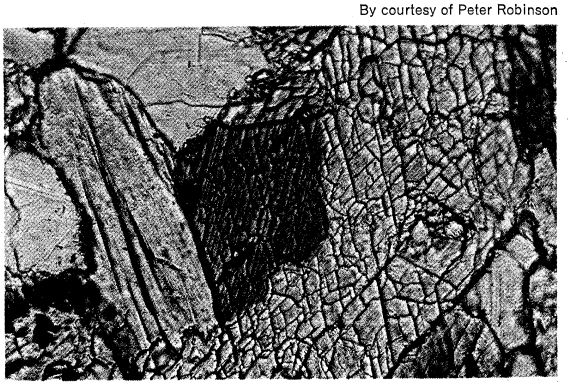
(C.J.G.)

Amphiboles

Definition and general properties

The amphiboles constitute an important group of rock-forming silicate minerals that are closely related to each other in physical and chemical characteristics and in crystal structure. They occur in a wide variety of igneous rocks (originating from a molten condition) and metamorphic (altered) rocks, less commonly in sedimentary rocks, and rarely in meteorites. Common forms include long prismatic crystals, radiating columnar aggregates, silky fibres, or irregularly shaped grains that typically exhibit well-developed planes of prismatic cleavage (tendency to split parallel to certain crystallographic directions). These planes meet at angles of about 55° and 125° and yield cleavage fragments that are diamond-shaped in cross section.

By courtesy of Peter Robinson



Photomicrograph of a three-amphibole assemblage from the Partridge formation of Middle Ordovician age near Quabbin Reservoir, Massachusetts. The light brown grain filling the centre of the micrograph is anthophyllite exhibiting incipient prismatic cleavage characteristic of amphiboles. The darker brown grain, which is an oriented intergrowth enclosed by anthophyllite, is hornblende showing cleavage and exsolution lamellae of $P2_1/m$ cummingtonite. Subjacent to the hornblende is a smaller brownish grain of $P2_1/m$ cummingtonite with fine exsolution lamellae of hornblende.

Amphiboles may be white, gray, green, brown, black, blue, or violet; iron-poor varieties, however, can be colourless or nearly so. They exhibit a hardness on the Mohs scale (on which talc equals 1 and diamond 10) between 5 and 6.5, yield water when heated in a closed tube, fuse in a flame with difficulty, and dissolve slowly in hydrofluoric acid but are virtually unaffected by other acids. The optical property of pleochroism (colour absorption varying with direction), which sometimes aids in mineral identification, usually is distinct, increasing in intensity with increasing transition metal ion (charged atom) content; it may be weak or absent, however, in amphiboles lacking these ions.

Economic importance

Few amphibole minerals are of economic importance. Fibrous anthophyllite, actinolite, and crocidolite (a variety of magnesioriebeckite–riebeckite) are used as substitutes for chrysotile asbestos. Their heat-resistant properties and tensile strength, however, are less than that of chrysotile, and they are not as suitable for spinning. The aggregate called jade, used extensively by Oriental and Mayan peoples for ornamental purposes, is of two types. The more highly prized type consists of the pyroxene mineral jadeite. The other, usually called nephrite jade, is an amphibole and consists of a yellowish or spinach green, translucent-to-opaque, compact, dense aggregate of finely interfelted tufts of actinolite fibres.

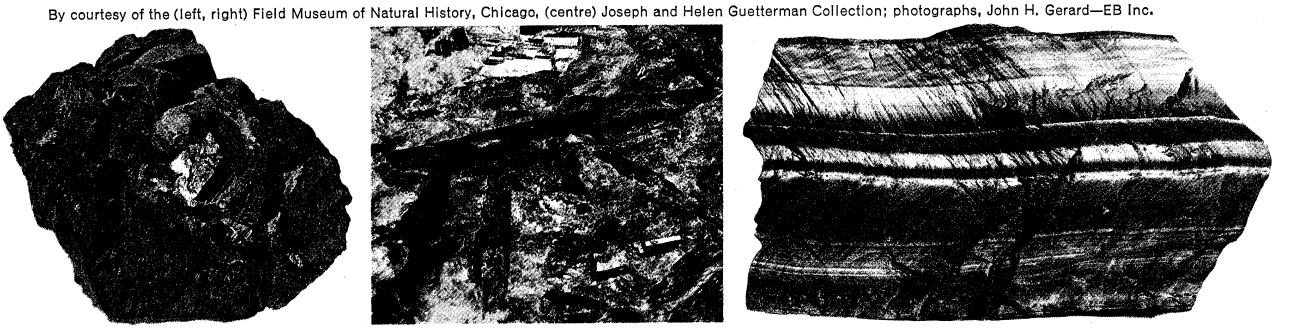
This article treats the chemical composition, crystal structure, and occurrence of the amphiboles in nature and covers experimental synthesis of this mineral group. For further information on details of crystal structure, see CRYSTALLOGRAPHY, and for coverage of an allied mineral group, see PYROXENES. For the relation of amphiboles to other rock-forming minerals and to minerals in general, see SILICATE MINERALS and MINERALS, respectively.

Chemical composition. The chemistry of the amphibole group is complex and variable because of the extensive ionic substitution that may take place between certain end-members (a solid-solution series, which is a single crystalline phase that varies between finite compositional limits, is said to have end-members for those limits). A typical amphibole often contains ten or more of the most abundant elements in the Earth's crust. A generalized structural formula that embraces these many substitutions is:

$$W_{0-1}X_2Y_5(Z_4O_{11})_2(OH,F,O,Cl)_2,$$

in which W refers to ions of radius one Å (one angstrom unit [Å] equals 10^{-8} centimetre) and larger, chiefly potassium (K^+) and sodium (Na^+), which may or may not be present; X refers to ions of radius 0.7 to 1.1 angstroms, chiefly calcium (Ca^{2+}), sodium (Na^+), ferrous iron (Fe^{2+}), magnesium (Mg^{2+}), manganese (Mn^{2+}), and lithium (Li^+); Y refers to ions of radius 0.5 to 0.9 angstrom, chiefly magnesium (Mg^{2+}), ferrous iron (Fe^{2+}), ferric iron (Fe^{3+}), aluminum (Al^{3+}), manganese (Mn^{2+}), and titanium (Ti^{4+}); and Z refers to small, highly charged cations, chiefly silicon (Si^{4+}) but to a lesser extent aluminum (Al^{3+}) and ferric iron (Fe^{3+}), and rarely beryllium (Be^{2+}). The type and extent of ionic substitution is determined, in general, by the radius of the ion, by the requirement for electrostatic neutrality (*i.e.*, that the positive and negative charges balance) in the final formula unit, and by the chemical environment and pressure-temperature conditions that existed during formation of the containing rock. Amphibole minerals can be conveniently classified chemically into one of three groups: magnesium–iron-, sodic-, or calcic-amphibole, depending on whether Mg-Fe, Na, or Ca is the predominating X cation. Selected end-member

By courtesy of the (left, right) Field Museum of Natural History, Chicago, (centre) Joseph and Helen Guetterman Collection; photographs, John H. Gerard—EB Inc.



Three important members of the amphibole group.
(Left) Hornblende from Chester, Massachusetts. (Centre) Actinolite crystals with talc from Washington. (Right) Crocidolite (tigereye) from Orange River, South Africa.

Selected (Mg, Fe)-, Na-, and Ca-Amphibole End-Members with Characterizing Chemical and Physical Properties

amphibole	W_{0-1}	X_2	Y_5	Z_8O_{22}	$(OH,F,O)_2$	specific gravity	refractive indices			crystal system	space group symmetry	unit cell parameters			
							α	β	γ			a	b	c	β
(Mg, Fe)-															
Anthophyllite	0	Mg_2	Mg_5	Si_8	$(OH)_2$	2.95	1.587	1.602	1.613	orthorhombic	Pnma	18.61	18.01	5.24	90°
Gedrite	0	Mg_2	Mg_3Al_2	Si_6Al_2	$(OH)_2$	3.18	1.642	1.655	1.661	orthorhombic	Pnma	18.53	17.74	5.25	90°
Protoamphibole	0	$(Mg, Li)_2$	Mg_5	Si_8	F_2	2.92	1.575	1.587	1.593	orthorhombic	Pmnm	9.33	17.88	5.29	90°
Cummingtonite	0	Fe_2	Mg_5	Si_8	$(OH)_2$	3.10	1.643	1.650	1.663	monoclinic	C2/m	9.30	18.14	5.19	102.1°
Grunerite	0	Fe	Fe_5	Si_8	$(OH)_2$	3.54	1.688	1.711	1.731	monoclinic	C2/m	9.59	18.43	5.34	101.9°
$P2_1/m$ clino-amphibole	0	$(Fe, Mn)_2$	$(Mg, Fe)_5$	Si_8	$(OH)_2$		1.621	1.632	1.643	monoclinic	$P2_1/m$	9.55	18.01	5.30	102.6°
Na-															
Glaucophane II	0	Na_2	Mg_3Al_2	Si_8	$(OH)_2$	3.03	1.594	1.612	1.618	monoclinic	C2/m	9.64	17.73	5.28	103.6°
Magnesioriebeckite	0	Na_2	$Mg_3Fe_2^{3+}$	Si_8	$(OH)_2$	3.15	1.656	1.671	1.672	monoclinic	C2/m	9.73	17.95	5.30	103.3°
Riebeckite	0	Na_2	$Fe_3^{2+}Fe_2^{3+}$	Si_8	$(OH)_2$	3.40	1.702	1.712	1.719	monoclinic	C2/m	9.73	18.06	5.33	103.3°
Ca-															
Richterite	Na	NaCa	Mg_5	Si_8	$(OH)_2$	3.0				monoclinic	C2/m	9.89	17.96	5.26	104.3°
Tremolite	0	Ca_2	Mg_5	Si_8	$(OH)_2$	2.97	1.600	1.614	1.627	monoclinic	C2/m	9.83	18.05	5.27	104.5°
Ferrotremolite	0	Ca_2	Fe_5	Si_8	$(OH)_2$	3.40	1.672	1.686	1.693	monoclinic	C2/m	9.97	18.34	5.30	104.5°
Pargasite*	Na	Ca_2	Mg_4Al	Si_6Al_2	$(OH)_2$	3.0	1.624		1.645	monoclinic	C2/m	9.91	17.99	5.26	105.3°
Ferropargasite*	Na	Ca_2	$Fe_4^{2+}Fe^{3+}$	Si_6Al_2	$(OH)_2$	3.4	1.700		1.718	monoclinic	C2/m	9.95	18.14	5.33	105.3°
Hastingsite*	Na	Ca_2	$Fe_4^{2+}Fe^{3+}$	Si_6Al_2	$(OH)_2$	3.5	1.702		1.728	monoclinic	C2/m	9.98	18.15	5.32	105.2°
Joesmithite	(Ca, Pb)	Ca_2	$(Mg, Fe^{2+}, Fe^{3+})_1$	Si_6Be_2	$(OH)^2$	3.83	1.747	1.765	1.78	monoclinic	P2/a	9.88	17.88	5.24	105.7°

*Hornblende.

representatives of each group are enumerated in the Table together with some characteristic physical properties.

Crystal structure. The fundamental structural unit in almost all silicate minerals is the SiO_4 tetrahedron consisting of four oxygen atoms that surround and are bonded to a Si-atom at the corners of a tetrahedron. The structure types of the different silicates can be related to the various ways the tetrahedra are linked together into larger units. In pyroxene, SiO_4 tetrahedra are linked in infinite single chains of SiO_3 composition whereas the amphiboles are made up of double chains, in which the tetrahedra alternately share two and three oxygen atoms; this arrangement leads to an oxygen/silicon ratio of 11 to 4 (see Figure 1). The crystal structure of amphibole is characterized by layers of interlocking (Z_4O_{11}) infinite double chains (Figure 2) consisting of tetrahedra linked together into arrays of O(3)-centred six-membered rings, O(3) representing hydroxyl, fluorine, oxygen, or (rarely) chlorine.

Amphibole chains and layers

In Figure 2, A and B represent idealized drawings of the structure of a clinoamphibole viewed down c (commonly the vertical or upright crystallographic axis) and a^* [a^* is perpendicular to (100)—a plane that intersects the a-axis at unit distance and is parallel to the other two crystallographic axes], respectively, together with the nomenclature of the atoms in the asymmetric unit. The chains in these drawings parallel (100) with their endless dimensions lying along c. Along the b-axis the chains are juxtaposed and alternate in orientation, with the apices of the ZO_4-tetrahedra pointing successively up and down along a^*. The O(4) oxygen anions of each chain dovetail

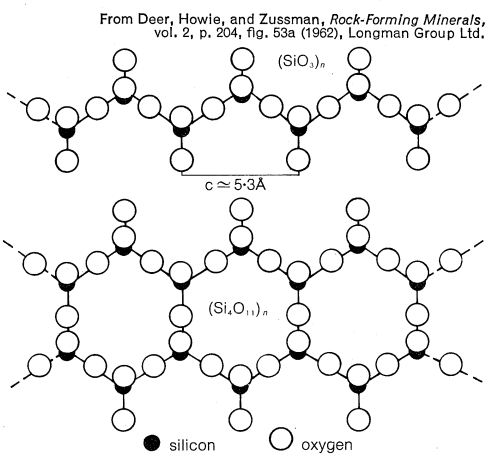


Figure 1: Comparison of amphibole band $(Si_4O_{11})_n$ and pyroxene chain $(SiO_3)_n$ (see text).

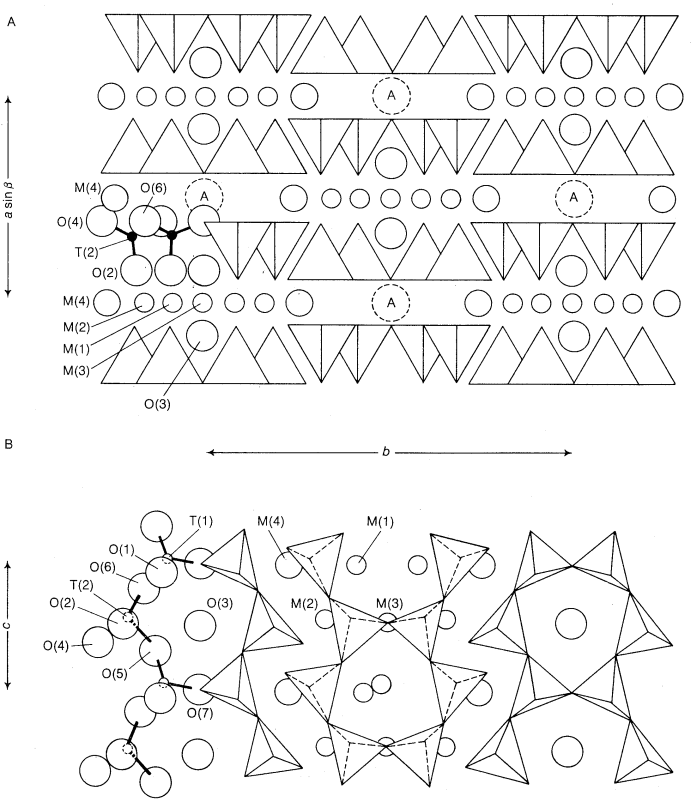


Figure 2: The crystal structure of C2/m clinoamphibole viewed (A) along the c-axis and (B) down the a-axis; T(1) and T(2) represent substitution sites (see text).

with the O(1) and O(2) oxygen anions of adjacent chains, locking them together into layers. The layers, stacked parallel to (100), are bonded together by X and Y cations that are coordinated (surrounded by and bonded to) between narrow hexagonal strips consisting of O(1), O(2), O(3), and O(4), the coordination being affected by a stagger of about $c/3$ between adjacent layers. The Y cations occur in more or less perfect octahedra designated M(1), M(2), and M(3), whereas the X cations are in irregular sixfold to eightfold coordination polyhedra designated M(4). When present, the W cations are distributed about the A-site, which is located in the relatively large cavity formed by eight oxygen anions at the backs of adjacent chains. The good to perfect prismatic cleavage, diagnostic of the mineral, passes through the structure parallel to these chains by stepping successively along their backs

through the *A*-site and thence to periphery of the next layer. The crystal structures and physical properties of many of the amphiboles are remarkably similar to those of the clinoamphibole shown. All consist of layers of interlocking double chains of ZO_4-tetrahedra bonded together by *X* and *Y* cations. Different structural types arise, however, because of different layered stacking sequences, the occurrence of chains of two slightly different geometries, or variations of cation ordering and composition.

Magnesium–iron amphiboles

Occurrence and distribution. The magnesium–iron amphiboles are for the most part confined to metamorphic rocks and crystallize in both the orthorhombic and monoclinic systems. The orthorhombic varieties are represented in nature by the anthophyllites, in which the Mg/Mg + Fe ratio may reach 0.4, and by the gedrites (essentially aluminous anthophyllites), in which aluminum may substitute for magnesium or iron in *Y* and for silicon in *Z*. The gedrites, however, can be quite iron-rich. Magnesium-rich anthophyllites are commonly found in contact zones where granitic dikes have intruded ultramafic (iron- and magnesium-rich) rocks and in metamorphosed siliceous dolomites (calcium magnesium carbonate rocks). Iron-rich anthophyllites occur in ultramafic bodies such as peridotites, which have been partially hydrated at moderate temperatures. In the presence of alumina, as in rocks that contain such minerals as cordierite or plagioclase, the amphibole composition approaches that of a gedrite. Environments that are relatively low in sodium, potassium, calcium, and manganese, however, are essential for crystallization of anthophyllite or gedrite alone.

The monoclinic varieties, called cummingtonites, commonly seem to require an Fe/Fe + Mg ratio of 0.4 or more. If this ratio is close to 1.0, the amphibole is called grunerite, a form that is stable only at the lower oxidation states. Manganese is normally the third most abundant cation in *X* and *Y*, whereas aluminum, sodium, and calcium are characteristically low. Cummingtonites are commonly found in iron formations, such as those in Labrador and Minnesota, where they are associated with quartz, iron oxides, carbonates, and iron-rich sheet silicates. The existence of a monoclinic form not commonly reported ($P2_1/m$) that can be changed by heating at 40° C (104° F) and one atmosphere pressure into the more commonly reported monoclinic form (C2/m) has been demonstrated. Assemblages that include orthorhombic and monoclinic amphiboles have been reported in zones between metamorphosed iron-rich volcanic rocks and clay-rich sediments. Nevertheless, because many natural environments contain significant amounts of sodium, potassium, and calcium, the magnesium–iron amphiboles are the least abundant of the three amphibole groups.

Sodic amphiboles

Sodic amphiboles are typified by the minerals glaucophane, magnesioriebeckite, and riebeckite (see the Table). Figure 3 illustrates their chemical variation in nature and the control of rock bulk composition on their chemistry. The ferroglaucophane corner in Figure 3 represents a composition that does not appear to be common.

Glaucophane and magnesioriebeckite occur principally in low-grade (altered at relatively low temperatures) metamorphic rocks. The association of glaucophane with lawsonite, $CaAl_2Si_2O_8 \cdot 2H_2O$, and/or one of the carbonate minerals (*e.g.*, calcite or aragonite, polymorphs of $CaCO_3$), constitutes one of the characteristic assemblages of the so-called blueschist metamorphic facies (rock types). Such rocks represent relatively high pressures (five to eight kilobars) and low temperatures of recrystallization (100° to 300° C; 210° to 570° F). Glaucophane is most abundant in rocks of basaltic bulk composition in this facies. In environments high in aluminum, magnesium, iron, and alkalis, solid solutions toward magnesioriebeckite result. These solid solutions, intermediate members of which are called crossite, may also occur at the low temperatures but less elevated pressures of the greenschist metamorphic facies. In most greenschists, however, the amphiboles are calcic varieties, so that the presence of sodic forms represents departures from basaltic composition toward high soda and low lime.

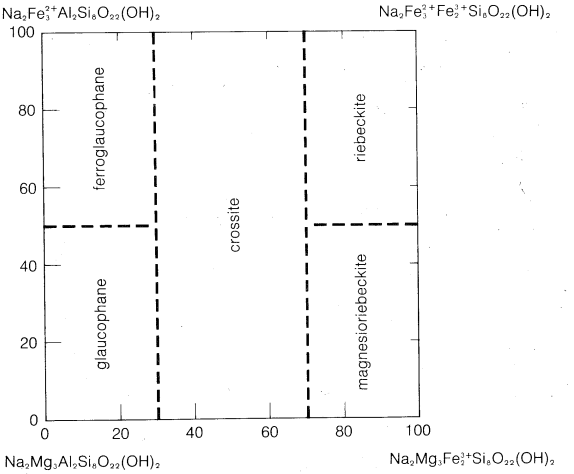


Figure 3: Variation in the chemical composition and nomenclature of the glaucophane–riebeckite amphiboles.
From Deer, Howie, and Zussman, *Rock-Forming Minerals*, vol. 2, p. 334, fig. 82 (1962); Longman Group Ltd.

Riebeckites have low thermal stability and are most common in the lowest temperature, last-to-crystallize, silica-rich igneous rocks, such as pegmatites. They also form in the greenschist and blueschist facies in recrystallized iron-rich cherts. The dark green to black amphibole that occurs in some granites and syenites, though often called riebeckite, is probably more nearly an arfvedsonite; it exhibits a higher soda content and higher Fe^{2+}/Fe^{3+} ratio than riebeckite. Authigenic (formed *in situ*) riebeckite–magnesioriebeckite has been reported from the Green River Shales and evaporites in Utah. The temperature of growth must have been in the range of 25° to 50° C (75° to 120° F), and, accordingly, these sodic forms may be stable at lower temperatures than any other members of the amphibole group. In contrast, there are no documented reports of magnesium–iron or calcium varieties forming naturally at temperatures as low as even 100° C. These latter species weather and decompose to clays and oxides, indicating that their lower temperature of stability lies above that of the sedimentary environment.

Calcic amphiboles

Calcic varieties are the most widespread and volumetrically abundant members of the amphibole group. For purposes of discussion, two subgroups may be distinguished: (1) the tremolite–ferrotremolite series (intermediate members being called actinolites), which contain little or no aluminum and lack *W* cations (vacant *A*-site), and (2) the hornblendes, which include the majority of other calcium amphiboles.

Tremolite is a common mineral in metamorphosed siliceous dolomites and limestones and in ultramafic rocks metamorphosed at low grade. It can be found in high-temperature carbonate-free rocks (sillimanite zone) but forms at much lower temperature in the presence of carbon dioxide, which is the common case. Thus, its occurrence normally implies only a low-grade metamorphic environment. Actinolites form in rocks of basaltic composition because such rocks contain more iron than carbonates. Actinolites are a common and diagnostic feature of greenschist facies conditions. The change from low-grade to high-grade metamorphic conditions is accompanied by a concomitant change in calcium-amphibole composition from actinolite to hornblende.

Hornblendes have a wide compositional range and are chemically the most complex of the natural amphiboles. Entire metamorphic terranes are composed chiefly of hornblende plus plagioclase feldspar (a rock called amphibolite). Higher temperature conditions of origin are commonly reflected by increased aluminum, titanium, sodium, and potassium content. The maximum amount of substitution of aluminum for silicon in *Z* is about 25 percent (two of the eight *Z* cations in the formula unit) with the aluminum ordering into the T(1) tetrahedra of the chain (Figure 2). The pargasite–hastingsites are hornblende end-members illustrative of the maximum substitution of aluminum for silicon.

Hornblendes have sufficient thermal stabilities to be stable in magmas; thus they are common and very abundant in such plutonic igneous rocks as diorites, quartz diorites, and granodiorites. Hornblendes also occur as phenocrysts (large single crystals) in andesitic lavas but normally are much less abundant in volcanic rocks because the volcanic conditions of relatively low water pressure and high temperature are outside amphibole stability. The iron-rich hornblende hastingsite is found in many granites, syenites, and certain other alkali-rich rocks.

Calcium amphiboles are found as detrital grains (those derived from pre-existing rocks and then transported) in heavy mineral (*i.e.*, of specific gravity greater than that of quartz [2.65] and other relatively light minerals) suites (sets) of sedimentary rocks. Actinolite seems to be less abundant than hornblende. In general, amphiboles represent a decreasing percentage of the suite with increasing age because of a relatively low resistance to weathering and post-depositional changes.

Synthesis of amphiboles

Experimental studies. Experimental studies of the stability and phase relations of amphiboles date back only to the 1950s. Since that time a number of the end-member compositions have been synthesized successfully. The decomposition of amphiboles at a given pressure with increasing temperature gives rise to assemblages that may include the phases pyroxene, quartz, olivine, plagioclase, garnet, oxides, liquid, and/or vapour. The chemically simpler amphiboles, such as anthophyllite or tremolite, may be represented by an assemblage of equivalent bulk composition consisting of only two or three minerals. The chemically more complex amphiboles, however, represented by hornblende, may either decompose or melt incongruently (react with the liquid phase) to as many as five other minerals. Thus, many rock compositions, if they could be recrystallized in the presence of water at 300° to 800° C (600° to 1,500° F) and pressures of 500 to 5,000 bars, would consist predominantly of amphibole.

Adequate experimental work has been completed to allow generalizations on the effect of compositional changes on amphibole thermal stability. One of the most dramatic increases in high temperature stability of amphibole comes about if magnesium is substituted for ferrous iron (200°–400° C [400°–750° F]). Equally important is the substitution of fluorine for the hydroxyl (OH) ion. Some synthetic fluoramphiboles appear to be stable up to 1,100°–1,200° C (2,000°–2,200° F) at one atmosphere pressure, which is 600°–800° C (1,100°–1,470° F) higher than for species containing only OH. Whether such relations hold at elevated pressures has yet to be determined. Substitution of ferric iron for aluminum and the introduction of sodium into the *A*-site lead to increases in stability of the order of tens of degrees Celsius only. If aluminum is introduced into *Y* and *Z* along with sodium in *W*, as in the change from tremolite to pargasite, thermal stability increases as much as 200° C (400° F).

For any amphibole containing elements of variable oxidation state (*e.g.*, iron, manganese, titanium), the oxygen fugacity (the thermodynamic equivalent of partial pressure) will strongly modify the stability relations. With increasing oxidation states, the upper temperature stability limit of ferropargasite, for example, can be reduced by more than 300° C (575° F). Moreover, under conditions oxidizing enough for the stable formation of hematite (Fe_2O_3), ferropargasite and ferrotremolite are unstable.

The pure iron end-members are rarely realized in nature. Experimental studies indicate, however, that this is not caused by the instability of such structures. The principal causes are: (1) the iron members have much more limited thermal ranges compared to their magnesium counterparts, (2) the oxygen fugacity–temperature conditions necessary for the formation of the iron members occur rarely in nature, and (3) rock bulk compositions normally are not sufficiently iron-rich to permit the formation of such amphiboles.

Geological evidence suggests that the oceans and atmosphere of the Earth developed with time through emission or release of gases from the mantle (the zone beneath the crust) during volcanism. Accordingly, interest has centred on amphiboles as possible compounds containing water in the mantle. Work shows that these phases have lowered thermal stabilities at fluid pressures of 15 to 30 kilobars compared to pressures up to 15 kilobars. Nevertheless, the stable existence of amphiboles to depths of 50 to 70 kilometres (30 to 45 miles) in the Earth is not precluded. The storage of water in amphiboles under upper mantle conditions is thus a very real possibility (see EARTH, STRUCTURE AND COMPOSITION OF).

Amphiboles in the Earth's mantle

BIBLIOGRAPHY. W.G. ERNST, *Amphiboles* (1968), a good review of the crystal chemistry, experimentally determined phase relations, and occurrence of amphiboles—cell parameters and indexes of refraction of a number of synthetic end-member amphiboles are included; B.E. LEAKE, *A Catalogue of Analyzed Calciferous and Subcalciferous Amphiboles Together with Their Nomenclature and Associated Minerals* (1968), a compilation of 1,200 post-1890 wet chemical (gravimetric) analyses of natural amphiboles that contain 6 weight percent or more CaO—also includes a brief discussion of occurrence and chemistry; W.A. DEER, R.A. HOWIE, and J. ZUSSMAN, *Rock-Forming Minerals*, vol. 2, *Chain Silicates* (1962), a compendium of mineralogic data for amphiboles useful in petrology; C. KLEIN, JR., "Coexisting Amphiboles," *J. Petrology*, 9:281–330 (1968), presents extensive new compositional data on coexisting amphibole pairs that help to define the range of miscibility between (Mg,Fe)-and Ca-, and (Mg,Fe)- and Na- amphibole groups; J.J. PAPIKE *et al.*, *Pyroxenes and Amphiboles: Crystal Chemistry and Phase Petrology* (1969), contains studies by X-ray diffraction, optical, electron paramagnetic resonance, Mössbauer, electron microprobe, and synthesis techniques of the amphiboles; E.J.W. WHITTAKER, "The Crystal Chemistry of the Amphiboles," *Acta Crystallogr.*, 13:291–298 (1960), one of the first modern discussions of amphibole crystal chemistry.

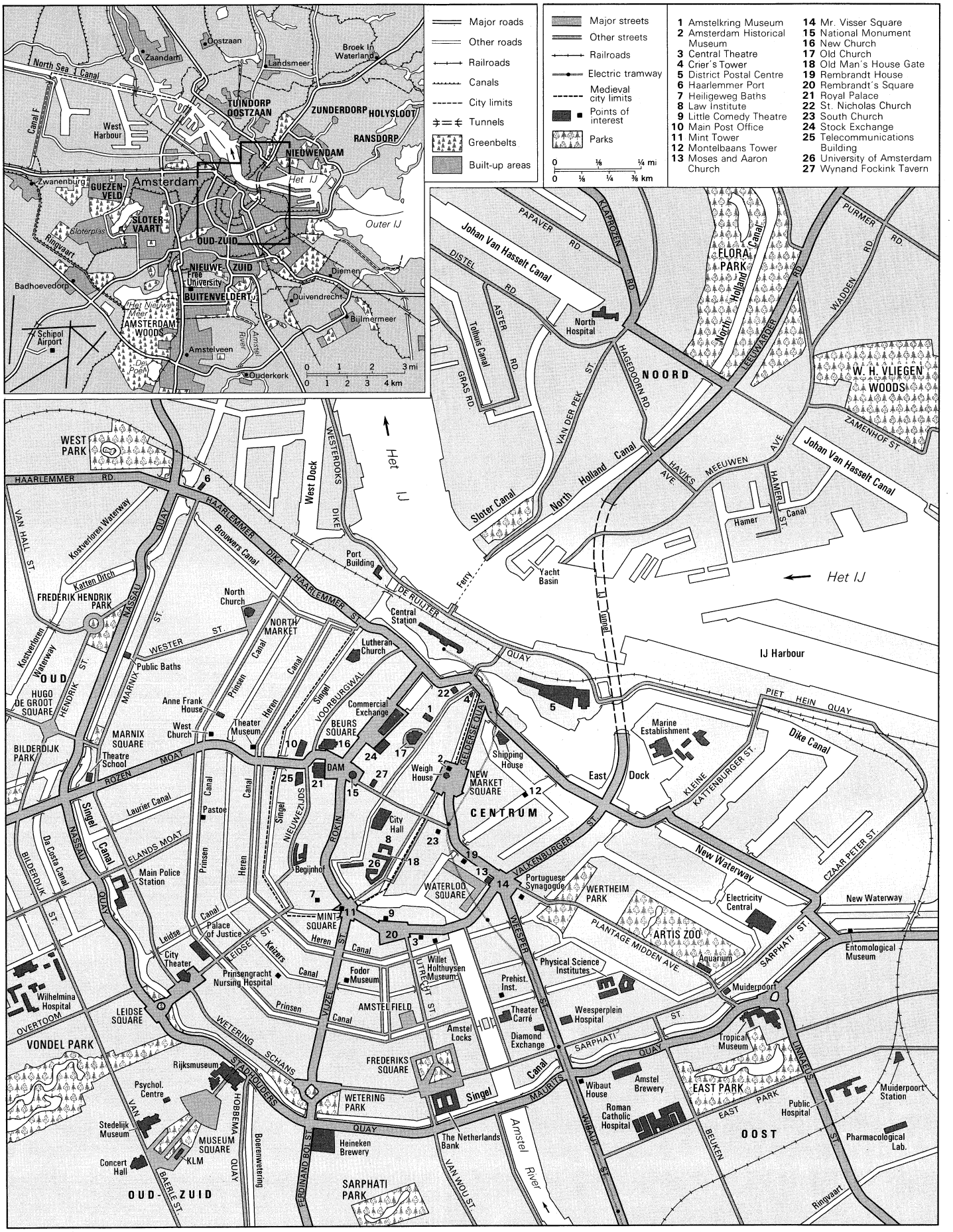
(G.V.G./M.C.G.)

Amsterdam

In talking of Amsterdam as "a city rich in history, outstanding in art and colourful in appearance," John J. Murray, a U.S. writer summarizing his impressions when he visited the city for the first time, undoubtedly used the right words. But he could have added: an overcrowded metropolis with the modern big-city problems of air pollution, traffic congestion, and housing shortages. It is easy to describe it as a romantic city and to praise the eternal beauty of the centuries-old canals, the ancient patrician houses, and the atmosphere of freedom and tolerance, but Amsterdam is also a modern city that has yet to work out its own solutions to today's problems.

Amsterdam, the home of almost 810,000 people in the early 1970s, is the nominal capital of The Netherlands but not the seat of the government, which is at The Hague ('s-Gravenhage). Queen Juliana, for example, lives mainly at Soestdijk, only visiting the Royal Palace on the Dam square occasionally. Amsterdam lacks the monumental architecture found in other capitals. There are no wide squares suitable for big parades, and there are no triumphal arches or imposing statues. The secret of Amsterdam's greatness is to be found in small, narrow streets where ordinary people live, occupied with their own day-to-day problems. There are many reminders of the glorious past—the gabled houses, the almshouses, the noble brick facades covered by sandstone, the richly decorated cornices, the towers and churches, and the music of carillons and barrel organs—but today's city is supported by other and harder facts that leave little room for romantic dreams. It is a city of 90 islands connected by about 600 bridges and is as well the main wholesale and industrial centre of The Netherlands and its principal retail centre. It is a city in which 500,000 bicyclists risk their lives each day, one that in the early 1970s had two universities and 25,000 students, over 100 Chinese restaurants, 1,000 houseboats, and 1,300,000 tourists annually.

Rich in history, outstanding in art, and colourful in appearance, yes—but the colour is no longer Rembrandt's mysterious glow but rather the soft blue light of nearly 200,000 television sets in as many cozy living rooms. (For related information, see NETHERLANDS, THE.)



Central Amsterdam and (inset) its metropolitan area.

History. Though modern historians do not exclude the possibility that during the Roman domination some form of settlement was developed at the mouth of the Amstel River, evidence of such has never been found. So far as is known, Amsterdam originated as a small fishing village in the 13th century AD. To protect themselves from floods, the early inhabitants had to build dikes on both sides of the river, and in about 1270 they built a dam between these dikes, the Amstel-dam.

Even then, merchant ships from Amsterdam sailed as far as the Baltic Sea and laid the foundation of the future trade centre, acting as a link between northern countries and Flanders (roughly modern Belgium). The city was under the jurisdiction of the counts of Holland, and one of them, Count Floris V, granted the *homines manentes apud Amestelledamme* ("the people living near the Amsteldam") a toll privilege in 1275. In this document the name Amsterdam is mentioned for the first time. The city rapidly extended its business, and in 1489, as a sign of gratitude for the support given by the city to the Burgundian-Austrian monarchs, Emperor Maximilian I allowed Amsterdam to adorn its armorial bearings with the imperial crown. By then Holland's greatest commercial town and the granary of the northern Netherlands, Amsterdam was a centre of wealth and influence in Europe.

In the 16th century there was religious and political resistance in the Netherlands against Spanish oppression. Amsterdam hesitated to accept the leadership of William I the Silent, prince of Orange, but in 1578 there was a bloodless revolution. The magistrates, together with the majority of the Roman Catholic priests, were deported; the religious communities were secularized; and the old Catholic Church underwent reform.

Amsterdam was still a small town with no more than about 30,000 inhabitants, but things changed quickly, especially when, in 1585, Antwerp (in modern Belgium) was conquered by Spanish troops, and the Scheldt (Schelde) River was closed. Much of the trade concentrated in this port then moved to Amsterdam, and with the Flemish merchantmen soon came hundreds of Jews expelled from Portugal, followed by their co-religionists from the area of modern Germany and eastern Europe. The city soon became a trading metropolis: merchant ships from Amsterdam not only sailed to the Baltic and the Mediterranean but also found the long seaway to the East Indies and established colonies in South America and South Africa. This complex of activities during the Netherlands' long struggle for independence brought prosperity to shipbuilding and other industries, including banking, communications, the arts, and science.

Development as a trading centre

At this time, the still externally medieval town developed into a big city, and in 1612 the City Council decided upon a new extension—the Three Canals Plan. Furthermore, the city needed a new and stately town hall, and an architect, Jacob van Campen, built one in the Dam square in the shadow of the Nieuwe Kerk (New Church). In 1632 the Athenaeum Illustre (to become the University of Amsterdam in the 19th century) was erected, and, when, in 1648, the Treaty of Münster ended the Eighty Years' War (1568–1648) with Spain, Amsterdam was the financial, trading, and cultural centre of the world, lending money to foreign kings and emperors and thus influencing the politics of many countries.

It was then inevitable that the City Council would come into conflict with the political centre of the republic, because the country was no longer ruled by the States General in The Hague but by a few burgomasters and merchants in Amsterdam. This situation led to political difficulties with William II, prince of Orange, who even planned to besiege the city. Amsterdam, nevertheless, maintained its position for many years. Decline gradually came in the 18th century: London and Hamburg surpassed Amsterdam as trade centres, but the city remained the financial heart of Europe. Serious conflict arose when the new political ideas that led to the French Revolution found acceptance. In 1787 Amsterdam was occupied by the Prussians who backed the policy of William V, prince of Orange. The French, welcomed as liberators in 1795, brought freedom, but within a few years trade and shipping nearly stopped because of Napoleon's constricting trade policies. In 1806 Napoleon proclaimed the Netherlands a kingdom, with Amsterdam as its capital, but already by 1810 the country was incorporated in the French Empire. The Russian Cossacks entered the town in 1813, and, on March 30, 1814, William VI, prince of Orange, was inaugurated as William I, king of The Netherlands in Amsterdam's Nieuwe Kerk.

The Napoleonic era

Several attempts to restore prosperity to the city failed because of the harbour's disastrous condition. Economic rehabilitation came only after the digging of the North Sea Canal in the years 1865–76. The population expanded rapidly (500,000 by 1900), with the East Indian trade remaining the backbone of its economic activities. A sudden end came in 1940, however, when the German Army occupied the country. Allied bombers attacked industrial areas several times during World War II, but the city's most severe loss was the deportation of about 70,000 Jewish inhabitants. In May 1945 Amsterdam was liberated by Canadian troops, and, on March 29, 1946, Queen Wilhelmina granted the town the right to add to its armorial bearings the device "Heroic, Resolute, Compassionate," in recognition of the citizens' attitude during the German occupation.

The contemporary city. Amsterdam is situated at the mouth and on the south side of the IJ, an inland arm of the former Zuiderzee, now IJsselmeer (so called since 1932, when a dam separated it from the sea), and connected by canal with the North Sea. It is divided by the canalized Amstel River into two main sections.

The medieval town lies on either side of the Amstel at the city's centre, enclosed by the semicircular Singel gracht. Outside the latter are the three main canals (*grachten*), dating from the 17th century; the Herengracht, Keizersgracht, and Prinsengracht. Within this area three smaller canals run north–south. One tower (the Schreierstoren) of the old fortifications still stands, although it is now utilized by the port authority.

The old part of Amsterdam has many ancient buildings, among which the Oude Kerk (Old Church), built in the 13th century, the Nieuwe Kerk (New Church, 1408, rebuilt 1645) are outstanding. Next to the Nieuwe Kerk is the 17th-century town hall, now Royal Palace, in classical Palladian style. Other notable buildings include the Munttoren (Mint Tower), with a 17th-century upper part superimposed on a medieval gate; the Zuiderkerk, or South Church (1611); the Westerkerk (1631), where Rembrandt is buried; the Trippenhuis, housing the Royal Netherlands Academy of Sciences; and the Oudemanhuispoort (Old Man's House Gate), now one of the municipal university's main buildings. The Sint Nicolaas Kerk (1886), the Central Station (1889), the Beurs (Commercial Exchange Building, 1903), and the Shipping House (1916) are more modern in style. The former Jewish quarter, in the east of the old town, boasts the Portuguese Synagogue (1670) and the Rembrandthuis (the home of the painter Rembrandt), now a museum. The old town's three main squares are the Dam, the Leidseplein, and the Rembrandtsplein. Fine 17th- and 18th-century patrician houses line the canals.

In the 17th century the Singelgracht enclosed the three canals. Along it now run three quays: the Nassaukade, Stadhouderskade, and Mauritskade. Buildings outside the Singelgracht rose from the late 19th century onward and include the Rijksmuseum (State Museum, 1876–85); the Concert Hall (home of the Concertgebouw Orchestra); the Municipal (Stedilijk) Museum (1895); the Olympic Stadium (1928); and the Amstel Station (1939).

Development plans. In 1928 the Amsterdam City Council decided to set up an organization to deal with town development as a separate part of the Public Works Department.

A scheme called the General Extension Plan, accepted in 1935 by the City Council, was realized after World War II when the urgent need for houses was an incentive to build as quickly as possible. Simultaneously, a further detail of the master plan was realized: a large forest covering an area of 2,250 acres (910 hectares) in the southwestern part of the city and situated more than 13 feet

The General Extension Plan



Sint Nicolaas Kerk, Amsterdam, with sightseeing boats in the harbour. The port is in the left background.
Louis Goldman—Rapho-Guillumette

(four metres) below sea level was turned into a summer recreation centre for over 100,000 people.

Soon after World War II certain developments made it necessary to revise the General Extension Plan: the residential areas were obviously inadequate to accommodate future inhabitants because of the lowered marriage age, and this required considerably more and larger dwellings than foreseen in 1935. Further space, moreover, was required for traffic and parking as well as for schools, service industries, and recreation, while the urban renewal of the old city, including the 19th-century residential areas, required much replacement space for dwellings and business in the outer areas.

So far, Amsterdam's town-planning problems differ little from those of any major city in the world, though the town still retains its 17th-century characteristics while combining the cultural and economic activities of the 20th century. The particular problem is that a city over 300 years old must integrate with the natural features of the area in which the town expansion is to take place. The streets are narrow, and the great concentration of employment produces heavy radial traffic flows. Public traffic carries about 25 percent of the passenger load and private cars only 10 percent; more than 60 percent of all the journeys to and from this central area were for many years made by the most popular means of Amsterdam transport, the bicycle. As distances grow and large radial highways from the outer areas to the city are not desired, the public-transport system needs to be radically renewed. Therefore, a metropolitan railway system has been developed, its total network to have a length of 55 miles (88 kilometres), of which 17 will be underground.

Town expansion since 1945 has resulted in the building of about 65,000 dwellings, but this has meant at the same time a doubling of the urban area. There still will remain a considerable housing shortage when all the plans, now under construction, are realized. The main problem for the years to come is the renewal of about 75,000 dwellings in the 19th-century area, which implies an annual clearance of about 1,900 dwellings in a 40-year period beginning in 1975. A very large number of these dwellings cannot be replaced in Amsterdam itself, and, therefore, in the planning an overflow is extended to other neighbouring centres, such as Purmerend and Hoorn, about ten and 20 miles northward, respectively. At the same time, new cities have to be constructed in the polders (land reclaimed from the sea) of the former Zuiderzee, east of Amsterdam. The General Extension Plan of 1935 came to be regarded as a modest beginning when it was realized in about 20 years, and the plan of the total renewal of the city will have to be executed in the period between 1975 and 2025.

Demography. In 1947 the urban district of Amsterdam covered an area of 14,100 acres (5,700 hectares), while the number of inhabitants was 800,000. Nearly 18 years later the area had expanded to 27,000 acres, but the population had only grown to 866,000; however, by 1972 the area had increased to 42,000 acres and the population had decreased to almost 810,000, a remarkable phenomenon not uncommon in most cities of western Europe.

Migration from the inner city

The population can be roughly divided into 390,000 men and 420,000 women: 116,000 old people (65 years and over), 240,000 male and 90,000 female workers, and 180,000 schoolchildren. The birth rate went down between 1946 and 1970, while the death rate increased over the same period. Population changes resulting from migration have shown negative figures since 1951 and have reached now an average loss of 10,000 per annum. The highest figures in the statistics of causes of death are those of heart disease and malignant neoplasms (morbid growths). There is an average of 22 marriages and four divorces daily. About 23,000 foreign workers are temporarily resident in Amsterdam because of the city's chronic manpower shortage.

Economic life. The importance of Amsterdam as an economic centre can be demonstrated by a few statistics. Roughly 6.1 percent of the population of The Netherlands live in the city, whereas 9.2 percent of the national production is from Amsterdam. Of the total amount spent on wages and salaries in The Netherlands, 9.7 percent is paid out in Amsterdam, and the imports and exports of goods and services transacted there comprise 7.25 percent of the national figure.

The origins of most of these activities can be traced back several hundred years. The Amsterdam Exchange Bank was founded in 1609 in order to correct the chaotic mon-

etary affairs of the day. It soon developed into the largest clearinghouse in Europe and made the city an international financial centre. A few years earlier, in 1602, the Dutch East India Company was founded, and the lively trade in shares laid the basis for modern dealings in stocks. For this reason the Amsterdam Stock Exchange can be considered the oldest in the world. Nowadays more foreign securities are quoted there than in any other exchange in Europe, and Amsterdam, the home exchange for leading Dutch securities such as Royal Dutch/Shell, Philips Electric Company, and Unilever, occupies an honourable place after New York, London, and Zürich. The main banks and insurance companies in The Netherlands have their roots in Amsterdam; about 23 percent of all coverage in The Netherlands is taken out in Amsterdam, while 35.6 percent of the production value of the Dutch banking establishments is effected here.

In wholesale trade Amsterdam has played an important role since the 14th century. In the past the port was mainly a harbour for tropical products (coffee, tobacco, tea, and rubber), but, after the waterway connections with the German hinterland were improved following World War II and the ties with the former Indonesian colonies cut loose in 1949, the port developed into a transit port for bulk cargo (ore, coal, cereals, timber, and oil). By 1968 the harbour mouth at IJmuiden (on the coast some 15 miles west-northwest of Amsterdam) had been thoroughly modernized, and the port gave access to ships weighing up to 85,000 tons. Each year some 8,000 ships, carrying about 23,000,000 tons of goods, berth at the well-equipped quays. A study was being made in the early 1970s for the construction of an outer port immediately south of IJmuiden to meet the further demands for accommodation of ocean carriers larger than 100,000 tons. In addition to this, about 4,000 acres of deepwater-fronted industrial sites were ready for occupation by modern process industries. The port provides a living for more than one-third of the Amsterdam population.

Changing role of the port

In civil aviation Amsterdam has occupied a prominent place right from the start, thanks to the proximity of Schipol Airport and KLM Royal Dutch Airlines' large share of international air transport. In 1967 an entirely new airport was put into operation, and Schipol now ranks sixth in European airports with regard to passenger traffic (5,200,000 per annum) and fourth in the field of freight traffic. The distance from Schipol to the city is only five miles.

Amsterdam, moreover, is the largest industrial city in The Netherlands, with over 7,700 industrial establishments employing more than 112,000 men and women, who produce an annual income of 5,000,000,000 guilders ($1,400,000,000). The city accounts for 13.4 percent of the production value of the printing and publishing business, 8 percent of metal goods and mechanical engineering, and 11 percent of the conveyance industry. A special attraction for tourists is the diamond industry. The art of diamond cutting was carried to Amsterdam in the 1570s, when refugees from Antwerp settled here. Two thousand diamond workers of Jewish origin were deported and murdered during World War II by the Germans, and large quantities of precious stones disappeared during the occupation. With foreign assistance and government support, however, the industry was rebuilt, and now there are again 20 diamond factories, and 80 percent (in weight) of the total world production of industrial diamonds is produced here.

Seven percent of the municipal income comes from the tourist trade: 2,400,000 overnight stays (1970) demonstrate the importance of this source of revenue. There are over 200 hotels, some 15 youth hostels, several camping sites, more than 400 boardinghouses, and dozens of Dutch and foreign restaurants. The main attraction for tourists is the Rijksmuseum, with 1,400,000 visitors annually, followed by Artis Zoo (630,000 visitors) and the canal trips (2,000,000 passengers). The city is an important centre for congresses and exhibitions. There are 12,500 retail businesses, employing 40,000 people, while 40 percent of the population lives within a 35-mile radius of Amsterdam and do their shopping in the city centre.

Political and governmental institutions. According to the constitution, every municipality in The Netherlands is headed by a council, the size of which depends on the number of inhabitants. The Amsterdam City Council has 45 members, who are chosen for a four-year period. The municipal executive consists of the burgomaster and eight aldermen. The former is not an elected official: he is appointed by the crown for a period of six years, and, though he presides at the meetings of the Council, he is not a member of this institution. The Council has no say in his appointment, but he usually represents the largest political group in the Council. For a quarter of a century after 1946, the Amsterdam burgomaster was always a member of the Socialist Party.

Methods of political appointment

The aldermen are chosen by and from the members of the Council. They, too, have a four-year period of office and receive an income, in contrast to the Council members, who get an attendance fee.

As a result of the many and very different political parties, the City Council represents more than ten groups, divided in 1972, for example, as follows: Socialists, 12; Communists, eight; Liberals, six; Roman Catholics, five; Democrats, three; Christians, three; a radical youth group calling themselves Gnomes, four; Pacifists, one; and three independent members. There were three Socialist, two Communist, two Catholic, and one Christian aldermen.

As a rule, the Council meetings are open to the public. The meeting place is at the town hall—the Prinsenhof, formerly the Admiralty Court. In the 15th century this was a convent, but after alteration in 1578 it was converted into a suitable "inn for Princes and High Personages." From 1655 to 1808, the seat of the Council was at the Dam square, where the medieval town hall was replaced by Jacob van Campen's building. When the French king Louis-Napoleon chose this as his residence the Council moved to the Prinsenhof.

In 1926 the chairman of the board of The Netherlands Trading Company offered his private house to the burgomaster: Herengracht 502 then became the official residence. It was built by an unknown architect in 1672 and commissioned by one of the directors of the Dutch East India Company.

The various tasks of the municipality cover a wide territory: public transport, public works (including acquisition and allocation of ground and buildings), public health, housing, electricity and gas, the port, markets, police, fire brigade, cleaning, social services, waterworks, education, old people's welfare work, and churchyards. The city has its own clearing bank, credit bank, advertising department, printing shop, swimming pools, theatre, archive department, museums, abattoir, and orphanage, all of which are administered and controlled by about 30,000 municipal employees.

Cultural life and recreation. As a centre of art, Amsterdam has much to offer. The Rijksmuseum is famous for its collection of 17th-century Dutch art (Rembrandt, Frans Hals, Vermeer, Jan Steen). In 250 rooms 1,000,000 art objects, including 5,000 paintings, are exhibited. The Stedelijk Museum is a leading international centre of modern art, representing such artists as the Spanish Pablo Picasso, the Russian Marc Chagall, the French Henri Matisse, the Dutch Karel Appel, the Dutch-American de Kooning, and the American Robert Rauschenberg. A third museum, opened in 1972, is dedicated to the work of Vincent van Gogh and his contemporaries, such as Paul Gauguin and Henri de Toulouse-Lautrec. Of the many other museums, the Amsterdam Historical Museum, which is located in the former Sint Lucien Convent and was built around 1420, is notable.

Museums

Besides all this, the city is a home for hundreds of painters, sculptors, authors, musicians, and actors. There are about 140 curio and antique shops.

Amsterdam is famous for its Concertgebouw Orchestra, composed of 117 musicians with 155 instruments. Finally, Amsterdam is situated in a region with many possibilities for recreation. There is not only the Amsterdam Forest, already mentioned, but also the seaside resort of Zandvoort, due west, the Sloterplas (Sloter Lake), in the heart of the western suburbs, the neighbourhood of the

former Zuiderzee, and many smaller lakes to the south and the north of the city, where various water sports can be practiced.

BIBLIOGRAPHY. One of the earliest descriptions of Amsterdam may be found in *Descrittione di tutti i Paesi Bassi* (1581; Eng. trans., *The Description of the Low Countreys and of the Province Thereof*, 1593), by the Italian historian LODOVICO GUICCIARDINI. More exact details are in JOANNES PONTANUS, *Historische Beschrijvinghe der seer wijt beroemde Coop-stadt Amsterdam* (1614); PHILIPP VON ZESEN, *Beschreibung der Stadt Amsterdam* (1664); and in JAN WAGENAAR, *Amsterdam, in zyne Opkomst, Aanwas, Geschiedenissen, . . .*, 5 vol. (1760–67). See also JOHN J. MURRAY, *Amsterdam in the Age of Rembrandt* (1967). A modern history of the city was written by HAJO BRUGMANS, *De Geschiedenis van Amsterdam* (1925). The rise of the city in the Golden Age is covered by VIOLET BARBOUR, *Capitalism in Amsterdam in the Seventeenth Century* (1950). Many details of cultural life in the 17th century are presented by PIERRE DESCARGUES in *Rembrandt et Saskia à Amsterdam* (1965). Works dealing with specific aspects are B.A.M. KROON *et al.*, *Bewaard in het hart*, with English summary (1965), on architecture; J.J. VAN DER VELDE, *Stadsontwikkeling van Amsterdam 1939–1967*, also with English summary (1968), on town planning in modern times; and G.H. KNAP, *Havenboek Amsterdam* (1969; Eng. trans., *The Port of Amsterdam*, 1970). The best guides for tourists are HENDRIK F. WIJNMAN, *Historische Gids van Amsterdam* (1971); and BRYCE ATTWELL, *Amsterdam* (1968), with parallel English, Dutch, German, and French text.

(Ev.W.)

Amu Darya (River)

The longest river of Soviet Central Asia, the Amu Darya (Amudarya in Russian, Amudarja in the transliteration system of the Akademiya Nauk), one of the great rivers of history and of major contemporary economic significance, is conventionally considered to run from the confluence of the Vakhsh and Pyandzh rivers to its mouth on the southern shore of the Aral Sea. It has a length of 879 miles (1,415 kilometres); but its length is 1,578 miles (2,540 kilometres), if measured from the remotest sources of its longest headstream, the Daryā-ye Vākhjīr in the Eastern Pamirs. Traditionally known to the Western world from Greek and Roman times as the Oxus, and called Jayḥūn by the Arabs, its present name allegedly derives from the city of Āmul, which is said to have occupied the site of modern Chardzhou.

The river basin. The Amu Darya Basin extends for 600 miles from north to south and for over 900 miles from east to west. It borders on the Syrdarya Basin in the north, on the Tarim Basin in the east, and on the Indus and Helmand basins in the south. Of its total area of 179,700 square miles (465,500 square kilometres), only 88,000 square miles belong to the regions whence the river is fed; that is to say, to the mountain ranges or to the Zeravshan and Kashkadarya reservoir zones. The mountainous part of the basin lies among the drifting, permanent snows and glaciers of the Pamirs (with the adjacent Alay Range [Alaysky Khrebet] to the northwest) and of the Hindu Kush, where heights average from 16,400 to 18,000 feet and in some places reach 19,000 and even 23,000 feet. Its plains form part of the Turani Lowland (Turanskaya Nizmennost) within which the desert of Kyzylkum and part of the Kara-Kum Desert also lie. Below its conventional source the river is joined by only three tributaries: from the left by the Surkhob, and from the right by the Kafirnigan and Surkhandarya. Thenceforth, the river continues to its mouth without receiving a single tributary, losing much of its water in evaporation and drainage or to irrigation on the way.

The five sections of the river

The river course. The Amu Darya may be divided into five parts: (1) from the confluence of the Pyandzh and Vakhsh rivers to the city of Kerki; (2) thence to the Ilchik Ravine; (3) from there to the Tyuyamuyun Ravine; (4) to the village of Takhiyatas; and (5) to the Delta. In its first section the river spreads lakelike to wind through a valley crisscrossed by channels and with many lakes and marshes. In its second section the valley varies from 2.5 to 16 miles wide, with the river branching around many islands. From the Ilchik to the Tyuyamuyun Ravine the river valley narrows to widths of from one to 2.5 miles, and within it the river branches and bends very much less. In its fourth section the river widens into an area of marked erosion, while in its final section—from the village of Takhiyatas to the Delta—many distributaries, characterized by temporary flooding, begin to form.

Climate. The mountainous area of the Amu Darya is characterized by heavy rainfall in winter and spring and by sharp variations of air temperature. The outlying districts of the mountain range, for example, may receive as much as 39 inches (1,000 millimetres) of rain per year, and at about 11,500 feet above sea level average temperatures for July are 59° F (15° C) and for November −4° F (−20° C). In the foothills the corresponding temperatures are 86° F (30° C) and 36° F (2° C), respectively. The climate of the plains, conversely, is very dry with average July temperatures of from 79° F (26° C) to 90° F (32° C) and a maximum temperature of almost 122° F (50° C). Average temperatures for the same area in November range from 32° F to 43° F (0° C to 6° C) at Chardzhou, and at Turtkul from 25° F to 43° F (−4° to +6° C). Frost forms at temperatures of −4° F to −13° F (−20° to −25° C). Rainfall on the plains is slight—perhaps only eight inches (200 millimetres) per year—with rainfall on the lower reaches of less than four inches (100 millimetres) per year.

Vegetation and fauna. Junipers and poplars grow down to the river's edge in the mountain regions where bushes of sweetbrier and blackberries also abound. Willows, buckthorn, and oleasters predominate lower down while on the river's middle course, the *tugai*, or dense flood-meadow forest, made up of different varieties of willow and poplar proliferates. Willows, oleasters, and poplars continue along the river's lowest reaches to form an impenetrable tangle at the river's reed-covered delta. Fish found in the Amu Darya include varieties of sturgeon, carp, barbel, and trout, while riverain animals include boar, wildcat, jackal, fox, and hare. There are also 211 species of birds.

Human ecology. The Tadzhik people raise cattle in the remote pastures of the mountainous regions. On the lower reaches of the river, various Turkmen peoples predominate, but Uzbeks, Kara-Kalpaks, and, to a lesser extent, Tatars and Kazakhs are also represented. There are numerous Russian and other Slavic immigrants from the west, and some Koreans from the east.

History. Well-known as it was in antiquity and to the Arabian geographers of the 2nd and 3rd century, the Amu Darya (Oxus), nevertheless, received but little attention until the reign of Peter I. Though the first relatively authentic map was made in 1734, systematic research of the river was begun only at the end of the 19th century. At the end of the 1920s a map of the entire Amu Darya Basin was published in Tashkent. Detailed studies by Soviet scientists are still going on.

Hydrology. The Amu Darya is largely dependent on the waters of its tributaries, the Pyandzh and the Vakhsh, the greater part of the flow of which derives from melting snow and ice. Increasing from March to May, when snow melts on the plains and rainfall increases, the river's flow is further augmented in June, July, and August as the ice and snow of the mountain ranges thaw. It gradually abates from September to February. Ice forms along the banks of the upper reaches, as do sludge and drifting ice. Lower down near the delta, however, Nukus is icebound for from two to two and a half months. Beginning to disperse at the beginning of February in the middle reaches and elsewhere toward the end of March, ice floes jam the river downstream. In its upper course the river's flow is stable; in its lower course it is much less so. The amount of matter suspended in the water is very high. Near Kerki, for example, it has averaged 210,000,000 tons per year since 1960 and, near Nukus, 130,000,000 tons. At the same places ratios of alluvium of 0.25 pounds and 0.17 pounds, respectively per cubic foot have been recorded. The water also contains hydrocarbonate.

The river's alluvial content

Irrigation, navigation, and power. The diversion to irrigation of water from the main stream and from its

tributaries is desirable—particularly in summer—for, though more than 100 canals water the cultivated lands that border its lower course, and more are being built, water supplies elsewhere remain inadequate.

The unstable river bed and shoals so hamper navigation that though the Amu Darya is navigable from its mouth to Termez little transport is conveyed by it. A complex system of dams has been erected, mainly on the lower course, to protect the cultivated fields from floods. Further dams, reservoirs, and barrages are at the planning stage or are already under construction. A large hydroelectric station is being built at Vakhsh, and others are planned to utilize the waters at the Vakhsh and Pyandzh Falls. (A.A.K.)

Amur River

The Amur River (Chinese Hei-lung Chiang) in eastern Asia forms part of the frontier between the Union of Soviet Socialist Republics and the People's Republic of China. It rises on the northern border of Northeast China and Inner Mongolia with the Soviet Union and flows generally east and southeast along that border to Khabarovsk, where it turns northeastward and flows to the Tatar Strait, separating Siberia from the island of Sakhalin. It is the longest river in the Soviet Far East. Its Chinese name, Hei-lung Chiang, means Black Dragon River; its Mongol name, Kharamuren, means Black River; among local tribes it is known as Mamu.

The Amur begins at the confluence of the Shilka and the Argun rivers, at a point 1,755 miles (2,824 kilometres) from its mouth. The Shilka rises more than 340 miles farther inland at the junction of the Siberian Ingoda River and the Mongolian Onon River. The Argun River rises in the Chinese autonomous region of Inner Mongolia, about 1,000 miles from its confluence with the Shilka. The entire river system of the Amur includes numerous rivers and lakes; among the rivers are five major ones—the Zeya, Bureya, Sungari, Ussuri, and Amgun.

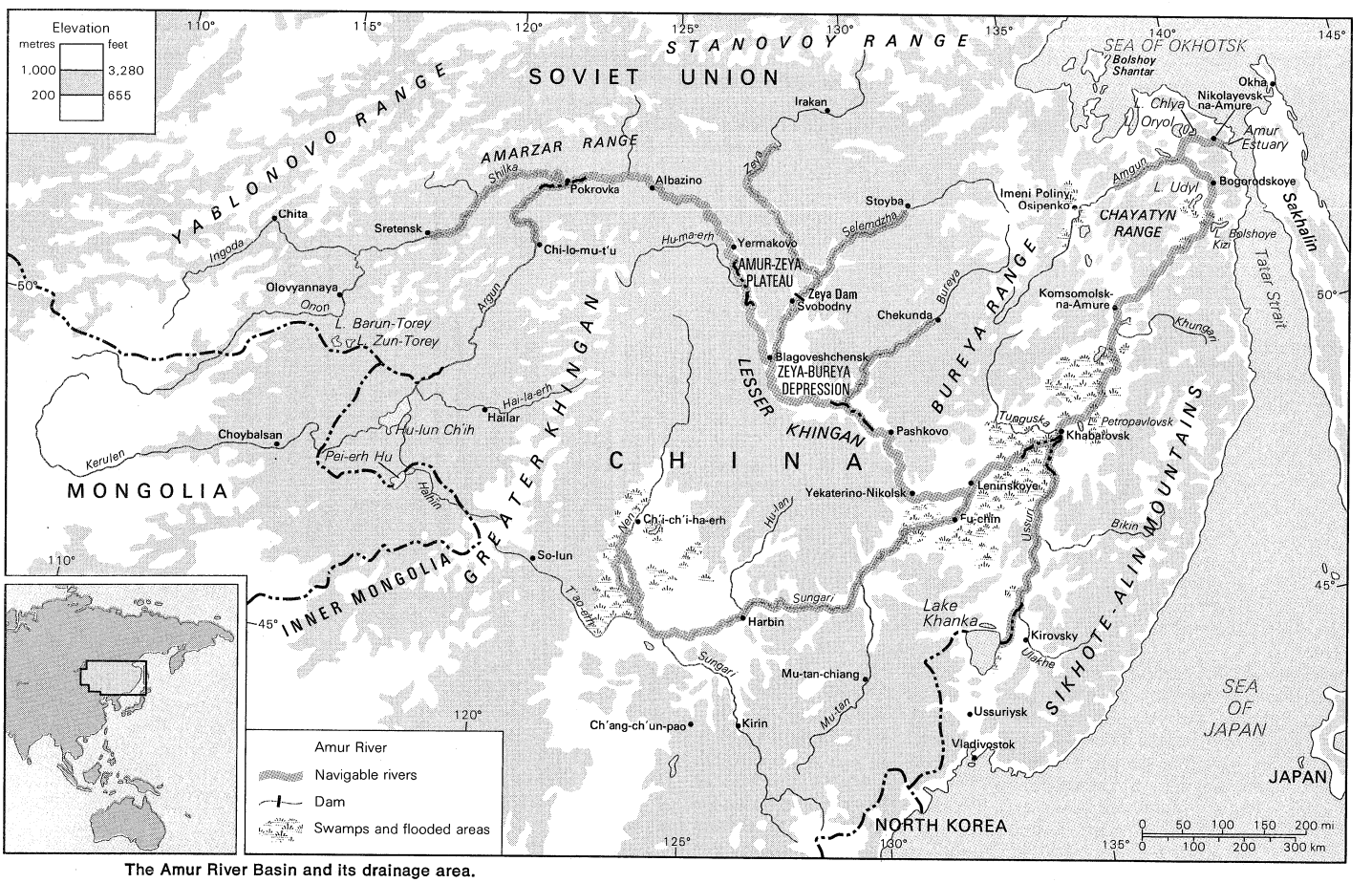
The Amur Basin. It is customary to divide the river into three sections: the Upper, Middle, and Lower Amur. The Upper Amur begins at the juncture of the Shilka and Argun and ends at the mouth of the Zeya (Blagoveshchensk), about 550 miles downstream. The Middle Amur extends about 600 miles to Khabarovsk. The Lower Amur, from Khabarovsk to the mouth, is about 600 miles.

The Upper Amur

The Upper Amur flows through a mountain valley between the spurs of the Greater Khingan, which is covered by thick larch woods, and the pine-clad slopes of the Amarzarsky mountain ridge. Near Albazino the mountains part, and the river enters open-plateau country. The winding river has built up sandbanks, many of them really spits of land formed at looping bends of the river, which in places has cut deeply into the earth. The terraced slopes of the valley testify that in remote times the land was flat and the Amur wandered through wide floodlands. Below Yermakovo lie the Burning Mountains, rocky precipices made up of complex layers of spontaneously igniting carbonaceous, clayey shales that are continually steaming and sometimes give off flames.

Middle Amur flows into the Zeya-Bureya Depression. The left slope of the valley merges imperceptibly with the plain, while the right slope—steep and high—borders the Lesser Khingan Range. Below the Bureya River the plain narrows gradually, and near Pashkovo the river runs past spurs of the Bureya Range. Farther on it flows along a narrow ravine-like passage through the Lesser Khingan mountains, its depth and speed sharply increasing.

The Lower Amur runs between low, overflowing banks into a vast marsh, the surface of which is broken by channels and dotted with lakes and ponds; the riverbed branches often, and the channel becomes very wide. Near Novoye the Sungari, the Amur's largest tributary, pours in its yellow, muddy waters. Near Khabarovsk it is joined by the Ussuri. With these accessions to its waters, it overflows widely over the flat, marshy ground of the tidal



The Amur River Basin and its drainage area.

valley. The riverbed becomes a labyrinth of branches, channels, offshoots, former riverbeds, islands, sandbanks, and spits. During high-water seasons these sections overflow, and the region becomes an enormous lake. At Khabarovsk the Amur is only 230 miles from the shores of the Sea of Japan; but, diverted by the Sikhote-Alin Mountains, it runs northward for 600 miles before it finds the sea. Near Komsomolsk-na-Amure the plain gradually narrows, and the river flows for 90 miles among mountains into a scenic forest valley. It then enters the Udil Kizinsky hollow with wide, flat, marsh-ridden ground. In the hollow lie two large lakes—the Big Kizi and the Udyl. Near Bogorodskoye the hollow is closed in by the Chayatyn Range, and the river flows out onto a low-lying plain, where the Amgun, the last of its important tributaries, joins the Amur on its left bank. It enters the sea through a wide, bell-shaped estuary, which is about 30 miles long.

Climate. The Amur Basin has a monsoon climate—a seasonal alternation of winds from the mainland and the ocean. In winter, the dry, cold air from Siberia brings clear, dry weather with strong frost. Snowfall is generally light. In summer the moist ocean winds predominate, raising the water level in the Amur and its main tributaries.

Autumns are warm and dry. Average temperatures, in January, range from −11° F (−24° C) in the south to −27° F (−33° C) in the north. The average July temperature in the south (at Blagoveshchensk) is 70° F (21° C) and in the north about 64° F (18° C). Precipitation in the Amur Basin is uneven, being heaviest in the maritime section, where it ranges between 24 and 36 inches (600 and 900 millimetres) annually. In the middle regions the annual rate does not exceed 24 inches, and in the western, continental regions it averages 12–16 inches. The peak months are July and August, when there are four or five inches of rainfall.

Wide range of precipitation

Fish life. The Amur is rich in fish. The lower river has about 100 species of fish, the upper river 60, surpassing even large European rivers such as the Volga and the Danube. About 25 or 30 species are of commercial value. Northern forms include the Siberian salmon, the sig, and the burbot; southern forms include the Chinese perch and the white amur. There are about 20 indigenous species of carp and broadhead. A biological peculiarity of the Amur fishes is the large number of species that develop in the sea and therefore escape exposure to the sharp changes in water level that occur in the river in summer.

Settlement. The territories of the Amur were originally populated by hunting and cattle-breeding nomadic tribes, such as the Buryats, Yakuts, Nanais, Gilyaks, Udegeys, and Oroks, From the 18th century onward the area north of the Amur was settled by Russians along with Ukrainians, Belorussians, Tatars, Latvians, and others. South of the Amur live Chinese, Mongols, Manchus, and many other peoples.

The Russians began to explore the Amur Basin in the 17th century. Maksim Perfilyev passed through the region in 1638. In the spring of 1644 Vasily Poyarkov set off on a three-year exploration of the Amur; he covered much of the river basin and estuary. Yerofey P. Khabarov followed in 1649–51; Khabarovsk—the second largest city in the Soviet Far East—is named for him.

In 1849–55 an expedition under the leadership of a naval officer, Gennady I. Nevelskoy, made the discovery that Sakhalin is an island and that, therefore, the Amur is accessible from the south and not alone from the north as the Russians had previously supposed. This was the beginning of the systematic study of the water system of the Amur. A series of large expeditions was carried out in the first part of the 20th century (1901–09, 1927–29, 1932, 1935–36). In 1952–58 the Academy of Sciences of the U.S.S.R. made a thorough exploration of the Amur.

Hydrology. The river is fed principally by the monsoon rains that fall in summer and autumn. The rainwater finds its way quickly into the river, resulting in a period of flooding from May to October. There are usually several times of high water when the river is from 16 to 26 feet (5 to 8 metres) above its usual level. In particularly rainy years the high-water level may be as much as 45 feet above the usual level in the Upper Amur. Farther downstream the variations are less; in the lower reaches, the maximum rise is about 8 feet. In August and September the waters begin to abate. The lowest level of the river is reached in March and April, before the spring flood, which is fed mainly by runoff of melted snow and which is much smaller than the floods that occur in summer and autumn. Those floods often cause serious economic loss.

Seasonal flooding

The mean discharge of the Amur is 380,000 cubic feet (10,900 cubic metres) per second. The rate of flow near Komsomolsk-na-Amure is about 348,200 cubic feet (9,860 cubic metres) per second and at the mouth about 383,800 (10,900). The rate varies from as low as 5,300 cubic feet (150 cubic metres) per second to 7,000 (200) in winter near Khabarovsk; the highest rate ever recorded was 1,400,000 (40,000) in 1897.

Ice forms in the Amur in the second half of October. The Upper Amur becomes icebound at the beginning, the Lower Amur in the second half of November. The lower reaches of the river open at the end of April, the upper reaches at the beginning of May. Ice jams often occur in the sharp bends of the river, raising the water level by as much as 50 feet.

At the mouth of the Sungari, the water in the Amur is comparatively clear, but lower down it becomes muddy. In a year the river carries down about 20,000,000 tons of sediments.

Economic resources. The Amur is navigable throughout its course for about half the year. The large tributaries—the Zeya, Bureya, Sungari, Tunguska, Ussuri, and Amgun—are navigable, as are some of the lakes. The main ports include Pokrovka, Blagoveshchensk, Leninskoye, Khabarovsk, Komsomolsk-na-Amure, and Nikolayevsk-na-Amure. The potential hydroelectric power of the Amur River Basin is estimated at roughly 45,000,000 kilowatts, but little of this potential has so far been developed. (A.P.M.)

Analgesic

Since the dawn of history, man has searched for ways to relieve pain, a search that has resulted in the development of a class of drugs known as analgesics. Analgesics may be defined as drugs that relieve pain without blocking the conduction of nerve impulses or altering markedly the functioning of the sensory apparatus. Because analgesics relieve pain without markedly interfering with physiological processes, they are among the most important drugs available for medical use.

The analgesics may be roughly divided into two groups. In one group are the strong analgesics, which relieve severe pain and act mainly on the central nervous system. In the second group are the mild analgesics, which relieve minor pains, such as headaches; mild analgesics probably act mainly via peripheral mechanisms. The mild analgesics are also antipyretic (fever reducing) and anti-inflammatory; strong analgesics have neither of these properties. As antipyretics, mild analgesics reduce the high body temperature that is associated with fever; as anti-inflammatory agents they reduce inflammation and swelling.

General features. *Historical background.* The strong analgesics, often referred to as narcotic analgesics, are opium derivatives of the poppy *Papaver somniferum.* Since 1803, when pure morphine was first obtained from opium, thousands of synthetic and semisynthetic substances with the pharmacological properties of morphine have been prepared. The narcotic analgesics are dealt with in detail in the article NARCOTIC.

The mild analgesics were also first derived from natural products. In 1763, willow bark, because its bitter taste is similar to that of the more expensive quinine, was introduced as a substitute for quinine for the reduction of fever. Willow bark, which had also been used by the ancients to relieve pain and fever, owes its effectiveness

Willow bark

to one of its components, salicin, which is a derivative of salicylic alcohol. Salicylic acid was introduced into medicine soon after its development in 1874. The use of aspirin, a derivative of salicylic acid, became widespread in 1899 after the accidental discovery of its superiority to salicylic acid. The simple nature of the mild analgesics and the ease of their synthesis led to the preparation of large numbers of derivatives. The mild analgesics introduced in the 20th century have been the result of both the accidental discovery of clinical properties of specific derivatives and large-scale research and development programs by pharmaceutical companies (see PHARMACEUTICALS, PRODUCTION OF).

Mechanism of action and physiological effects. The specific mechanism of action of the analgesics has not yet been established. The body functions that are differentially affected by the strong and mild analgesics, however, have helped to clarify their effectiveness in relieving different types of pain. Unlike the mild analgesics, the strong ones elevate the physiological threshold to painful stimuli, generally tend to cause sleepiness, and usually do not have much anti-inflammatory activity. One important component of pain, the individual appreciation of the pain experience, consists of several factors—*e.g.*, emotional states and memory of prior experience—all of which occur simultaneously. This component of pain is affected by mild and, particularly, by the strong analgesics.

Absorption, metabolism, and excretion. The strong analgesics are, in general, poorly absorbed from the gastrointestinal tract and, with a few exceptions—such as codeine, methadone, anileridine, and propoxyphene—are given by injection. The mild analgesics are generally well absorbed from the gastrointestinal tract and are usually given by mouth.

The strong, or narcotic, analgesics—such as morphine—are rapidly metabolized; they are excreted mainly in the urine. Simple synthetic ones, such as meperidine, are more extensively metabolized in the liver, and only about 5 percent of the drug is excreted unchanged. The mild analgesics are, in general, extensively metabolized in the liver and excreted in the urine; for example, phenacetin is converted in the liver to acetaminophen, and acetylsalicylic acid (aspirin) is excreted in the urine as salicyluric acid.

Toxicity. The major acute toxic reaction to the narcotic analgesics is depression of the respiratory centre in the brain; chronic toxicity is indicated by tolerance to the drug (*i.e.*, increasing doses are required to obtain the same effect) as well as by psychic and physical dependence. The mild analgesics have a relatively low toxicity, although acute aspirin overdosage frequently occurs. Signs and symptoms of aspirin overdosage include nausea, ringing in the ears, and diarrhea; in addition, the balance of electrolytes, the salts in body fluids, is upset, and the respiratory centre is at first stimulated, then depressed. Death may occur as a result of respiratory paralysis.

Chronic use of the salicylates often results in gastrointestinal irritation and bleeding. Allergic reactions to the salicylates are rare, considering the enormous amounts consumed. Analgesics such as phenacetin may cause methemoglobinemia, a disease in which the blood pigment hemoglobin is unable to function in respiration, after prolonged use; as a result of the reduced oxygen-carrying capacity of the blood, cyanosis and anoxemia may also occur. When used in large doses for extended periods, phenacetin may also cause kidney damage; evidence indicates that phenacetin may be habit forming. Certain analgesics (*e.g.*, aminopyrine) may cause agranulocytosis (breakdown of certain white blood cells). The use of the drug indomethacin, which is one of the newer mild analgesics, is associated with such undesirable effects as a relatively high incidence of headache and gastrointestinal disturbances.

The toxic effects of narcotic analgesics are dealt with in the article NARCOTIC.

Types. *Salicylates.* Only three salicylates are commonly used as analgesics: sodium salicylate, the sodium salt of salicylic acid; acetylsalicylic acid, or aspirin, the acetic-acid ester of salicylic acid; and salicylamide, the amide of salicylic acid. Methyl salicylate (oil of wintergreen) is used locally as an ingredient in a wide variety of liniments. Aspirin

The salicylates are commercially available alone or in combination with other analgesics, stimulants, and buffers (neutralizing agents). Except when salicylates are combined with strong analgesics, such as codeine or *d*-propoxyphene, no good evidence exists that combinations offer any therapeutic advantage over the salicylate alone. The salicylates are useful in the relief of headache and muscle and joint aches. The swelling and pain associated with rheumatoid arthritis are effectively relieved by aspirin; indeed, it is the drug of choice in the treatment of this disease. Aspirin has some value in the reduction of fever and the relief of pain associated with mild infections; it does not, however, modify or shorten the duration of a disease.

Aspirin and sodium salicylate have an important role in the treatment of rheumatic fever, which may cause damage to the heart; it has been estimated that more than 90 percent of the cardiac disorders occurring in patients under 20 years of age are caused by this disease. High doses of aspirin or sodium salicylate result in a dramatic lowering of the fever and relieve the pain and swelling of the joints, which are prominent symptoms of the acute phase of the disease.

Of the three commonly used salicylates, aspirin is preferred for most therapeutic uses. It is broken down in the body to salicylic acid but causes less gastrointestinal irritation than does sodium salicylate. Although salicylamide is not converted to salicylic acid to any extent and thus irritates the gastrointestinal tract even less, its analgesic and anti-inflammatory activities are not as pronounced as are those of aspirin.

In general, the salicylate analgesics are effective, safe, and inexpensive. Since they are freely available, however, the tendency is for them to be overused. The fact that over 25,000,000 pounds (about 10,000,000 kilograms) of aspirin are produced annually in the United States attests to the popularity of this drug.

Aniline derivatives. A second class of mild analgesics is the aniline derivatives. Acetanilide (acetylaniline) was first introduced into medical practice in 1886 following the discovery of its fever-reducing (antipyretic) action in a patient whose pharmacist had accidentally added it to a prescription. The drug became widely used for its analgesic and antipyretic effects but produced a high incidence of the blood disease methemoglobinemia. It has therefore been replaced by the less toxic acetophenetidin, or phenacetin, which is widely used alone and in combination with aspirin and either caffeine or codeine. The combinations are known as APC mixtures; those with caffeine were at one time among the most popular headache remedies. Aspirin is now sometimes substituted for phenacetin as a result of the increasing awareness of the possible kidney damage that may be caused by high doses and prolonged use of acetophenetidin. The discovery in 1949 that acetophenetidin and acetanilide both are converted in the body to acetaminophen (N-acetyl-*p*-aminophenol) led to the introduction of this compound as a replacement for acetophenetidin, either alone or in a variety of mixtures including aspirin and caffeine. Evidence indicates that acetaminophen causes less methemoglobinemia than does acetophenetidin. Acetaminophen

Although the aniline compounds have analgesic and antipyretic activity equivalent to that found for the salicylates, their anti-inflammatory activity is somewhat less. The aniline analgesics offer relatively little advantage over the salicylates; and, although the aniline analgesics produce less gastrointestinal irritation, they do cause the formation of altered blood pigments. Moreover, kidney damage after prolonged use of these analgesics at high doses has been reported.

Pyrazolone derivatives. In 1884 the pyrazolone analgesics, in the form of the compound antipyrine, were introduced; antipyrine was followed in a few years by aminopyrine. Antipyrine, as indicated by the name, was

first used as an antipyretic; since the discovery of the analgesic properties of the pyrazolone derivatives, however, they have been used mainly to relieve pain. Both antipyrine and aminopyrine are potent analgesics and anti-inflammatory agents; however, because they are associated with a relatively high incidence of blood abnormalities, particularly agranulocytosis, these drugs now are seldom used.

Structural modification of the pyrazolone analgesics has resulted in the introduction of phenylbutazone, the clinical usefulness of which was accidentally discovered. When the drug was used in combination with aminopyrine, the therapeutic effect of the combination was greater than that of aminopyrine alone. Although phenylbutazone has some utility in the treatment of rheumatoid arthritis and gout, this drug is effective for only a short time.

The side effects of these newer pyrazalone derivatives are similar to those of aminopyrine, particularly the effects on blood-cell formation; for this reason, the use of such derivatives is limited and necessitates monitoring the patient's blood.

Indomethacin and mefenamic acid. Two relatively new mild analgesics are indomethacin and mefenamic acid. Indomethacin, which has potent anti-inflammatory and analgesic activity, is effective in the treatment of rheumatoid arthritis; it is indicated for patients in whom pain cannot be successfully controlled with aspirin. The drug has a high incidence of side effects associated with its use, however, including gastrointestinal disturbances, dizziness, mental confusion, and severe morning headache.

Mefenamic acid is similar to aspirin in its analgesic, anti-inflammatory, and antipyretic activity but causes a higher incidence of side effects, the most important of which are possible renal damage and damage to the blood-forming elements.

Methotrimeprazine and pentazocine. Two non-narcotic strong analgesics are also considered here, methotrimeprazine and pentazocine. Methotrimeprazine has potent analgesic properties. Although it does not produce physical dependence as do the narcotics, its use is limited because it induces pronounced lowering of the blood pressure. For this reason it is primarily used in bedridden patients.

Pentazocine, a weak narcotic antagonist with minimum addiction liability, is an effective analgesic for severe pain, and it may be administered orally or by injection. This weak narcotic represents the first clinically acceptable strong analgesic with a markedly reduced addiction potential. The successful development of pentazocine has stimulated further research in the development of strong analgesics.

Treatment for gout

Colchicine. Colchicine, the first drug useful in the treatment of gout, a hereditary disease caused by improper uric acid metabolism (uric acid is the main excretory product in man of proteins and nucleic acids), was introduced in 1763. Colchicine is derived from the plant *Colchicum autumnale*, or meadow saffron. The drug, which may be used orally or injected, provides relief of pain in acute attacks of gout, which is characterized by pain, swelling, and the deposition of needlelike crystals of sodium urate in the joints. *Colchicum* provides no relief from other types of pain, including those also characterized by swollen joints. The mechanism by which it exerts its analgesic action in gout patients has not yet been established.

The side effects that have been associated with the use of colchicine include severe diarrhea and other gastrointestinal disturbances (see also the article JOINT DISEASES AND INJURIES).

BIBLIOGRAPHY. L.S. GOODMAN and A.Z. GILMAN (eds.), *The Pharmacological Basis of Therapeutics*, 4th ed. (1970); and JOSEPH R. DEPALMA, *Drill's Pharmacology in Medicine*, 4th ed. (1971), two encyclopaedic compendia giving detailed descriptions of the knowledge available on drugs that have therapeutic usefulness; JOHN C. KRANTZ, JR., *et al.*, *The Pharmacologic Principles of Medical Practice*, 7th ed. (1969), a standard textbook slanted toward the medical student, emphasizing the therapeutic usefulness of the drugs in a less detailed fashion; WINDSOR C. CUTTING, *Handbook of Pharmacology*, 4th ed. (1969), an outline containing little detail but with the essential facts; chapters in WALTER S. ROOT and FREDERICK G. HOFMANN (eds.), *Physiological Pharmacology* (1963), critical, detailed reviews of the subject matter of analgesics, although now somewhat dated; GEORGE DE STEVENS (ed.), *Analgesics* (1965), a detailed review, mainly of the strong analgesics.

(L.S.H.)

Analysis, Complex

Analysis is the name given to the branch of mathematics involving the notion of limit. For example, the series $a + ax + ax^2 + \cdots$, in which there is an infinite number of terms, belongs to analysis. If the partial sums are considered $S_1 = a$, $S_2 = a + ax$, $S_3 = a + ax + ax^2, \cdots$, then as more and more terms are included in the partial sum the closer it approaches some limiting value. Thus analysis includes the differential and integral calculus (see ANALYSIS, REAL).

Complex analysis arose from studies of analytic functions of complex variables. A function $f(z)$ is said to be analytic in a two-dimensional region if, at each point of the region, the derivative $f'(z)$ exists. Analytic functions are also called holomorphic, monogenic, regular, or regular-analytic functions. Historical usage, as well as varieties of context, recommend the different terms for the same concept, as will be seen below.

The theory of analytic functions was created by the French mathematician Augustin-Louis Cauchy, who, in 1814, began the task of integrating (see ANALYSIS, REAL) between complex limits. His results were first made known in 1825. He used his new ideas to lay the foundation for a theory of ordinary and partial differential equations (see DIFFERENTIAL EQUATIONS), including applications to mathematical physics. In 1851 Bernhard Riemann of Germany outlined a theory in which the analytic function is defined by a pair of harmonic functions (functions that satisfy Laplace's equation; see DIFFERENTIAL EQUATIONS: *Partial differential equations*) joined by the Cauchy-Riemann equations. The values of the function form a geometric entity that is called a Riemann surface. In the same paper the idea of conformal mapping (see below) was introduced.

His contemporary, Karl Weierstrass of Germany, took the view that analytic functions are defined by power series (a series in which each term is a constant multiplied by the independent variable to some power; see ANALYSIS, FOURIER: *Series*) that are linked by a method of analytic continuation.

The material of this article is divided into the following sections.

Theory of analytic functions of one complex variable
- Foundations
  - Complex numbers and the complex plane
  - Functions
  - Power series
- The Cauchy theory
  - Cauchy's integral theorem
  - Representation theorems
  - Residue theorem
- Entire and meromorphic functions
- Elliptic functions
  - Origins
  - Properties
  - Theta functions
  - The Weierstrass theory

Theory of analytic functions of several complex variables
- Development of complex analysis
- Holomorphic functions of several complex variables
- Conformal maps
- Complex manifolds and analytic spaces
- Cohomology with coefficients in sheaves
- Complex line bundles; complex vector bundles
- The Riemann-Roch theorem
- Analytic continuation, domains of holomorphy
  - Runge's theorem
  - The Cousin problems
- Deformations of complex structures
- Automorphisms of analytic spaces and automorphic functions

Potential theory
Historical survey of problems and concepts
Basic concepts and results of classical potential theory
Harmonic functions
Polyharmonic functions
Sub- and superharmonic functions
Polar sets and capacity
Thin sets and fine topology on $\mathfrak{R}^n$
Dirichlet problem
Green function; Green space
The general classical convergence theorem
Dirichlet principle
Energy
Martin boundary
Axiomatizations and probability theory
Local axiomatic theories without basic kernels
Dirichlet spaces
Hunt kernels
Relations with probability theory

Theory of analytic functions of one complex variable

FOUNDATIONS

Complex numbers and the complex plane. Complex numbers are numbers of the form $a + bi$, in which a and b are real numbers and i is the square root of -1. They can be written also as ordered pairs of real numbers (a, b). The complex numbers form a field (see ALGEBRAIC STRUCTURES). The elements of this field are ordered pairs of real numbers (a, b) for which equality, addition, multiplication, and scalar multiplication are defined by well-specified rules involving ordered pairs (see Box, equations 1). The associative, commutative, and distributive laws (see ALGEBRA, ELEMENTARY AND MULTIVARIATE) hold; (0, 0) acts as zero, (1, 0) as unity. From one of these laws the result that an arbitrary ordered pair can be written as a linear combination of two fixed ordered pairs each having components one or zero in some order (see 2) is obtained, using the classical notation. The numbers $(a, 0)$ are identified with the real numbers a so that $(1, 0) = 1$, and $(0, 1) = i$ may be written.

The notation (a, b) suggests the geometric representation of the complex number $a + bi$ by the point (a, b) in an ordinary Euclidean plane (see GEOMETRY, ANALYTIC AND TRIGONOMETRIC). The x-axis is then called the real axis, and the y-axis the imaginary one. An alternate representation is by the vector (see ANALYSIS, VECTOR AND TENSOR) joining (0, 0) with (a, b). The length r of this vector is called the absolute value of the complex number $c = a + bi$ and is written $|c| = r$. The angle, symbolized here by the Greek letter theta, θ, that the vector makes with the positive real axis is called the argument, $\theta = \arg c$ (see Figure 1). Thus an arbitrary complex number can be written in trigonometric or exponential form (see 3). The point $a - bi = \bar{c}$ is called the conjugate of c. In the geometric representation of complex numbers, addition becomes vector addition. The absolute value of a product is the product of the absolute values, and the argument is congruent to the sum of the arguments, modulo 2π (see NUMBER THEORY).

The plane used for this representation is known as the complex plane. The complex variable is usually denoted by $z = x + iy$. There is an extensive geometry of the complex numbers based on this representation. In its trigonometric form z can be expressed by $z = r(\cos\theta + i\sin\theta)$. Given another complex number z', it can be shown that the product $z \cdot z'$ accords with the general relationship $z^n = r^n(\cos n\theta + i\sin n\theta)$, which is known as de Moivre's formula. It holds for any real number n and can be used to compute the nth roots of a complex number by means of the expression $n\theta = \phi + k2\pi$, in which k is an integer and ϕ is the angle involved in the circular plot (Figure 2) of radius r. For the following discussion it is important to realize that the equation $|z - c| = r$ represents a circle with centre at c and radius r and that the inequality, the absolute value of $z - c$ less than r—that is, $|z - c| < r$—defines the interior of the circle, often called a disk.

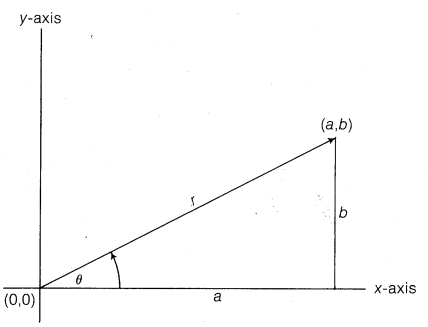


Figure 1: Geometric representation of a complex number $a + bi$ by the vector r that joins the origin (0, 0) and the point (a, b) in the x, y plane, making an angle θ with the x-axis. As indicated by the relations shown, $r = \sqrt{a^2 + b^2}$ and $\theta = \arctan b/a$.

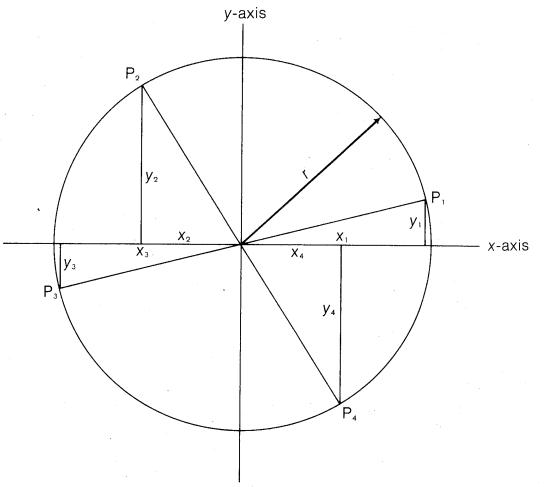


Figure 2: Graphic illustration of the solution for the nth roots of a complex number when $n = 4$. The circle is of radius r, and the points P_1, P_2, P_3, P_4 are obtained by finding four values of θ from the expression $\theta = \frac{\phi}{4} + k\frac{360^\circ}{4}$ when k takes on the values 0, 1, 2, 3.

Functions. The idea underlying the function concept is that of a correspondence, or of a mapping of one set of objects D, called the domain, upon another set of objects R, called the range. If there is a unique definite object Q of R corresponding to every object P of D, then Q is said to be a function of P and is written $Q = f(P)$. If R and D are sets of real numbers, then the function is a function Q of a real variable.

For functions of complex variables, D and R are sets of complex numbers, and D will then be a domain in a narrower sense, namely, an open connected set (often called a region). This means that for every point of D there is a small neighbourhood that is also in D, and that any two points of D may be connected by a broken line in D.

A function $f(z)$ defined in D may be continuous there. This means that if z is near to z_0 in D, then $f(z)$ is near to $f(z_0)$ in R. The differentiable functions form a subclass of the continuous functions. In the present case $f(z)$ is said to be differentiable at a point z_0 in D if the ratio (see 4) of the difference of the function evaluated at 2 points to the difference of the 2 points approaches a finite limit when z approaches z_0, the limit being independent of how z approaches z_0. This limit is known as the derivative of $f(z)$ at $z = z_0$ and is denoted by $f'(z_0)$ (see ANALYSIS, REAL). If $f'(z)$ exists everywhere in D, the function $f(z)$ is said to be holomorphic or analytic in D. The term analytic, however, will be used below in a more general sense.

The formal rules of the calculus apply to complex-valued functions. Thus, sums and products of functions holomorphic in D are also holomorphic in D, although in the case of a quotient the points at which the denominator is zero must be excepted. If n is a positive integer, z^n is holomorphic in the finite plane; this implies that every polynomial is a holomorphic function and that

rational functions are holomorphic except at the zeros of the denominators.

If $f(z) = U(x, y) + iV(x, y)$ is holomorphic in D, then the real-valued functions U and V have partial derivatives with respect to x and y that satisfy the Cauchy-Riemann equations relating these partial derivatives (see 5) in which subscripts indicate differentiation with respect to x and y. Conversely, if a pair of real-valued functions has continuous partials in D that satisfy the Cauchy-Riemann equations, then $f = U + iV$ is holomorphic in D. The functions U and V are known as conjugate harmonic or logarithmic potential functions. They satisfy Laplace's equation (named after Pierre-Simon Laplace, an 18th–19th-century French mathematician) in which the sum of two second partial derivatives is restricted to equal zero (see 6). This is the two-dimensional analogue of the equation satisfied by the Newtonian potential.

Power series. The basic class of holomorphic functions is formed by the power series; that is, a series of the form in which only constants times nonnegative integral powers of z minus a appear (see 7). Such a series converges for a given z if the sequence of partial sums converges to a finite limit. This may happen only for $z = a$ or, at the other extreme, for all values of z. In general there exists a positive number R, known as the radius of convergence of the series, such that it converges (even absolutely) for $|z - a| < R$ and diverges for every z outside the circle $|z - a| = R$, known as the circle of convergence. The value of R is given by a limiting form for its reciprocal (see 8).

The sum of the series in $|z - a| < R$ is a holomorphic function $F(z)$, the derivative of which can be obtained by termwise differentiation of the power series. The resulting power series has the same radius of convergence and may be differentiated as often as desired. Thus $F(z)$ has derivatives of all orders. According to Weierstrass, the power series defines an element of an analytic function. Further elements can be obtained by elementary rearrangements of the power series (see 7).

If $|b - a| < R$, $z - a = (z - b) + (b - a)$ can be written, and the nth power of this expression can be expanded by the binomial theorem, multiplied by a_n and added. Collection of terms results in Taylor's series (named after the 17th–18th-century English mathematician Brook Taylor); that is, the result of such collection is a power series for f in which the coefficients are constructed from values of derivatives of f (see 9) and that converges at least for $|z - b| < R - |a - b|$. If the radius of convergence exceeds this value, then the new series defines a holomorphic function in a disk that overlaps the old one. In the intersection of the two disks the functions coincide; outside, each function defines the analytic continuation of the other. The new element (see 9) can be rearranged in the same manner and gives rise to further new elements. The totality of such elements (with some adjunctions; see below) constitutes the analytic function defined by the first power series (7). In this process, it is possible to get back to the starting point with an element distinct from this series (7). The number of distinct elements having the same centre $z = a$ is at most countable.

It may happen that the power series (7) is not continuous outside its circle of convergence, which in this case is said to be the natural boundary of the function $F(z)$. In a certain sense this is the rule rather than the exception; the phenomenon is apt to present itself if the coefficients exhibit recurrent nonperiodic irregularities, such as very long gaps between terms, or if the coefficients have only a finite number of distinct values that do not recur periodically.

THE CAUCHY THEORY

Cauchy's integral theorem. A rectifiable curve C is given by $z = g(t)$, for t greater than or equal to zero and less than or equal to unity; that is, $0 \leqslant t \leqslant 1$, in which $g(t)$ is continuous and of bounded variation. If $f(z)$ is continuous on C, then the Riemann-Stieltjes integral (Thomas Jan Stieltjes was a 19th-century Holland-born French mathematician) of $f[g(t)]$ with respect to $g(t)$ (see 10) exists. If D is a simply connected domain with $f(z)$ holomorphic in D, and, further, if C is a simple, closed, rectifiable curve in D, then the integral of the function around the curve is zero (see 11).

This basic theorem, announced by Cauchy in 1825, is the source of most of the results in analytic-function theory. First, it gives a representation of $f(z)$ in terms of its boundary values on C. If z is interior to C, then the function evaluated at that point has a representation in terms of an integral extended around the boundary (see 12). The integral can be differentiated with respect to z under the sign of integration as often as desired (see 13). Thus the existence of a first derivative implies the existence of derivatives of all orders.

Representation theorems. Cauchy's formula (see 12) can also be used to obtain expansions of $f(z)$ in infinite series. The Cauchy kernel $(t - z)^{-1}$ is expanded in a series of the desired type; multiplication by $f(t)$ and termwise integration then gives the corresponding series for $f(z)$. In this manner, for instance, Taylor's series (9), with F replaced by f, which converges in the largest circle with centre at b in which $f(z)$ is holomorphic, is obtained. Laurent's series is obtained in an analogous manner (see 14). It is valid in the largest annulus (a ring-

(1) $$\begin{cases} (a, b) = (c, d) \text{ if and only if } a = c, b = d \\ (a, b) + (c, d) = (a + c, b + d) \\ (a, b) \cdot (c, d) = (ac - bd, ad + bc) \\ c \cdot (a, b) = (ac, bc) \end{cases}$$

(2) $$(a, b) = a \cdot (1, 0) + b \cdot (0, 1) = a + bi$$

(3) $$c = r(\cos\theta + i \sin\theta) = re^{i\theta}$$

(4) $$\frac{f(z) - f(z_0)}{z - z_0}$$

(5) $$U_x = V_y, \qquad U_y = -V_x$$

(6) $$F_{xx} + F_{yy} = 0$$

(7) $$\sum_{n=0}^{\infty} a_n (z - a)^n$$

(8) $$\frac{1}{R} = \limsup |a_n|^{1/n}$$

(9) $$F(z) = \sum_{k=0}^{\infty} \frac{F^{(k)}(b)(z - b)^k}{k!}$$

(10) $$\int_C f(z)\, dz = \int_0^1 f[g(t)]\, dg(t)$$

(11) $$\int_C f(z)\, dz = 0$$

(12) $$f(z) = \frac{1}{2\pi i} \int_C \frac{f(t)\, dt}{t - z}$$

(13) $$f^{(n)}(z) = \frac{n!}{2\pi i} \int_C \frac{f(t)\, dt}{(t - z)^{n+1}}$$

(14) $$\begin{cases} f(z) = \sum_{-\infty}^{\infty} a_n (z - b)^n \\ a_n = (2\pi i)^{-1} \int_C f(t)(t - b)^{-n-1}\, dt \end{cases}$$

$$(15)\qquad \int_C f(z)\,dz = 2\pi i \,\Sigma a_{-1,k}$$

$$(16)\qquad \begin{cases} \displaystyle\int \frac{dx}{\sqrt{(1-x^2)(1-k^2x^2)}} \\ \displaystyle\int \sqrt{\frac{1-k^2x^2}{1-x^2}}\,dx \\ \displaystyle\int \frac{dx}{(1+nx^2)\sqrt{(1-x^2)(1-k^2x^2)}} \end{cases}$$

$$(17)\qquad u = \int_0^x \frac{ds}{\sqrt{(1-s^2)(1-k^2s^2)}}$$

shaped region) $0 \leqslant R_1 < |z-b| < R_2 \leqslant \infty$ in which $f(z)$ is holomorphic. Various expansions of holomorphic functions in so-called interpolation series are obtainable by the same method.

If in the Laurent series (14) $R_1 = 0$, the point $z = b$ is an isolated singularity of $f(z)$. This singularity is removable if $a_n = 0$ for $n < 0$. It is a pole of order m if $a_{-m} \neq 0$ but a_{-k} is identically equal to zero—that is, $a_{-k} \equiv 0$, for $k > m$; it is an essential singularity in the remaining case. The coefficient a_{-1} is known as the residue of $f(z)$ at $z = b$. In the case of a pole, $|f(z)|$ is uniformly large in the neighbourhood of $z = b$. The function $f(z)$ is assigned the value ∞ (infinity) at $z = b$, and the expansion (14) is adjoined to the elements that define the analytic function.

At an essential singularity no value can be assigned to the function; in fact, $f(z)$ assumes every value infinitely often, with at most one exception, in every neighbourhood of $z = b$ (which was proved by Émile Picard, a French mathematician, in 1879).

Residue theorem. Cauchy's theorem (see 11) must be modified if $f(z)$ is not holomorphic inside C. If the only singularities are isolated and none are on C, then the integral of the function around the curve is expressed in terms of a sum (see 15), in which the summation extends over the residues of the integrand inside C. This residue theorem of Cauchy also has a large number of applications. Cauchy used it to evaluate definite integrals —a problem of considerable practical importance in electrical engineering, physics, and other fields. He also used his residue theorem for the first rigorous discussion of series expansions arising in boundary value problems of mathematical physics, and this is still a valuable method. The residue theorem also plays a central role in the theory of elliptic functions.

Another important application of the residue theorem is the principle of the argument. Suppose that inside C there are no singularities other than poles of $f(z)$ and that no pole or zero lies on C. If the residue theorem is applied to the logarithmic derivative of $f(z)$, it is found that the change in the argument of $f(z)$ as z describes C in the positive sense equals 2π times the difference in the number of zeros and poles of $f(z)$ inside C.

ENTIRE AND MEROMORPHIC FUNCTIONS

A function $f(z)$ is entire (integral in British terminology) if it is holomorphic in the finite plane. It is transcendental if it does not reduce to a polynomial. In 1876 Weierstrass showed how to construct an entire function of given zeros. An entire function without zeros is of the form $\exp[g(z)]$, where $g(z)$ is also entire. Entire functions are classified according to their order where $0 \leqslant \rho \leqslant \infty$. The order is closely related to the asymptotic behaviour of the coefficients of the defining power series and imposes an upper bound on the frequency of the zeros.

An entire function takes on every value infinitely often, with at most one exception. This theorem, proved by Émile Picard in 1879 by use of the elliptic modular function, has been in the centre of attention ever since. Elementary proofs have been found by the French mathematician Émile Borel and by the German mathematician Edmund Landau, and the theorem also follows from a mapping theorem due to A. Bloch, proved in 1924. It is now a simple consequence of an elaborate theory of the distribution of the values assumed by an entire or meromorphic function developed by F. Nevanlinna and Rolf Nevanlinna of Finland starting in 1924.

A function having only polar singularities is said to be meromorphic. The representation of such a function by partial fractions was given by the Swedish mathematician Magnus Gösta Mittag-Leffler in 1877.

ELLIPTIC FUNCTIONS

One example of analytic functions is the class of functions known as elliptic functions. They have many applications in algebraic geometry, topology, the algebraic theory of numbers, and other fields of mathematics. Also, they have influenced the development of the general theory of functions of complex variables.

Origins. It is a familiar result of elementary calculus that if R is a linear or quadratic function of a variable x, then the integral of any rational function of x and $\sqrt{R}$ is an elementary function. In the next higher case, when R is a cubic or quartic with no repeated factor, no such result is possible. It was gradually recognized by mathematicians of the 18th century that these integrals $\int f(x, \sqrt{R})dx$, in which f is any rational function of x and $\sqrt{R}$, were essentially new functions requiring special investigation. They were named elliptic integrals because a special case of such integrals occurred in the problem of finding the length of arc of an ellipse. An important formula connecting such integrals, equivalent to one of what are now called addition theorems, was discovered as early as 1761 by Leonhard Euler of Switzerland. The first systematic treatment of the new integrals, however, is a legacy of Adrien-Marie Legendre, a French mathematician, who proved, among other results, that any elliptic integral can be reduced to the sum of an elementary function and of constant multiples of integrals in three standard forms (see 16) involving the variable x as well as k and n, in which k and n are constants called the modulus and the parameter, respectively. These three integrals are called Legendre's standard (or normal) elliptic integrals of the first, second, and third kinds, respectively.

But before Legendre finished his *Traité des fonctions elliptiques et des intégrales Eulériennes* in 1826, Niels Henrik Abel of Norway, in competition with Karl Gustav Jacob Jacobi of Germany, revolutionized the subject. If a variable u is expressible as an indefinite integral of one of the three standard forms in which k appears (see 17) and if the parameter k is 0 in this integral, then $u = \sin^{-1}x$ or $x = \sin u$. In 1825 Abel recognized this fact and inverted the relation between u and x (for k not necessarily 0) by treating x as a function of u instead of following Legendre in treating the integral u as a function of x. The function x obtained in this manner is denoted by sn u or, more precisely, by sn (u, k). Two more functions, cn u and dn u, are defined similarly; together with sn u they are called Jacobi's elliptic functions, as distinguished from Legendre's elliptic integrals, which, however, Legendre himself called *fonctions elliptiques*. From the analogy with trigonometric functions (see GEOMETRY, ANALYTIC AND TRIGONOMETRIC) Abel and Jacobi immediately obtained important properties of the new functions: their double periodicity (see below *Properties*) and the addition theorems that express sn $(u + v)$, etc., as rational functions of sn u, cn u, dn u, and sn v, cn v, dn v. It is now known that many of the fundamental results of Abel and Jacobi had been anticipated, but not published, by Carl Friedrich Gauss, a German mathematician.

Properties. Jacobi's sn u, cn u, dn u are simple cases of a more general class of functions having the same or similar properties, namely, elliptic functions. Let ω_1, ω_2 be any complex numbers such that $\omega_2/\omega_1 = \tau$ is not real. A function $f(u)$ of a complex variable u is said to be doubly periodic in ω_1, ω_2 if $f(u + \omega_1) = f(u + \omega_2) = f(u)$. An elliptic function is a single-valued doubly periodic

analytic function $f(u)$ that has no singularities other than poles for any finite value of u. To obtain the properties of such a function it is convenient to use the geometrical representation of complex numbers by points in a plane (the Argand diagram, named for the Swiss mathematician Jean-Robert Argand). Any parallelogram with vertices at α, $\alpha + \omega_1$, $\alpha + \omega_1 + \omega_2$, $\alpha + \omega_2$ (α arbitrary) is called a parallelogram of the periods; if the properties of $f(u)$ are known throughout any one parallelogram, then from the periodicity its properties are known everywhere, so that it suffices to study one parallelogram. The function $f(u)$ may also be considered as an analytic function on the torus obtained by identifying opposite sides of a parallelogram of periods. This interpretation of elliptic functions as functions on tori is the starting point of the modern analytic approach to the theory.

A number of important general properties of elliptic functions were established by Joseph Liouville of France in the 19th century, mainly by the method of contour integration, including the following:

An elliptic function $f(u)$ that has no pole in a parallelogram of the periods is a constant.

The sum of the residues of $f(u)$ at the poles in a parallelogram is zero.

The function $f(u)$ has at least two poles, or a multiple pole, in a parallelogram.

The number of poles in a parallelogram is equal to the number of zeros and is also equal to the number of points at which the function has any assigned value.

The sum of the values of u at the poles in a parallelogram is either equal to the sum of the values of u at the zeros or differs from it by some period.

These theorems require obvious modifications if the poles, etc., are multiple or if any of them lie on the boundary of the parallelogram, and in the trivial case when $f(u)$ is merely a constant.

Theta functions. To study the properties of his elliptic functions, Jacobi introduced four allied functions $\theta_j(u, \tau)$, $j = 1, 2, 3, 4$, called theta functions. The first, $\theta_1(u, \tau)$, is defined by a series in which a typical term includes as factors a power of an exponential and a sinusoidal term (see 18) and θ_2, θ_3, θ_4 are defined by similar series. When u is replaced by $u + 1/2$ or $u + \tau/2$, any θ is converted into another θ multiplied by a simple factor so that the quotients of certain θ's are doubly periodic in 1, τ. Thus Jacobi could express his elliptic functions as quotients of θ's and obtain most of the fundamental properties of elliptic functions by using theta functions. The series defining the θ's converge with great rapidity when $|q|$ is small, so that they are conveniently used in numerical calculations of elliptic functions.

The Weierstrass theory. Weierstrass founded his theory of elliptic functions on a function $p(u)$ defined by the reciprocal of u squared added to a sum in which summands also include reciprocals of squares expressed in terms of u and ω (see 19) in which the sum is taken over all numbers $\omega \neq 0$ of the form $m\omega_1 + n\omega_2$, $m, n = 0, \pm 1, \pm 2, \cdots$, ω_1 and ω_2 being any given complex numbers such that ω_2/ω_1 is not real. The function $p(u)$ is an elliptic function with periods ω_1, ω_2; it has poles of order 2 at the points ω and at 0, is regular elsewhere, and satisfies a differential equation that is quadratic in the derivative of the function and cubic in the function itself, coefficients being g_2 and g_3 as well as the constant 4 (see 20), where g_2, g_3 are constants depending only on the periods. Thus $p(u)$ is the inverse function of the elliptic integral that involves the cubic just mentioned (see 21). Closely connected with $p(u)$ are the functions $\zeta(u)$, defined as a doubly infinite series with simple poles, and $\sigma(u)$, defined as a doubly infinite product with no singularities. The three functions are connected by equations (see 22). Neither ζ nor σ is periodic, but they undergo simple changes when u is increased by a period; ζ can be taken as the integral of the second kind, and the integral of the third kind is simply expressible in terms of $\sigma(u)$.

The Liouville theory leads readily to the addition theorem for $p(u)$, to formulas for the addition of half periods, and to many other results, the development being in general much more symmetrical and elegant than that for the Jacobian functions. With $\sigma(u)$ are associated three other functions $\sigma_j(u)$, $j = 1, 2, 3$, these four functions being closely related to the four θ's. The Weierstrassian and Jacobian functions are connected by equations of the form involving six constants (see 23), with similar formulas for the other σ's. A, B, C, a, b, c being constants, the properties of either set of functions can be derived from the others.

Any elliptic function $f(u)$ can be expressed in any one of three standard forms: (1) as a product of σ's divided by another product of σ's; (2) as a linear combination of ζ's and their derivatives; (3) as a rational function of $p(u)$ and $p'(u)$. From the second form the integral of $f(u)$ is at once expressible in terms of ζ and σ.

Letting $x = p(u)$, $y = p'(u)$, then $u \leftrightarrow (x, y)$ defines a one-to-one correspondence between points u in a parallelogram of periods and points (x, y) on the plane algebraic curve defined by $y^2 = 4x^3 - g_2x - g_3$, and the elliptic functions with the given periods may be considered as rational functions on this curve. This observation enabled mathematicians to develop the theory of elliptic functions entirely within the realms of algebra and algebraic geometry without recourse to the theory of functions of a complex variable. Thus the scope of the theory of elliptic functions can be broadened considerably. (K.Iw.)

(18) $$\begin{cases}\theta_1(u,\tau) = 2\sum_{n=0}^{\infty}(-1)^n q^{(n+\frac{1}{2})^2}\sin(2n+1)\pi u \\ q = e^{i\pi\tau}\end{cases}$$

(19) $$p(u) = \frac{1}{u^2} + \sum{}'\left(\frac{1}{(u-\omega)^2} - \frac{1}{\omega^2}\right)$$

(20) $$p'^2(u) = 4p^3(u) - g_2p(u) - g_3$$

(21) $$u = \int_{\infty}^{x} \frac{ds}{\sqrt{4s^3 - g_2s - g_3}}$$

(22) $$\zeta'(u) = -p(u), \qquad \sigma'(u)/\sigma(u) = \zeta(u)$$

(23) $$\begin{cases}p(u) = A + B/\mathrm{sn}^2(bu) \\ \sigma(u) = Ce^{au^2}\theta_1(cu)\end{cases}$$

Theory of analytic functions of several complex variables

DEVELOPMENT OF COMPLEX ANALYSIS

Complex analysis arose from the study of analytic functions of one or several complex variables. Beginning with Riemann's fundamental discoveries in the middle of the 19th century, there was a gradual but continual shift of emphasis toward the investigation of global properties of analytic functions by topological (see TOPOLOGY, GENERAL) and algebro-geometrical methods, with the result that complex analysis became rather intricately interwoven with the other principal mathematical disciplines and could be viewed only against the background of mathematical development as a whole, which is characterized in the 20th century by a remarkable synthesis of ideas.

Analysis of the properties of functions of one complex variable (complex analysis in one dimension) had a remarkable growth during the second half of the 19th century as a result of the brilliant research of Riemann and others. Shortly after 1900 the general principle of uniformization, conceived by Felix Klein of Germany and Henri Poincaré of France, was established by P. Koebe. These developments were presented as a unified theory in 1913 by Hermann Weyl, a German-born U.S. mathematician.

During the latter part of the 19th and the beginning of the 20th century, the theory of functions of several complex variables made its appearance, as did algebraic

geometry, especially the theory of algebraic surfaces founded by Max Noether of Germany, Émile Picard, Henri Poincaré, and, last but not least, the Italian geometers Guido Castelnuovo, Federigo Enriques, and, somewhat later, Francesco Severi. In 1895 P. Cousin investigated certain global properties of analytic functions on domains in complex n-space and brought to the attention of mathematicians two problems that were later to bear his name; the study of these and related problems led to some of the most important developments of complex analysis during the first half of the 20th century.

In the years following the publication of Weyl's book, the geometrical concepts, first introduced into complex analysis by Riemann, were studied for their own sake and vastly generalized as a result of the growth of topology as a branch of mathematics. Solomon Lefschetz, a Russian-born U.S. mathematician, made a profound study of the topological properties of algebraic surfaces in a tract that exerted considerable influence on the later development of complex analysis, topology, and algebraic geometry. Élie-Joseph Cartan, a mathematician in France, developed and applied systematically the theory of exterior differential forms that merged with topology in the fundamental theorem proved by the Swiss mathematician Georges de Rham in 1932 and helped give birth to the concept of cohomology in algebraic topology (see TOPOLOGY, ALGEBRAIC). S. Bergman discovered an important type of Hermitian metric on subdomains of complex n-space that was subsequently generalized by Erich Ernst Kähler of West Germany and became known as a Kähler metric.

Beginning about 1932, the British mathematician Sir William Vallance Douglas Hodge generalized the theory of harmonic functions on Riemann surfaces to harmonic differential forms on higher dimensional compact Riemannian manifolds, and on complex manifolds that possess Kähler metrics. Hodge's theory was developed and perfected during the 1940s by K. Kodaira, who also generalized the theory to open manifolds and introduced systematically the method of orthogonal projection in Hilbert space (named after David Hilbert, a mathematician in Germany).

Simultaneous with the development of the theory of harmonic differential forms, an intensive investigation of questions related to the Cousin problems was made by K. Oka, Heinrich Adolf Louis Behnke and Karl Stein of Germany, and Henri Cartan of France, and others. Although the methods used by these investigators differed somewhat from those of Hodge, Kodaira, and de Rham, the time was becoming ripe for a remarkable confluence of ideas, and this took place shortly after 1950 with the introduction into complex analysis of new notions from algebraic topology, namely, fibre bundles and cohomology with coefficients in sheaves (see TOPOLOGY, ALGEBRAIC).

The idea of a fibre bundle was first recognized during the period 1935–40, while the notion of a sheaf was introduced into topology a few years later by the French mathematician Jean Leray (similar concepts, never precisely defined, had occurred earlier in various special contexts). The introduction of these concepts had an explosive effect on the development of complex analysis during the following years.

A third advance of great importance in complex analysis, which took place during the period 1935–50, was the development, principally by the German mathematician Carl Ludwig Siegel, of the theory of automorphic functions of several complex variables.

HOLOMORPHIC FUNCTIONS OF SEVERAL COMPLEX VARIABLES

If $\boldsymbol{C}$ denotes the set of all complex numbers, $\boldsymbol{C}^n$ is used to denote the set of n-tuples $z = (z_1, \cdots, z_n)$. Here n is an integer and each of $z_1, \cdots, z_n$ is a complex number. For any z in $\boldsymbol{C}^n$, $|z|$ denotes the positive square root of a sum of absolute values of components (see 24); here $z = (z_1, \cdots, z_n)$, and $|z_j|$ is the absolute value of the complex number z_j for $j = 1, \cdots, n$.

Now if Ω is a subset of $\boldsymbol{C}^n$, it is called an open set if, for any point a in Ω, all points of $\boldsymbol{C}^n$ sufficiently close to a actually lie in Ω (that is, there is a positive number ρ such that any z in $\boldsymbol{C}^n$ with $|z - a| < \rho$ belongs to Ω). An open set Ω is called a domain if it is connected; that is, if any two points of Ω can be joined by an arc lying entirely in Ω.

$$(24) \qquad (|z_1|^2 + \cdots + |z_n|^2)^{1/2}$$

$$(25) \qquad \sum_{k_1, \cdots, k_n} c_{k_1 \cdots k_n} (z_1 - a_1)^{k_1} \cdots (z_n - a_n)^{k_n}$$

$$(26) \qquad f(z) = (f_1(z), \cdots, f_m(z))$$

If a complex-valued function f is defined on the open set Ω [f associates to any point z in Ω a complex number $f(z)$], f is said to be holomorphic on Ω if the following holds:

For any point $a = (a_1, \cdots, a_n)$ in Ω, there is a family $\{c_{k_1 \cdots k_n}\}$ of complex numbers, one for each n-tuple $(k_1, \cdots, k_n)$ of nonnegative integers, such that, for each z in $\boldsymbol{C}^n$, sufficiently near a, the series of generalized power series type with coefficients from the family of complex numbers (see 25) converges to the value $f(z)$ of the function f at z.

Some basic properties of holomorphic functions are the following:

1. A holomorphic function is continuous; it possesses partial derivatives of any order; and each partial derivative is again a holomorphic function.
2. If f and g are two holomorphic functions on the domain Ω (Ω is open and connected), and if $f(z) = g(z)$ for all z in some neighbourhood of a point of Ω, then f and g are identical.
3. In the maximum principle, f is taken to be holomorphic on Ω, in which Ω is a domain in $\boldsymbol{C}^n$. Furthermore, there is a point a in Ω such that $|f(a)| \geqslant |f(z)|$ for any z in Ω. It follows that f is a constant ($f(z)$ is independent of z).

For later use, the following notation is introduced. Ω is an open set in $\boldsymbol{C}^n$, and $f\colon \Omega \to \boldsymbol{C}^m$ a map of Ω into $\boldsymbol{C}^m$ (n and m are positive integers). For each z in Ω, the value $f(z)$ is a point in $\boldsymbol{C}^m$, hence is an m-tuple (see 26). This defines m functions $f_1, \cdots, f_m$ on Ω. The map $f\colon \Omega \to \boldsymbol{C}^m$ is said to be holomorphic if each of $f_1, \cdots, f_m$ is a holomorphic function on Ω.

Another basic fact concerning holomorphic functions can be stated as follows:

4. Ω is taken to be open in $\boldsymbol{C}^n$, and $f\colon \Omega \to \boldsymbol{C}^m$ is a holomorphic map. Furthermore, Ω' is taken to be an open set in $\boldsymbol{C}^m$ containing all the points $f(z)$ as z runs over Ω. Finally, ϕ is a holomorphic function on Ω'. Then the function $\phi \circ f$ defined on Ω' by $\phi \circ f(z) = \phi[f(z)]$ is holomorphic on Ω'.

A topological space is a set on which the notion of open set is given. Hence, on a topological space, the notion of a continuous function is defined. A map of one topological space onto another is said to be bicontinuous, or to be a homeomorphism, if it is one-one, continuous, and the inverse map is also continuous. Two topological spaces are said to be homeomorphic if they are connected by a homeomorphism. Homeomorphic spaces are topologically equivalent; *i.e.*, their intrinsic topological properties are the same.

A map of one open set of $\boldsymbol{C}^n$ onto another is said to be biholomorphic if it is one-one and holomorphic (which implies that the inverse is also holomorphic). The ball of $\boldsymbol{C}^n$ cannot be mapped biholomorphically onto the polycylinder of $\boldsymbol{C}^n$ if $n > 1$. Thus, topologically equivalent domains are not generally holomorphically equivalent, except under certain circumstances in the case when n is equal to 1.

CONFORMAL MAPS

The study of conformal maps arose from the necessity of constructing geographic maps. Because the surface of the Earth is approximately a sphere, it is readily seen that it cannot be represented on a flat surface with preservation of the proportion of all distances. The next best

property to be preserved is the relative shape of configurations in the neighbourhood of each point; this, in a more precise mathematical formulation, amounts to preservation of the magnitude and sense of angles at each point. Such a map is called conformal. Examples of conformal maps include the familiar Mercator (which was named after Gerardus Mercator, a Flemish mathematician and geographer of the 16th century) and stereographic projections. In a conformal map of one open set of the plane onto another, it is possible to verify that the map is given locally by a biholomorphic transformation of one complex variable.

On the other hand, a biholomorphic map of one open set of $\boldsymbol{C}^n$ onto another is not generally conformal if $n > 1$. Throughout this section it will be supposed that $n = 1$.

The preceding local characterization of a conformal map disposes only of the most trivial aspect of the map because the characterization gives no information concerning the existence of a conformal map of one given plane domain onto another. Because it is impossible, except in the most trivial cases, to construct conformal maps explicitly, attention has centred mainly on the problems of proving their existence and of determining their properties.

A necessary condition in order that there should exist a biholomorphic (one-one conformal) map of one plane domain onto another is that there should exist a bicontinuous map (homeomorphism) of the one domain onto the other, and it is a remarkable fact that this necessary condition is in some cases sufficient.

Riemann mapping theorem

If D is a domain of $\boldsymbol{C}$ that is not the whole of $\boldsymbol{C}$, and D is homeomorphic to the unit disk $|z| < 1$, then there exists a unique biholomorphic (one-one conformal) map of D onto the unit disk $|z| < 1$. It has the property that a prescribed point Z_0 of D goes into the centre of the disk, and a given direction at Z_0 transforms into the direction of the positive real axis.

Although the whole plane $\boldsymbol{C}$ is homeomorphic to the unit disk, it is impossible to map it conformally onto the unit disk; this is a consequence of a classical theorem of Liouville.

The symbol S is used to denote the class of functions f of the complex variable z that are holomorphic in the unit disk, that map it one-one conformally onto domains of $\boldsymbol{C}$, and that are normalized by the condition that their power series in z, convergent for $|z| < 1$, are of the form with lowest power z being 1 and with coefficient of z being 1 (see 27). In his proof of the general uniformization principle (of which the Riemann mapping theorem is a special case), Koebe introduced the class S and studied some of its properties. If the class S is suitably topologized, then, for each integer $n > 1$, the map $F^n\colon S \to \boldsymbol{C}^n$, which sends the function f of S into the point $(a_2, a_3, \cdots, a_{n+1})$ of $\boldsymbol{C}^n$, determined by the n coefficients $a_2, a_3, \cdots, a_{n+1}$ in the power series development of f, is continuous. If the image of S under the map F is denoted by B^n, then the set B^n, homeomorphic to the closed unit ball $|w^1|^2 + |w^2|^2 + \cdots + |w^n|^2 \leqslant 1$ of $\boldsymbol{C}^n$, is called the nth coefficient region of S. The structure of the class S is completely determined by a knowledge of these coefficient regions and the maps F^n.

In 1916 the German mathematician Ludwig Georg Elias Moses Bieberbach proved that B^1 is the region $|a_2| \leqslant 2$ and, soon thereafter, posed the problem of investigating the regions B^n for all n. He further conjectured that $|a_n| \leqslant n$ for all n (the Bieberbach conjecture). This problem and conjecture were later followed by an intensive study of the class S.

COMPLEX MANIFOLDS AND ANALYTIC SPACES

For a given (Hausdorff) topological space X, it is supposed that U is an open set on X and that ϕ is a homeomorphism of U onto an open set Ω in $\boldsymbol{C}^n$. Then given a function f on U with complex values, it is tempting to say that f is holomorphic on U if the function $f \circ \phi^{-1}$ on Ω, obtained by identifying U with Ω by means of ϕ, is holomorphic on Ω. This definition, however, is not intrinsic, in the sense that it depends very much on the choice of the homeomorphism ϕ. If ψ is another homeomorphism of U onto an open set Ω' in $\boldsymbol{C}^n$, it is possible to write a simple relationship (see 28) in which $h\colon \Omega \to \Omega'$ is the homeomorphism $\phi \circ \psi^{-1}$. If it is known that the map h is holomorphic, then the fact that $f \circ \phi^{-1}$ is holomorphic leads to the conclusion that the same fact is also true of $f \circ \psi^{-1}$.

These remarks motivate the following definition.

X is a Hausdorff topological space. $U_1, \cdots, U_k, \cdots$ is a sequence of open sets on X and, for each k, a homeomorphism ϕ_k of U_k exists onto an open set Ω_k in $\boldsymbol{C}^n$ (n is a fixed positive integer). The sequence of pairs $\{(U_k, \phi_k)\}$ defines on X the structure of a complex manifold of (complex) dimension n if the following conditions are fulfilled:

(i) Every point of X lies in at least one of the U_k.

(ii) For every pair of indices k, l for which U_k and U_l have a point in common (that is, the intersection $U_k \cap U_l$ is non-empty), it is the case that the map h_{kl} from the image of $U_k \cap U_l$ in Ω_k to its image in Ω_l (see 29) is holomorphic.

It has emerged since the 1950s that, general though it is, the concept of a complex manifold is still not sufficient to carry out certain important constructions. This necessitates the introduction of the so-called analytic spaces. To this end, the following definition, which is of importance by itself, is given:

Ω is an open set in $\boldsymbol{C}^n$ and A, a subset of Ω. A is called an analytic set if, for any point a of Ω, there is a positive number r and holomorphic functions $f_1, \cdots, f_q$ in the set U of z in $\boldsymbol{C}^n$ with $|z - a| < r$ such that the points of A lying in U are precisely those z in U for which $f_1(z), \cdots, f_q(z)$ are all zero; in other words, A is described locally by holomorphic equations $f_1 = \cdots = f_q = 0$.

$$f(z) = z + a_2z^2 + a_3z^3 + a_4z^4 + \cdots + a_nz^n + \cdots \tag{27}$$

$$f \circ \psi^{-1} = (f \circ \phi^{-1}) \circ h \tag{28}$$

$$h_{kl}(z) = \phi_l(\phi_k^{-1}(z)) \tag{29}$$

$$h_{kl} = \phi_l \circ \phi_k^{-1} \tag{30}$$

Given an analytic set A in an open set Ω in $\boldsymbol{C}^n$, holomorphic functions and mappings on A can be defined. For instance, a map ϕ of A into $\boldsymbol{C}^m$ is called holomorphic if the following holds:

For any point a of A, there is an open set U in Ω containing a and a holomorphic map Φ of U into $\boldsymbol{C}^m$ the restriction of which to $A \cap U$ is precisely ϕ; in other words, ϕ can be locally extended to a holomorphic map of an open set in Ω into $\boldsymbol{C}^m$.

An analytic space is now defined in the same way as a complex manifold; the condition that the space look locally like an open set in $\boldsymbol{C}^n$ is replaced by the condition that it look locally like an analytic set. More precisely, this may be stated in the following manner:

X is a Hausdorff topological space, and $U_1, \cdots, U_k, \cdots$ is a sequence of open sets in X such that any point of X lies in some U_k. For each k, a homeomorphism ϕ_k of U_k onto an analytic set A_k in an open set Ω_k in some $\boldsymbol{C}^{n_k}$ is given. If, whenever $U_k \cap U_l$ is non-empty, a map (see 30) of the image of $U_k \cap U_l$ in A_k into its image in A_l is holomorphic, then $\{(U_k, \phi_k)\}$ defines on X the structure of an analytic space.

If X is an analytic space defined by a sequence $\{(U_k, \phi_k)\}$ it is possible to define holomorphic functions and mappings, analytic sets and the like on X. For instance, if Y is another analytic space defined by a sequence $\{(V_k, \psi_k)\}$, a continuous map $\phi\colon X \to Y$ is said to be holomorphic if the following holds:

a is a point of X and $b = \phi(a)$. Any pair of indices k, l are chosen so that a lies in U_k and b in V_l. An open set U in U_k is also chosen containing a and such that $\phi(U)$ is contained in V_l. If U_k and V_l are identified respective-

ly with analytic sets A_k and B_l, then U becomes an open set in A_k and ϕ a continuous map of U into B_l. It is required that this map be holomorphic.

Holomorphic functions are defined in the same way.

Analytic sets are defined, exactly as in an open set in $\boldsymbol{C}^n$, as being locally defined by holomorphic equations.

The general theory of analytic spaces deals principally with two types of questions. The first concerns the geometric properties of an analytic space. Typical questions are the following:

To what extent is an analytic space different from a complex manifold? Does it have, in some reasonable sense, a model that is a complex manifold?

To what extent are the local geometric properties of the space near a point characterized by the algebraic properties of the set of holomorphic functions defined in some neighbourhood of that point?

The second type of question concerns holomorphic mappings of one analytic space into another. A typical example is the following.

If $\phi\colon X \to Y$ is a holomorphic map, what is the structure of the level sets, that is, the set of points x in X in which $\phi(x)$ is a given point y in Y? In particular, how does this set depend on y?

The majority of these questions can be answered precisely and satisfactorily. These solutions, however, are extremely technical. Here only one result will be cited, a famous theorem due (essentially) to the German mathematician Reinhold Remmert.

Let $\phi\colon X \to Y$ be a closed holomorphic map (that is, the image of any closed set in X is a closed set in Y). Then, the image of any analytic set in X is an analytic set in Y.

It should be said that much of the work on these and related problems is based on the so-called Weierstrass preparation theorem (actually due to Cauchy, 1831). Historically, it can be said that the theory of functions of several complex variables was born with this discovery, although examples had appeared earlier.

COHOMOLOGY WITH COEFFICIENTS IN SHEAVES

An n-dimensional complex analytic manifold V may be covered by a finite or (countably) infinite set $\mathbf{U} = \{U_i\}$ of open sets U_i. Their non-empty intersections are denoted by $U_{ij} = U_i \cap U_j$ (in which $U_i \cap U_j$ signifies the intersection of U_i and U_j), $U_{ijk} = U_i \cap U_j \cap U_k, \cdots$. Cochains with coefficients, or values, in the complex numbers, of dimensions 0, 1, 2, $\cdots$ are defined by associating to each U_i, U_{ij}, U_{ijk}, $\cdots$ complex numbers c_i, c_{ij}, c_{ijk}, $\cdots$ depending skew-symmetrically on the indices. For purposes of illustration, a 1-cochain $c = \{c_{ij}\}$ (cochain of dimension 1) can be considered. Its coboundary δc is a 2-cochain defined by associating to each U_{ijk} a number (see 31). A cochain c is a cocycle if its co-

$$(31) \qquad (\delta c)_{ijk} = c_{ij} + c_{jk} + c_{ki}$$

boundary vanishes. The coboundary of a $(q-1)$-cochain is a q-cocycle (cocycle of dimension q), and it is said to be cohomologous to zero. The vector space of q-cocycles *modulo* the q-coboundaries defines the cohomology of what is called the "nerve" of the covering $\mathbf{U}$ with coefficients in the complex numbers, and it is denoted by $H^q(\mathbf{U}, \boldsymbol{C})$. The cohomology $H^q(\mathbf{U}, \boldsymbol{C})$ depends on the covering $\mathbf{U} = \{U_i\}$ of V by the overlapping open sets U_i; by passage to finer coverings or by a suitable normalization of the covering, however, this cohomology can be made independent of the choice of the covering, and the result is the q-dimensional coholmology of V with coefficients in the complex numbers $\boldsymbol{C}$, which is denoted by $H^q(V, \boldsymbol{C})$.

If now c_i, c_{ij}, c_{ijk}, $\cdots$ are replaced by holomorphic functions f_i, f_{ij}, f_{ijk}, $\cdots$ on U_i, U_{ij}, U_{ijk}, $\cdots$, cochains with coefficients in the holomorphic functions will be obtained, and the resulting cohomology is called the cohomology of V with coefficients in the sheaf Ω of holomorphic functions (see TOPOLOGY, ALGEBRAIC); this cohomology, of dimension q, is denoted by $H^q(V, \Omega)$. Over any point of V the sheaf Ω consists of all functions f that are holomorphic in some neighbourhood, depending on f, of the point. In defining this cohomology the fact was used that holomorphic functions can be added, their sum being again a holomorphic function. More generally, the cohomology can be defined with coefficients in any sheaf of Abelian groups (see ALGEBRAIC STRUCTURES).

Without attempting to define sheaves in general, it can be said that various types of cohomology can be defined with coefficients in sheaves related, in various ways, to the complex analytic structure. Further examples will be indicated in the following section.

COMPLEX LINE BUNDLES; COMPLEX VECTOR BUNDLES

A divisor on a complex manifold V is defined by assigning, to each open set U_i of a covering $\{U_i\}$ of V, a meromorphic function f_i on U_i such that $f_i/f_j = f_{ij}$ is a nonvanishing holomorphic function on $U_{ij} = U_i \cap U_j$. Clearly $\{f_{ij}\}$ defines a 1-cocycle in the multiplicative sense, namely $f_{ij} \cdot f_{jk} \cdot f_{ki} = 1$. Hence a divisor defines an element (class) of the cohomology $H^1(V, \Omega^*)$ where Ω^* denotes the multiplicative sheaf of nonvanishing holomorphic functions. This class is trivial if there exists a 0-cochain $\{g_i\}$, with coefficients in Ω^*, satisfying $f_{ij} = g_j/g_i$; then $f = f_i \cdot g_i = f_j \cdot g_j$ in U_{ij} defines a meromorphic function f on V, the divisor of which is equal to the given one.

The cohomology class of the 1-cocycle $\{f_{ij}\}$ corresponds to a class of complex analytically equivalent complex line bundles (complex analytic fibre bundles the fibre of which is $\boldsymbol{C}$; see TOPOLOGY, ALGEBRAIC). A bundle F of this class may be constructed by taking the products $U_i \times \boldsymbol{C}$ the points of which are pairs (z_i, ζ_i), in which z_i is a point of U_i, ζ_i a point of $\boldsymbol{C}$, and indentifying the points (z_i, ζ_i), (z_j, ζ_j), of $U_i \times \boldsymbol{C}$, $U_j \times \boldsymbol{C}$, respectively, if z_i and z_j are the same point Z of U_{ij} and $\zeta_i = f_{ij}(Z) \cdot \zeta_j$. The functions f_{ij} are called the transition functions of the bundle F over V. There is a holomorphic map $\pi\colon F \to V$ that is induced by the projection $U_i \times \boldsymbol{C} \to U_i$ sending (z_i, ζ_i) into z_i.

A holomorphic section s of the line bundle F over an open set U in V is a holomorphic map $s\colon U \to F$ such that s followed by π is the identity map of U. Because the holomorphic sections of F over U form a complex vector space, the cohomology of V with coefficients in the sheaf of sections of a complex line bundle F over V is defined; this cohomology, of dimension q, will be denoted by $H^q(V, F)$.

A complex line bundle is a special case of a complex (analytic) vector bundle. The fibre of a complex vector bundle is $\boldsymbol{C}^m$ and its transition functions f_{ij} are $(m \times m)$ matrices of holomorphic functions with nonvanishing determinant at each point of U_{ij}. Cohomology with coefficients in the sheaf of holomorphic sections of an arbitrary complex vector bundle F is defined because the sections form a complex vector space. For each q, the cohomology $H^q(V, F)$ is a complex vector space, which is finite-dimensional if V is compact.

THE RIEMANN-ROCH THEOREM

Let F be a complex line bundle, over a compact Riemann surface V, defined by a divisor which is given by meromorphic functions f_i in each open set U_i of a covering $\{U_i\}$ of V. Because f_i/f_j is a nonvanishing holomorphic function in the intersection U_{ij} of U_i and U_j, a point in U_{ij} that is a zero or pole of f_i of a certain order is also a zero or pole, respectively, of f_j of the same order. The finitely many points of V in which the functions $f_i, f_j, \cdots$ have zeros, or poles, are therefore defined, and to each zero or pole there is assigned an integer, positive if the point is a zero and negative if it is a pole, the absolute value of which defines the order of the zero or pole. The sum of all these integers is a topological invariant of the line bundle F that is called its characteristic class; it will be denoted by $c(F)$.

A compact Riemann surface is homeomorphic to the

(topological) surface obtained from a sphere by cutting g pairs of holes in it and attaching to each pair of holes a handle, thereby forming a total of g handles, in which g is an integer 0, 1, 2, $\cdots$. The nonnegative integer g, which is a topological invariant of the Riemann surface, is called its genus. For example, the sphere with one handle, which is topologically the same as a torus, has genus equal to 1.

The classical Riemann-Roch theorem, named for the German mathematicians Bernhard Riemann and G. Roch, expressed in the language of sheaf cohomology, concerns the difference of dimensions of cohomologies (see 32) in which $\dim H^q(V, F)$ signifies the dimension of the cohomology $H^q(V, F)$. The term $H^0(V, F)$ is the (vector) space of holomorphic sections of F. Because $\dim H^1(V, F)$ is a nonnegative integer, the Riemann-Roch inequality (see 33) is obtained. In particular, if $c(F) \geqslant g + 1$, then $\dim H^0(V, F) \geqslant 2$ and there exist at least two linearly independent holomorphic sections s_1, s_2 of F over V the ratio of which s_1/s_2 is a nonconstant meromorphic function f on V. Thus the Riemann-Roch theorem, for a Riemann surface, yields information concerning the existence of meromorphic functions with given divisor on V.

In 1954 the German mathematician Friedrich Ernst Peter Hirzebruch generalized the Riemann-Roch theorem to complex analytic vector bundles F over projective manifolds V of arbitrary complex dimension. The Euler-Poincaré characteristic $\chi(V, F)$ of the cohomology of V with coefficients in the sheaf of holomorphic sections of the complex vector bundle F (see 34) may be introduced. It will be noted that the Riemann-Roch theorem for a complex line bundle F over a compact Riemann surface V expresses $\chi(V, F)$ in terms of topological invariants of V and F. The Riemann-Roch theorem of Hirzebruch expresses $\chi(V, F)$ in terms of topological invariants of V and F, called characteristic classes, or Chern classes of V and F (after the Chinese-born U.S. mathematician Shiing-shen Chern, who first gave a precise definition of these classes).

Hirzebruch's proof of his theorem, which is based on a deep and remarkable theory of the French mathematician René Thom, is nonalgebraic in character. A formulation of the Riemann-Roch theorem that contains Hirzebruch's theorem as a special case has been proved by Alexandre Grothendieck of France, using only algebraic methods.

A knowledge of the value of $\chi(V, F)$ yields less information, concerning the existence of holomorphic sections of F, as the complex dimension n of V becomes larger.

ANALYTIC CONTINUATION, DOMAINS OF HOLOMORPHY

If D is a domain in the plane C, and D' is any domain containing D, then, if $D \neq D'$, there is a holomorphic function on D that cannot be extended holomorphically to D' (choose a point a in D', which is not in D, and consider the function $1/(z - a)$). In a remarkable series of papers beginning in 1908, F. Hartogs showed that this statement is completely false for domains in $\boldsymbol{C}^n$. These papers raised some central questions, which are still of importance today. To state but one result of Hartogs, here is one of the most surprising:

D is taken to be a domain in $\boldsymbol{C}^n$, $n \geqslant 2$, and it is supposed that $\boldsymbol{C}^n - D$ (the points of $\boldsymbol{C}^n$ not in D) is connected. U is an open set in $\boldsymbol{C}^n$ containing the boundary of D. For any holomorphic function f on U, there is a holomorphic function F on D that equals f in some neighbourhood of the boundary of D.

Now, for a domain D in C, if a single function on D is considered, its maximum domain of existence is no longer a plane domain, but is, in a natural way, a domain spread over C; for example, the function $(1 + z)^{1/2}$ in the disc $D = \{|z| < 1\}$. It turns out that a similar situation arises when one considers the family of all holomorphic functions on a domain in $\boldsymbol{C}^n$, $n \geqslant 2$.

This is made precise as follows: X denotes a Hausdorff topological space and ϕ: $X \to \boldsymbol{C}^n$ is a continuous map. The pair (X, ϕ) is called a domain over $\boldsymbol{C}^n$ if each point of X has a neighbourhood U such that ϕ is a homeomorphism of U onto an open set in $\boldsymbol{C}^n$, and if, in addition, X is connected. Then X carries, naturally, the structure of a complex manifold (in this case it is only necessary to take for the U_k in the definition of a complex manifold sufficiently small open sets in X, and to take for the ϕ_k simply the map ϕ).

The central theorem, proved first by Peter Thullen of Switzerland, can be stated as follows:

For a given domain D in $\boldsymbol{C}^n$, there is a domain (X, ϕ) over $\boldsymbol{C}^n$ containing D as an open subset such that (1) for any holomorphic function f on D, there is a unique holomorphic function F on X the restriction of which to D is f, (2) (X, ϕ) is the largest domain over $\boldsymbol{C}^n$ with this property.

A similar statement can be proved when one starts with a domain (D, ϕ_0) over $\boldsymbol{C}^n$. These maximal domains of existence are called domains of holomorphy.

The question arises as to whether one can characterize domains of holomorphy differently. Furthermore, there is the obvious question of whether other basic theorems of one complex variable (such as the Runge approximation theorem and Magnus Gösta Mittag-Leffler's theorem that on any plane domain one can construct meromorphic functions with prescribed singularities) have analogues in several variables. In a beautiful series of papers, K. Oka recognized that all these problems are related, and gave very satisfactory solutions. Oka's papers have probably influenced most of the papers written on these questions since.

One of Oka's fundamental theorems, special cases of which had been obtained earlier by Henri Cartan and Peter Thullen, may be stated as follows:

A domain (X, ϕ) over $\boldsymbol{C}^n$ is a domain of holomorphy if, and only if, for any infinite sequence $x_1, x_2, \cdots$, of points on X that has no point of accumulation in X, there is a holomorphic function f on X for which the sequence of complex numbers $f(x_1), f(x_2), \cdots$ is not bounded.

Oka gave other characterizations of domains of holomorphy; these, and their generalizations to analytic spaces are of the greatest importance. To describe one of these, the following notation is introduced: If $(z_1, \cdots, z_n)$ stands for a point in $\boldsymbol{C}^n$ and $z_j = x_j + iy_j$, in which x_j and y_j are real, then, for a differentiable function u defined on an open set in $\boldsymbol{C}^n$ define two partial differential operators (see 35).

The following theorem of Oka answers questions raised many years earlier by F. Hartogs and E.E. Levi.

A domain D in $\boldsymbol{C}^n$ is a domain of holomorphy if, and only if, there exists a twice continuously differentiable real-valued function u on D, which tends to infinity at every boundary point of D and has the property that an inequality holds (see 36) for any n-tuple of complex numbers $\lambda_1, \cdots, \lambda_n$. This result bears a striking resemblance to the following geometric result.

Recall that a domain in R^n, the space of n-tuples of real numbers, is called convex if, whenever two points lie in the domain, so also does the line segment joining them lie in the domain.

A domain D in R^n is convex if, and only if, there is a

(32) $$\dim H^0(V, F) - \dim H^1(V, F) = c(F) - g + 1$$

(33) $$\dim H^0(V, F) \geqslant c(F) - g + 1$$

(34) $$\chi(V, F) = \dim H^0(V, F) - \dim H^1(V, F) + \cdots + (-1)^n \dim H^n(V, F)$$

(35) $$\frac{\partial u}{\partial z_j} = \tfrac{1}{2}\left(\frac{\partial u}{\partial x_j} - i\frac{\partial u}{\partial y_j}\right), \qquad \frac{\partial u}{\partial \bar{z}_j} = \tfrac{1}{2}\left(\frac{\partial u}{\partial x_j} + i\frac{\partial u}{\partial y_j}\right)$$

(36) $$\sum_{k,l=1}^{n} \frac{\partial^2 u}{\partial z_k \partial \bar{z}_l} \lambda_k \bar{\lambda}_l \geqslant 0$$

function u on D, tending to infinity at each boundary point and such that an inequality formally similar to the above holds (see 37) for each n-tuple of real numbers $\lambda_1, \cdots, \lambda_n$.

Thus domains of holomorphy may be looked upon as complex analogues of convex domains.

The way in which classical theorems on functions of one complex variable generalize will now be described.

Runge's theorem. D is a domain of holomorphy in $\boldsymbol{C}^n$, and K a closed bounded subset of $\boldsymbol{C}^n$ contained in D. It is supposed that to every point a in D, not in K, it is possible to find a holomorphic function f on D with $|f(a)| > |f(z)|$ for any z in K. Then:

Let g be a function holomorphic on some open set containing K, and let $\varepsilon > 0$. Then, there is a function G holomorphic on D so that $|G(z) - g(z)| < \varepsilon$ for any z in K. When $D = \boldsymbol{C}^n$, one can take for G a polynomial in $z_1, \cdots, z_n$.

This was proved by André Weil of France for $n = 2$, by Oka in general.

The Cousin problems. Toward the end of the 19th century, a French mathematician, P. Cousin, generalized the classic theorems of Weierstrass and Mittag-Leffler to several variables. Again Oka solved these problems in great generality and made further progress possible. To describe these problems, meromorphic functions must be used: these are objects which, in the neighbourhood of any point, can be expressed as a quotient of two holomorphic functions with a denominator which is not identically zero near the given point (see above *Complex line bundles; complex vector bundles*).

The first Cousin problem. Given a domain D in $\boldsymbol{C}^n$, to each point a in D is assigned a small ball $|z - a| < r$, which is called B_a. It is supposed that for each a, a meromorphic function f_a in B_a is given. It is assumed that whenever two of these balls B_a, B_b meet, then f_a and f_b have the same singularities in $B_a \cap B_b$; that is, that $f_a - f_b$ is holomorphic on $B_a \cap B_b$.

Does there exist a meromorphic function f on D that has the same singularities as f_a in B_a for each a; that is, $f - f_a$ is holomorphic on B_a?

The second Cousin problem. D is a domain. B_a is a small ball about each point a in D. f_a is a holomorphic function on B_a. It is assumed that whenever two of the balls B_a, B_b meet, f_a and f_b have the same zeros in $B_a \cap B_b$; that is, there is a holomorphic function g on $B_a \cap B_b$ without zeros so that $f_a = gf_b$.

Does there exist a holomorphic function f on D having the same zeros as f_a in B_a for each a; that is, $f = hf_a$ in which h is holomorphic and has no zeros on B_a?

Here are Oka's solutions.

If D is a domain of holomorphy, the answer to the first Cousin problem is always: Yes.

The answer to the second, even if D is a domain of holomorphy, is: No, in general. One necessary and sufficient condition that it be solvable, however, is that there exist a continuous function having the same zeros as f_a in B_a.

This reduction of a problem concerning holomorphic objects to one concerning merely continuous objects has proved extremely fruitful, and has become a guiding principle in several questions in complex analysis.

Oka recognized that all the results described above are intimately related to the so-called global ideal theory. Here is a typical problem in this theory and a typical theorem.

Problem. Given a domain D of holomorphy and $f_1, \cdots, f_q$, holomorphic functions on D, if g is a holomorphic function on D, when do there exist holomorphic functions $\phi_1, \cdots, \phi_q$ so that g is a sum of function products (see 38) everywhere on D?

Theorem. This is the case if, and only if, for any point of D, one can find functions $\alpha_1, \cdots, \alpha_q$ holomorphic on some neighbourhood of that point such that g is a sum of products involving the α's (see 39) in that neighbourhood. In particular, if $f_1, \cdots, f_q$ have no common zeros, this is solvable for all g.

In proving these theorems, Oka was led to investigate solutions of a homogeneous equation (see 40) and thence to introduce objects that have led to the theory of coherent analytic sheaves. Cartan, partly in collaboration with the French mathematician Jean-Pierre Serre, used these sheaves to formulate, generalize, and solve corresponding problems on Stein spaces. These latter are far-reaching abstract generalizations of domains of holomorphy. This work has proved to be an enormously powerful and versatile tool in complex analysis and elsewhere.

DEFORMATIONS OF COMPLEX STRUCTURES

Deformation of the complex structure of a compact Riemann surface (1-dimensional complex manifold) is an idea dating back to Riemann who, in 1857, calculated the number of independent parameters on which the deformation of the structure of a compact Riemann surface depends and called it the number of moduli. The questions centring around the deformation of the complex structure of a Riemann surface have never lost their interest. The theory of quadratic differentials and extremal quasi-conformal maps was developed by Teichmüller (1938–44) in order to give a systematic theory of the deformation of Riemann surfaces. Teichmüller's methods were later perfected by Lars V. Ahlfors (Finnish-born), E. Rauch, Lipman Bers (Latvian-born), André Weil (French), and others.

The deformation of the complex structure of higher-dimensional complex manifolds seems to have been considered first in 1888 by Noether, who calculated the number of moduli for a class of algebraic surfaces in projective 3-space. In contrast to the case of complex dimension 1, however, the deformation of the structure of higher-dimensional complex manifolds was neglected during the 70 years following Noether's work until the subject was taken up by A. Frölicher and Albert Nijenhuis (U.S.) and by Kodaira and Spencer.

Before beginning a description of the ideas of Kodaira and Spencer, a few remarks are in order:

A subset K of an analytic space is called compact if any infinite sequence of points in K has a convergent subsequence. For any sets A, B, $A \times B$ is used to denote the set of ordered pairs (a, b), in which a runs over the elements of A and b over those of B.

If A and B are analytic sets in open sets U and V in $\boldsymbol{C}^m$ and $\boldsymbol{C}^n$ respectively, then $A \times B$ is an analytic set in $U \times V$. It follows that if X and Y are analytic spaces, then $X \times Y$ has a natural structure of analytic space.

A holomorphic mapping $\pi\colon X \to S$ is considered in which X and S are analytic spaces; s_0 is a fixed point in S. The part of X lying over s_0 [that is, the set of points x of X with $\pi(x) = s_0$], is an analytic space X_0. This is the analytic space the deformations of which are being considered. For any s in S, X_s is taken to be the analytic space of points of X lying over s. It is necessary to think of the family X_s as representing a deformation of X_0 $(= X_{s_0})$. For this, the X_s should represent the same topo-

$$\text{(37)} \qquad \sum_{k,l=1}^{n} \frac{\partial^2 u}{\partial x_k \partial x_l} \lambda_k \lambda_l \geq 0$$

$$\text{(38)} \qquad g = \phi_1 f_1 + \cdots + \phi_q f_q$$

$$\text{(39)} \qquad g = \alpha_1 f_1 + \cdots + \alpha_q f_q$$

$$\text{(40)} \qquad \phi_1 f_1 + \cdots + \phi_q f_q = 0$$

$$\text{(41)} \qquad \begin{cases} \phi\colon Y \to Y_s \times U \\ \psi\colon Y_s \times U \to Y \end{cases}$$

$$\text{(42)} \qquad \theta_{ik}(t) = (\theta^1_{ik}(t), \cdots, \theta^\alpha_{ik}(t), \cdots, \theta^n_{ik}(t))$$

$$\text{(43)} \qquad \theta^\alpha_{ik}(t) = \theta^\alpha_{ik}(z_i, t) = \partial h^\alpha_{ik}(z_k, t)/\partial t$$

logical space as X_0. Also, the case in which X_0 is compact should be considered. Compactness is easily ensured by the requirement that π be proper; that is, for any compact subset K of S, the set of x in X for which $\pi(x)$ lies in K should be compact. An excellent way to ensure the former, as well as to describe the complex structure of X_s as being close to that of X_0, turns out to be that given in the following definition.

A deformation of the analytic space X_0 is a proper holomorphic map $\pi: X \to S$, S being connected and provided with a fixed point s_0 such that the part of X lying over s_0 is X_0 and such that the following holds:

For any point x in X, let $\pi(x) = s$. There exist neighbourhoods Y, Y_s of x in X, X_s, respectively, and one U of s in S such that Y can be identified with $Y_s \times U$ in such a way that π becomes the map sending (y, u) in $Y_s \times U$ onto u. More precisely: there are holomorphic maps ϕ, ψ (see 41) so that $\psi[\phi(y)] = y$ for any y in Y, $\phi[\psi(z)] = z$ for any z in $Y_s \times U$; and so that if $\phi(y) = (y_s, u)$, then $u = \pi(y)$.

Kodaira and Spencer introduced, when X and S are complex manifolds, a first order approximation of the deformation $\pi: X \to S$ at the point s_0, called the associated infinitesimal deformation; this is an approximation roughly in the same way that the partial derivatives of a function provide a first order approximation to the function at a point. They showed that many properties of the deformation in the vicinity of the point s_0 can be recovered from properties of the infinitesimal deformation. The advantage is that this latter is much more amenable to algebraic techniques.

These results and methods have been carried over to the case of arbitrary analytic spaces and local methods introduced that enable one to decide when a given holomorphic map is a deformation in the sense defined above.

A more or less systematic theory of deformation of the complex structure of higher-dimensional compact complex analytic manifolds given as stated above by Kodaira and Spencer, is now described briefly. They defined a differentiable family of compact complex manifolds as a fibre space $\mathfrak{F}$ over a connected differentiable manifold M the fibre V_t of which over the point t of M has a complex structure depending differentiably on the point t. Assume, for simplicity, that M is one-dimensional and let U be a sufficiently small interval containing the point $t = 0$ in its interior. Then there exists a covering $\{\mathfrak{F}_i\}$ of the portion of $\mathfrak{F}$ over U by open sets $\mathfrak{F}_i$ where $\mathfrak{F}_i$ is covered by the coordinates $(z_i, t) = (z_i^1, \cdots, z_i^n, t)$. Denote by $z_i = h_{ik}(z_k, t)$ the biholomorphic coordinate transformations that determine the complex structure of V_t; that is, $z_i^\alpha = h_{ik}^\alpha(z_k, t)$, in which α may range from 1 to n. Letting a notational device be used (see 42) in which each new term is defined by partial differentiation (see 43), a 1-cocycle $\{\theta_{ik}(t)\}$ is obtained that defines a cohomology class $\theta(t)$ of the 1-dimensional cohomology $H^1(V_t, \Theta_t)$ of V_t with coefficients in the sheaf Θ_t over V_t of holomorphic vector fields. The class $\theta(t)$, which is called the infinitesimal deformation of V_t, measures the dependence of the complex structure of V_t on the parameter t. In the case in which the fibres V_t are compact Riemann surfaces, $H^1(V_t, \Theta_t)$ is isomorphic to the vector space of quadratic differentials of the surface, in the sense of Teichmüller.

Proceeding in this fashion, a generalization of Riemann's concept of number of moduli was obtained. In contrast to the case of Riemann surfaces, however, this number is not defined for all compact complex manifolds. Particular examples of complex manifolds V_0, of complex dimension exceeding 2, show that not all cohomology classes of $H^1(V_0, \Theta_0)$ are the infinitesimal deformations of actual deformations of the structure of V_0. In fact, some classes of $H^1(V_0, \Theta_0)$ may be obstructed, the obstructions occurring as classes in $H^2(V_0, \Theta_0)$. If $H^2(V_0, \Theta_0) = 0$, then any class of $H^1(V_0, \Theta_0)$ is the infinitesimal deformation corresponding to a deformation of finite magnitude.

A classical problem in algebraic geometry concerning the completeness of characteristic systems is to show that all infinitesimal displacements correspond to finite deformations of a complex analytic submanifold of a fixed algebraic manifold. A condition introduced by Severi for the case of curves in algebraic surfaces was reformulated by Kodaira and Spencer in the language of sheaves and shown to be sufficient for completeness; *i.e.*, for the existence of a finite deformation of the submanifold corresponding to any given infinitesimal displacement of it.

Finally, the Italian-born U.S. mathematician E. Calabi and Eduardo Vesentini of Italy, in a paper published in 1960, proved that a compact complex manifold, which is uniformized by an irreducible Cartan domain of at least two complex dimensions, is (locally) rigid; *i.e.*, the number of its moduli is equal to zero. This result again underscores the difference between Riemann surfaces and complex manifolds of higher dimension.

$$\sigma(\tau(x)) = \tau(\sigma(x)) = x \tag{44}$$

$$\sigma \circ \tau(x) = \sigma(\tau(x)) \tag{45}$$

Returning to the problem of classifying complex structures on X_0, Kodaira and Spencer formulated the local problem and solved it in an important case. This was later done for an arbitrary compact complex manifold X_0 by K. Kuranishi. First, a preliminary remark must be made:

$\pi: X \to S$ is taken to be a deformation, X_0 the part of X over the point s_0 of S. S' is any analytic space, s'_0 a point of S' and $\phi: S' \to S$ a holomorphic map taking s'_0 to s_0. Then it is possible to construct a deformation $\pi': X' \to S'$ in which the part of X' lying over s' in S' is simply X_s, in which $s = \phi(s')$. The family $\pi': X' \to S'$ is said to be induced from $\pi: X \to S$ by the map ϕ.

Kuranishi's theorem is:

Given a compact complex manifold X_0, there exists a deformation $\pi: X \to S$ and s_0 in S so that X_0 is the part of X lying over s_0 and that has the following universal property:

If $\pi': X' \to S'$ is any deformation of X_0, X_0 being the part of X' over s'_0, then there is a neighbourhood U' of s'_0 in S' and a holomorphic map $\phi: U' \to S$ so that $\phi(s'_0) = s_0$ and the part of X' lying over U' is simply the family induced from $\pi: X \to S$ by the map ϕ. In a sense, this universal family contains all complex structures on X_0 close to the original structure.

Even in the case when X_0 is a complex manifold, the base space S of the universal family is not, in general, a complex manifold. This is only one of the many reasons for introducing analytic spaces.

Despite the efforts of many people, the corresponding problem for an arbitrary compact analytic space remains unsolved. This is certainly among the most important questions in complex analysis.

If one also admits for consideration non-compact analytic spaces, the problem becomes much more complicated, and very little indeed is known. It has turned out, however, that the general case holds the key to many questions; for instance, in the theory of holomorphic foliations (work initiated by the Swiss mathematician André Haefliger).

AUTOMORPHISMS OF ANALYTIC SPACES AND AUTOMORPHIC FUNCTIONS

If X is an analytic space and $\sigma: X \to X$ a holomorphic map of X into itself, then σ is called an automorphism of X if there is a holomorphic map $\tau: X \to X$ such that it acts formally like the inverse of σ (see 44) for all x in X. τ is then unique. It is called the inverse of σ and is denoted by σ^{-1}.

The set of all automorphisms of X forms a group: if σ and τ are automorphisms, so is the map $\sigma \circ \tau$ defined by composition (see 45). The map $1: X \to X$ defined by $1(x) = x$ for all x in X is an automorphism and $\sigma \circ 1 = 1 \circ \sigma$ for any automorphism σ. This group is denoted by $\mathrm{Aut}(X)$.

Γ denotes a subgroup of Aut(X); that is, a subset such that whenever two automorphisms σ, τ belong to Γ, so do $\sigma \circ \tau$ and σ^{-1}. For any subset E of X and any σ in Aut(X), $\sigma(E)$ denotes the set of points $\sigma(x)$ when x runs over the elements of E.

The subgroup Γ of Aut(X) is said to be properly discontinuous—or to act properly—if any point of X has a neighbourhood U so that there are only finitely many σ in Γ for which $\sigma(U)$ meets U. If this is the case, one looks, for any point x of X, at the set Γx [the set of $\sigma(x)$ when σ runs over Γ]; this is called the orbit of the point x under Γ. If two orbits meet, then they are identical, and so the orbits provide a partition of X into mutually exclusive sets.

A new set $Y = X/\Gamma$ is defined by taking as its elements the orbits of points in X. If Γ acts properly, Y can be made into a Hausdorff topological space by taking for its open sets those subsets U so that the union of the orbits represented by points of U is open in X. The map $\eta\colon X \to Y$ that associates to each point of X the orbit of that point is continuous.

The first basic result is due to Cartan: $Y = X/\Gamma$ carries a natural structure of an analytic space; the map $\eta\colon X \to Y$ is holomorphic. Further, a function f on an open set U in Y is holomorphic if, and only if, the function $f \circ \eta$ is holomorphic on the open set of points x in X for which $\eta(x)$ lies in U. Thus holomorphic (or meromorphic) functions on Y are those holomorphic (or meromorphic) functions on X that are invariant under Γ; that is, $f(\sigma(x)) = f(x)$ for any σ in Γ.

Various generalizations of this theorem of Cartan are known.

In what follows, it is supposed that X is a bounded domain in some $\boldsymbol{C}^n$.

Given a bounded domain X in $\boldsymbol{C}^n$, and a subgroup Γ of Aut(X) acting properly on X, one would like to construct meromorphic functions on $Y = X/\Gamma$; that is, functions on X that are invariant under Γ. Poincaré had the very fruitful idea of constructing first holomorphic functions on X that are less restricted, and showed how to construct these using certain infinite series.

If f is any nonzero holomorphic function, and σ is in Γ, $j_\sigma(x)$ is taken to be the meromorphic function on X defined by a quotient (see 46); $j_\sigma(x)$ measures the extent to which f is not invariant under Γ. The function $j_{\sigma\tau}$ corresponding to the composite $\sigma\tau$ can be expressed in terms of j_σ and j_τ (see 47). A family of holomorphic functions $\{j_\sigma\}$ satisfying this relation is called a factor of automorphy relative to Γ. If f, g are two holomorphic functions giving rise to the same factor of automorphy, the quotient f/g is a meromorphic function invariant under Γ. Thus the question arises as to whether, given a factor of automorphy, one can construct several functions f satisfying the relation above defining the factor on X; such an f is called an automorphic form relative to the factor of automorphy $\{j_\sigma\}$.

The most useful factor of automorphy is the following. X is a bounded domain, and σ is in Aut(X). For z in X, $\sigma(z)$ is an n-tuple (see 48). For any positive integer m, a factor of automorphy $\{j_\sigma{}^{(m)}\}$ is the mth power of the determinant of $\left(\frac{\partial \sigma_i}{\partial z_j}\right)$. Using the Poincaré series referred to above, Cartan showed that one can construct automorphic forms with prescribed values and derivatives at a point of X relative to the factor $\{j_\sigma{}^{(m)}\}$ for suitable values of m. This theorem gives great insight into the structure of the set of meromorphic functions on X/Γ when this space is compact.

In connection with his arithmetic and analytic study of quadratic forms, Carl Ludwig Siegel constructed many examples of the above (with X/Γ compact). Siegel's researches, however, also produced many beautiful examples in which X/Γ is not compact; these examples are of the greatest importance in many branches of mathematics. A systematic study of these examples has led to the theory of arithmetic subgroups of algebraic groups, which is being developed intensively. This theory is at the meeting place of a number of different branches of mathematics: complex analysis, differential geometry, number theory, algebraic geometry, Lie groups and algebras, and many others as well. (R.Na./Ed.)

$$(46)\qquad f(\sigma(x)) = j_\sigma(x) f(x)$$

$$(47)\qquad j_{\sigma\circ\tau}(x) = j_\sigma(\tau(x)) \cdot j_\tau(x)$$

$$(48)\qquad \sigma(z) = (\sigma_1(z), \cdots, \sigma_n(z))$$

$$(49)\qquad \Delta u = \sum_{i=1}^{3} \frac{\partial^2 u}{\partial u_i^2}$$

Potential theory

One of the most impressive examples of the interrelationship of theory and application of complex analysis is found in potential theory. Originally a part of physics, the body of knowledge surrounding the concept of the potential has been raised to a level of abstract theory by the contributions of many mathematicians. Since the mid-19th century, potential theory has gradually come to be studied as a part of complex ananlysis.

HISTORICAL SURVEY OF PROBLEMS AND CONCEPTS

Until about 1800, potential theory was actually that part of theoretical physics dealing with electrostatics (see ELECTRICITY) and Newtonian attraction between masses. The potential of a unit mass concentrated at a point X is equal at any other point x to the inverse $\frac{1}{|X-x|}$ of the distance $|X-x|$ to X. It is a harmonic function of $x (x \neq X)$; that is, it is equal at x to its mean value on any ball centred at x and not containing X, or, equivalently, it satisfies Laplace's equation (see 49), a second-order partial differential equation (see DIFFERENTIAL EQUATIONS). The potential of a general distribution of masses is then defined by multiplication and summation (actually by an operation of integration) and it is harmonic outside of the masses. Thus, physical problems became mathematical ones.

Important results due to the mathematicians Siméon-Denis Poisson (French) and George Green (British) were followed in 1840 by a fundamental work of Carl Friedrich Gauss (German), amazingly deep and rich, in which he solved for Euclidean space three connected problems that remain essential subject to various adaptations.

The first one is the problem of equilibrium; this is the determination of a distribution of a given mass on a conductor for which the potential is constant on it; this corresponds to minimizing the (potential) energy.

Given masses inside a closed conductor, the second problem asks how they can be distributed on the conductor so as to give the same potential outside. This is solved physically by the phenomenon of influence that creates on the conductor (on which one makes the potential zero) a distribution of masses such that the total potential is zero outside. Similar problems arise by commuting (interchanging) interior and exterior. This operation was later called *balayage* (sweeping-out process) by Poincaré.

The third problem is the determination in a domain of a harmonic function the limiting values of which agree on the boundary with a given function (first supposed to be finite continuous). This standard boundary-value problem was called the Dirichlet problem (after the German mathematician Peter Gustav Lejeune Dirichlet) by Riemann.

For lack of sufficient mathematical tools, the solutions given by Gauss were not rigorous and the results not quite correct or precise; actually they needed some restrictions. Because of these difficulties, the first two problems were put aside for a long time, but the Dirichlet problem, more important, was studied by different methods, which at first were not rigorous either. For example, an important method was based on the minimizing of a

so-called Dirichlet integral (Dirichlet principle). This was made rigorous by David Hilbert of Germany only around 1900. If so many great mathematicians gave different solutions, it is perhaps also because the restrictions on the boundary seemed unnecessary; actually they were necessary, as S. Zaremba of Poland and Henri-Léon Lebesgue, a French mathematician, noted after 1900.

Some other questions considered in this period include the analytic character of harmonic functions (expansion as power series), other types of boundary value problems (such as the Neumann problem, named for Karl Gottfried Neumann of Germany, for given normal derivatives), the connections of harmonic functions with functions of a complex variable, conformal mapping, Fourier series, similar developments for equations more general than the Laplace one, polyharmonic functions or more precisely harmonic functions of order p, satisfying to $\Delta^p u = \Delta(\cdots \Delta u) = 0$ with $p \mid \Delta$-operators.

A second period of the development of potential theory was characterized chiefly by the introduction, in the study of potential theory, of wider notions of measure and integral (those of Lebesgue and J.R. Radon), and of capacity, a notion inspired by electrostatics, first clearly defined and discussed by Norbert Wiener of the United States and Charles-Jean de la Vallée Poussin of Belgium, and greatly generalized much later by the French mathematician Gustave Choquet.

The use of integrals allowed a new development of the subject, chiefly by means of a larger definition of the potential of a general distribution of masses, viz., a measure μ. This potential is defined by the integral of a function N integrated with respect to y and to a measure μ (see 50) in which $N(x, y)$, called a kernel, is $\frac{1}{|x-y|^{n-2}}$ in the space $\mathfrak{R}^n$ with $n \geqslant 3$ (Newtonian kernel) and is $\log \frac{1}{|x-y|}$ in $\mathfrak{R}^2$ (logarithmic kernel).

Actually, any kernel of the form $N(|x-y|)$ can be used to define generalized potentials if $N(s)$ is a generating function of a mean value in the sense of Kolmogorov and Nagumo; s^{2-n} and $\log(1/s)$ are such generators. Integrals of this type with elliptic or Lorentzian metric were used by Marcel Riesz, a Hungarian-born Swedish mathematician, for the study of Laplace's and d'Alembert's equations in the 1930s and early 1940s. The solutions involve parameters, and the composition properties of the kernels lead to semi-groups of operators (see 77). The occurrence of such groups or semi-groups in the solutions of physical problems was first noted in 1924 by J. Hadamard of France and was regarded by him as a consequence of what he called the major premise of Huygens' principle.

A new setting and a generalized solution of the Dirichlet problem appeared (resulting from a remark of Lebesgue and results of Wiener), avoiding any restriction on the boundary; the boundary behaviour of the new solution and its relation with the previous one were made clear later by O.D. Kellogg and Griffith C. Evans of the U.S., thanks to the notion of capacity, and a more general problem with discontinuous boundary data was solved by Marcel-Émile Brelot of France.

New important tools, the concepts of subharmonic and superharmonic functions, were introduced by Frigyes Riesz, a Hungarian mathematician. These functions stand in relation to harmonic functions as arcs of convex and concave curves, respectively, do to their chords. They are locally equal modulo harmonic functions to potentials of negative and positive masses, respectively, in the general new sense.

A rigorous and more precise development of the ideas of Gauss was made in the famous thesis of the Swedish mathematician Otto Frostman (1935). A definitive convergence theorem on decreasing potentials of positive masses, given by the French mathematicians Marcel-Émile Brelot and Henri Cartan, was an essential tool for all refinements in potential theory. The sets of (outer) capacity zero emerged as the standard exceptional sets in potential theory and in its application to the functions of a complex variable.

(50) $\int N(x, y)\, d\mu(y)$

The third period from about 1940 to 1955 was marked by the role of topology in potential theory.

The notion of energy in Euclidean space was used more deeply by Henri Cartan. The measures were considered as points of a space, with the distance to the origin equal to the square root of the energy corresponding to the measure: in such a space the well-known methods of Hilbert space theory (geometry in a space with infinitely many coordinates) were applied, and a striking and powerful theory was developed. Later J. Deny continued this theory of finite energy, using, instead of measures, the more general distributions of the French mathematician Laurent Schwartz, but preserving the essential ideas.

Another example of the role of topology is the introduction of the notion of a thin set and equivalently of the fine topology in which all classical potentials become continuous. This notion due to Brelot-Cartan is adapted to the study of discontinuous potentials and nonsmooth boundaries and gives rise to new problems.

The most striking example is the introduction of the Martin boundary. The solution of the Dirichlet problem in a ball had led to the (integral) representation of a positive harmonic function in the ball by means of a positive measure on the boundary. A similar problem for a general Euclidean domain remained unsolved until 1941 when Martin introduced a quite new boundary and its useful minimal part. This work later inspired Choquet's profound study of extreme points of a compact convex set, which is one of the great discoveries in analysis after mid-20th century.

This period also marks the introduction in potential theory of spaces much more general than Euclidean spaces or certain manifolds and of more general kernels N than the Newtonian or logarithmic ones, N a suitable function of two variables (and even of more general concepts). Around 1940 certain basic properties of the classical theory were introduced as axioms in suitably chosen topological spaces and a discussion began of various generalizations and interactions between different axioms or principles.

Since about 1955 topological refinements and new axiomatizations have been made and the role of probability introduced. Fine topology was extended to the Martin boundary, and it became an essential tool for the study of the boundary behaviour of functions connected with potential theory.

Rather than speaking about the further developments with kernels $N(x, y)$, an axiomatic theory is sketched that is not based on such kernels. Given real-valued continuous functions are used, called harmonic functions, defined on every domain of a suitable topological space and satisfying some well-known properties of the classical theory that are considered as axioms. It was found possible to imitate the classical developments. In this direction, after some attempts, Joseph Leo Doob of the U.S. constructed a better base of an axiomatic theory but for probabilistic purposes. This inspired a detailed potential theory by Brelot and followers such as the French mathematician Rose-Marie Hervé. The solutions of partial differential equations of elliptic type more general than that of the Laplace equation satisfy the basic axioms under some restrictions. But it is not applicable to equations of a more general type like the classical heat equation. In order to include such equations, Heinz Bauer of West Germany and others extended the previous theory by weakening the axioms. Later the systematic search for all theories satisfying such local axioms was partly solved by Bony.

On the other hand, the concepts of energy and Dirichlet integral led Deny and the Swedish-born U.S. mathematician Arne Carl-August Beurling to another axiomatic theory of global character, called the theory of Dirichlet spaces (for complex functions). Problems similar to those of equilibrium and balayage were studied.

These axiomatic theories, including interesting cases of the last one, are actually contained in a fundamental probabilistic theory developed by Gilbert Hunt between 1957 and 1958. A function of a point x and of a set e, called a kernel $V(x, e)$, which transforms a nonnegative function f into a nonnegative function Vf, is considered. Under certain conditions, this V may be expressed by means of similar kernels of a certain family $\{P_t\}_{t\geq 0}$ by means of which the so-called excessive functions can be defined. These majorize their P_t-transforms. This $\{P_t\}$ has an essential probabilistic interpretation in terms of "Markov processes," named for the Russian mathematician Andrey Andreyevich Markov, of which the Brownian motion of physics is a particular case.

It was shown by P.A. Meyer and others that, given any of the above local axiomatic theories, it is possible to find a V for which the excessive functions of the corresponding $\{P_t\}$ are precisely the nonnegative hyperharmonic functions.

Various probabilistic notions like trajectories (of a particle) or martingales led to probabilistic interpretations of concepts and results of the above classical or axiomatic theories. The relations between probability and potential theory, known for a long time as "folklore" in the classical theory, deeply studied by Doob then Hunt and many others, have now produced a new and rich chapter in analysis. The latest researches are also concerned with a converse problem: given a family of functions, under what conditions is it possible for one to interpret them as potentials or superharmonic functions in the above local axiomatics or excessive functions in a global extension of Hunt theory?

BASIC CONCEPTS AND RESULTS OF CLASSICAL POTENTIAL THEORY

Using more advanced mathematical tools, the most important points of classical theory will now be outlined. A Euclidean space $\mathfrak{R}^n$ ($n \geqslant 2$) will be considered first, the case $n = 1$ being rather trivial.

Harmonic functions. A real finite continuous function u on an open set ω is said to be harmonic, if it is equal at each point x to its mean value on any ball (that is, a disc in $\mathfrak{R}^2$) $B_x{}^r$ (centre x, radius r), or equivalently (using the corresponding Lebesgue measure), on the boundary $\partial B_x{}^r (\bar{B}_x{}^r \subset \omega)$. This implies that u is C^∞ (infinitely differentiable) and, further, is equivalent to saying that $\Delta u = 0$ (local condition). Also, $\Delta u = 0$ in the sense of Schwartz is equivalent to saying that u is a function equal almost everywhere to a harmonic function.

The definition implies that a harmonic function that attains its maximum (or minimum) at a point x_0 is constant on a neighbourhood. This maximum or minimum principle is a key tool. A form that will be valid later in more general theories may be stated in the following manner: if at any boundary point (which may be the Aleksandrov point [named for the Soviet mathematician Aleksandr Danilovich Aleksandrov] at infinity) lim inf $u \geqslant 0$, then $u \geqslant 0$.

This implies that for a finite boundary function f the Dirichlet problem possesses at most one solution. The question of existence is more difficult. For a ball $B = B_0{}^R$ and a finite continuous function f, the solution is given by the Poisson integral (see 51) integrated with respect to the measure σ (σ being the normalized measure on ∂B proportional to angle or solid angle). For any f, $\bar{I}_f$ will denote the upper-integral which is $+\infty$, $-\infty$, or harmonic. Note that by changing $f\,d\sigma$ into any Radon measure $d\mu$ on ∂B, the corresponding $I_\mu(y)$ (called a Poisson-Stieltjes integral) is harmonic in B and represents in the general case the difference of two nonnegative harmonic functions.

The Poisson integral I_f implies many properties: a local criterion (by considering in the definition for any x only those $\partial B_x{}^r$ with $0 < r < r(x)$ an arbitrary function); analyticity (and various power series); the Harnack inequalities: u harmonic > 0 in $B_0{}^R$ implies upper and lower bounds for the ratio of u at y to u at zero (see 52), which in turn implies the equicontinuity at any point of any family of positive-bounded harmonic functions.

(51) $$I_f^B(y) = \int \frac{R^2 - |y|^2}{|x-y|^n} R^{n-2} f(x)\, d\sigma(x)$$

(52) $$R^{n-2}\frac{R-|y|}{(R+|y|)^{n-1}} \leqslant \frac{u(y)}{u(0)} \leqslant R^{n-2}\frac{R+|y|}{(R-|y|)^{n-1}}$$

(53) $$u(y) = H_1(y) + |y|^2 H_2(y) + \cdots + |y|^{2p-2} H_p(y)$$

(54) $$\overline{B}_x^r(w), \qquad u \geqslant \bar{I}_u^{B_x^r}$$

(55) $$h_x(y) - I_{h_x}^B(y) = G(x, y)$$

(56) $$\int G(x, y)\, d\mu(x), \qquad \mu \geqslant 0$$

(57) $$\int G(x, y)\, d\mu(x) + u^*(y)$$

(58) $$G(x, y) = |x-y|^{2-n}$$

(59) $$\int G(x, y)\, d\mu(x)$$

It should be emphasized that uniform convergence preserves harmonicity and implies local uniform convergence for any derivative, that in a domain (connected open set) an increasing sequence u_n of harmonic functions tends to $+\infty$ or to a harmonic function (uniformly locally in an obvious sense), and that there is an extension to an increasing directed family.

Polyharmonic functions. Definitions, by iteration of Δ, were previously given. Extensions were made systematically of many properties of harmonic and subharmonic functions. The old development of Almansi is mentioned for a function of ε^{2p} type, harmonic of order p in a domain containing 0, by means of harmonic functions H_i (see 53).

Sub- and superharmonic functions. In an open set ω, u is said to be hyperharmonic if: (i) u is lower semicontinuous; (ii) $u > -\infty$; (iii) for each ball contained in an open set, u has a specific lower bound (see 54; equivalently, only at x and for sufficiently small r; a local definition). In a domain, such a function is either identically $+\infty$ or finite on a dense set and in the first case is called superharmonic ($-u$ is called hypoharmonic and subharmonic, respectively).

The previous minimum-principle is still valid. Any set of superharmonic functions $\geqslant 0$ has a greatest harmonic minorant. The set of the harmonic functions $\geqslant 0$ is a complete lattice for the natural order. The set of the superharmonic functions $\geqslant 0$ is a complete upper semilattice for the natural order and also for the specific order defined by setting $u > v$ if $u = v +$ nonnegative superharmonic function.

Riesz representation. Any superharmonic function u is locally equal to $\int h(|x-y|)d\mu(x)$ plus a harmonic function, for a suitable measure $\mu \geqslant 0$.

For a ball B a specific function symmetric in x and y (see 55) is called the Green function. An integral of the Green function integrated with respect to x and μ (see 56) in which nonnegative μ is a Radon measure on B is superharmonic or $+\infty$. Any superharmonic function u on B with a harmonic minorant is equal to such an integral plus a function u^* (see 57) in which u^* is the greatest harmonic minorant of u on B. (μ, unique in both local and global representations, is called the associated measure.)

In $\mathfrak{R}^n$ ($n \geqslant 3$), the same result is valid in the whole $\mathfrak{R}^n$ by taking a specific form of the Green function (see 58).

If u is an e^2 function the result is easy to show, the general case is less easy unless Schwartz distributions are used. It is also possible, for the global form, to use a Choquet theorem on extreme elements.

The fact that $u^* = 0$ means that u is a Green potential (see 59).

Reduced function. On a fixed open set ω, a real function $\phi \geqslant 0$ and a set e are considered. The reduced function of ϕ relative to e (denoted $R_\phi{}^e$) is defined to be the lower envelope of all hyperharmonic nonnegative functions on ω majorizing ϕ on e. If $R_\phi{}^e(x)$ is the lim inf at x of $R_\phi{}^e$ then $R_\phi{}^e$ is hyperharmonic.

When ϕ is superharmonic, $R_\phi{}^e$ is superharmonic and is called the "swept-out" or "balayaged" function of ϕ relative to e.

Polar sets and capacity. A set e in an open set ω is called polar in ω if there exists a nonnegative superharmonic function in ω equal to $+\infty$ on e. When there exists a superharmonic $v > 0$ on ω, e is polar if $R_v{}^e \equiv 0$.

Note that if a set is locally polar, then it is polar in $\mathfrak{R}^n (n \geqslant 3)$ or in any domain where there exists a nonconstant superharmonic function $\geqslant 0$.

A superharmonic function $u \geqslant 0$ on $\omega - e$ where e is closed and locally polar has a unique continuation to ω as a superharmonic function. If u is bounded and harmonic, the continuation is harmonic.

A real-valued function ζ (finite or not) defined on the class of all subsets of a Hausdorff space E is called a general capacity if: (1) ζ is an increasing function; (2) for any increasing sequence $\{e_n\}$ of subsets of E, $\zeta(e_n) \to \zeta(Ue_n)$; (3) for any decreasing sequence of compact sets K_n, a similar convergent holds expressed in terms of intersection (see 60).

A K-analytic set of E is one that is the continuous image of $K_{\sigma\delta}$ set (countable intersection of K_σ sets; that is, of countable unions of compact sets) contained in a compact space, as, for example, the ordinary-Borel sets (named for the French mathematician Émile Borel) in a space like $\mathfrak{R}^n$. The celebrated Choquet capacitability theorem asserts that any K-analytic set e contained in a K_σ set is "ζ-capacitable" (K compact $\subset e$; see 61).

An example is where E is a ball on $\mathfrak{R}^n$ $(n \geqslant 3)$.

$C(K)$ denotes the supremum of $\{\mu(K), \mu \geqslant 0$, a measure on K whose Green potential is $\leqslant 1\}$.

It equals the infinum of $\{\nu(E) | \nu \geqslant 0$, a measure on E whose Green potential is $\geqslant 1$ on $K\}$. Letting $C_*(K)$ be an innercapacity (see 62) and $C^*(e)$ be an outercapacity (see 63) then C^* is a general capacity, $C^*(e) = C_*(e)$ for any Borel set e; and for any relatively compact set e, $C^*(e)$ is the total mass of the measure associated to $R_1{}^e$ in E (which is even a potential). As a consequence of this, in the space E, the sets of outercapacity 0 are thus the polar sets.

Thin sets and fine topology on $\mathfrak{R}^n$. A set e is thin at $x_0 \notin e$ if $x_0 \notin \bar{e}$ or if there exists a superharmonic function u in the neighbourhood of x_0 such that at x_0 it is strictly less than its inferior limit (see 64). The family of complements of sets thin at x form the family of neighbourhoods of x in the so-called "fine topology" which is the coarsest one, making all superharmonic functions continuous. It is finer than the Euclidean one. In $\mathfrak{R}^n$ the definition is extended by saying that e is thin at $x_0 \in e$ if $e - \{x_0\}$ is thin at x_0.

Thinness of e at x_0 is much stronger than the condition in which σ_r is the surface measure (see 65).

Locally polar sets are characterized as thin at every point, or sets of "fine isolated" points. In $\mathfrak{R}^2$, for a set thin at x_0, there exist arbitrarily small circles centred at x_0 and not intersecting the set.

As an example of application: if u is a nonnegative superharmonic function on an open set ω, with a boundary point x_0 such that $C\omega$ is thin at x_0, then it has a fine limit at x_0, and this coincides with the Euclidean limit on $\omega \setminus e$ where e is a suitable thin set.

Dirichlet problem. ω is taken to be a bounded open set of $\mathfrak{R}^n$ and f is a given extended real-valued function on $\partial\omega$. All hyperharmonic functions v on ω are considered, satisfying $\liminf_{y \in \omega,\, y \to x} v \geqslant f(x)$ and also $< -\infty$, for every $x \in \partial\omega$. The lower envelope is a function H_f that, in every component of ω, is $+\infty$, $-\infty$, or harmonic. The similar envelope $H_f = -H_{-f}$ is $\leqslant H_f$. In case both are equal and finite, the common value is denoted by H_f (solution) and f is said to be resolutive. This happens when f is continuous. Then $H_f(y)$ has, for a fixed y, a linear form $\int f d\rho_y{}^\omega$ with a measure $\rho_y{}^\omega$ called harmonic measure. For a domain ω, resolutivity of any f is equivalent to its $\rho_y{}^\omega$ -integrability (independent of y in ω).

A boundary point x is said to be regular if, for any finite continuous f, $H_f(y)$ tends to $f(x)$ at x. This is equivalent to saying that ω is not thin at x_0. It will be shown later that the irregular points form a locally polar set, as a consequence of a powerful convergence theorem.

Green function; Green space. It is possible to define on the space $\mathfrak{R}^n$ (compactified from $\mathfrak{R}^n$ with an Alexandroff point $\mathfrak{A}$ at infinity), harmonic, hyper, or superharmonic functions (even in the neighbourhoods of $\mathfrak{A}$) by a suitable mean value property, so as to preserve the minimum principle. The Newtonian and logarithmic kernels may be adapted and the local Riesz representation is preserved with a unique associated measure.

A $\mathfrak{G}$-space is defined to be a Hausdorff space with, for each point x, a homeomorphism of an open neighbourhood V of x onto an open set of $\mathfrak{R}^n$, such that for any two such neighbourhoods say V' and V'' the images of $V' \cap V''$ under the different homeomorphisms are isometric or also if $n = 2$ in conformal (but not necessarily orientation preserving) correspondence. Examples are $\mathfrak{R}^n$ and the classical Riemann surfaces.

$$\mathfrak{C}(K_n) \to \mathfrak{C}(\cap K_n) \tag{60}$$

$$\mathfrak{C}(e) = \sup_K \mathfrak{C}(K) \tag{61}$$

$$C_*(K) = \sup\{C(K) | K \subset e, \text{compact}\} \tag{62}$$

$$C^*(e) = \inf\{C_*(\omega) | e \subset \omega, \text{open}\} \tag{63}$$

$$u(x_0) < \liminf_{x \in e,\, x \to x_0} u(x) \tag{64}$$

$$\lim_{r \to 0} \sigma_r(e \cap \partial B_{x_0}^r) = 0 \tag{65}$$

This gives, on a $\mathfrak{G}$-space, local notions of harmonic, hyper, or superharmonic functions. When there exists a nonconstant superharmonic function greater than 0, one says that $\mathfrak{G}$ is a Green space. Examples are $\mathfrak{R}^n$ $(n \geqslant 3)$ or any bounded domain of $\mathfrak{R}^n$. Then there exists a smallest superharmonic function $\geqslant 0$ whose associated measure contains the unit mass at a point x. It is a generalization $G(x, y)$, still symmetric, of the Green function of a ball. The Green potential $\int G(x, y) d\mu(x)$ ($\mu \geqslant 0$ on Ω), when not everywhere $+\infty$, is superharmonic with the greatest harmonic minorant equal to 0. The previous theory may be extended. Here it is only noted that a point at infinity (that is, one whose local image is $\mathfrak{A}$) is nonpolar for $n > 2$ and that it is suitable to define the thinness of $e \in x_0$ at x_0 with the supplementary condition that x_0 is polar. On a Green space, locally polar is equivalent to polar; one says "quasi-everywhere" for "except on a locally polar set."

The general classical convergence theorem. Given on an open set ω (in a $\mathfrak{G}$-space) a family of locally lower-bounded superharmonic functions u_i, $\inf u_i$ (which is at every point the lim inf of inf u_i) is superharmonic and differs from inf u_i on a locally polar set.

Applications in a Green space Ω: (1) A superharmonic function $u \geqslant 0$ is considered. Then $R_u{}^e$ is equal to u quasi-everywhere on e, and harmonic outside of $\bar{e}$. If u is a potential or e is relatively compact, $R_u{}^e$ is a potential. In $\mathfrak{R}^3$, for example, for a nonpolar compact set e, $R_1{}^e$ is the Newtonian potential of a measure $\geqslant 0$ on e, with value 1, quasi-everywhere on e. This gives a solution of the equilibrium problem of Gauss, but with a constant value only quasi-everywhere on e and it is unique when the potential is bounded.

(2) The set of points where a set e is thin is polar. As an example, the set of irregular points of a Dirichlet problem for a relatively compact open set ω is polar.

(3) In treating a general form of the balayage, if a measure $\mu \geqslant 0$ with Green potential V is given, the balayaged potential $\hat{R}_V^e$ is characterized by the following properties:

i. This potential is equal to V quasi-everywhere on e;

ii. The associated measure b_μ is supported by the "base" B_e of e, which is the set of points where e is not thin $(b_\mu(CB_e) = 0)$.

Dirichlet principle. Consider first a bounded open set ω of $\mathfrak{R}^n$ and a C^∞-function f on ω, with a finite Dirichlet norm (see 66) and possessing a finite continuous extension on $\partial\omega$. Then there exists a harmonic function u on ω, unique up to a constant, such that $||u - f||$ is minimum; and it is H_f^ω.

This can be generalized by considering a $\mathfrak{G}$-space Ω without points at infinity, a domain ω relatively compact in Ω and a (BLD) function f on Ω that is a function which is a limit at every point, and also a limit according to the Dirichlet norm, of a Cauchy sequence of C^∞-functions with finite norms. Let f_ω be the restriction on ω. Then the minimizing function u is the projection of f_ω in the Hilbert space of the (BLD) functions on ω defined by the scalar product $\int(\text{grad } u, \text{grad } v)\, dx$, on the complete subspace of the harmonic (BLD) functions on ω; u is also the unique harmonic function whose continuation by f is (BLD) in Ω.

Energy. First $\mathfrak{R}^n$ $(n \geqslant 3)$ or a ball of $\mathfrak{R}^n$ is considered. Inspired by the kernels $|x - y|^{-\alpha}$, of M. Riesz, Cartan introduced with the classical potential U^ν the mutual energy of positive measures (see 67), the energy (μ, μ), and then the same notions for differences of measures of finite energy using for such differences a formal development of an integral (see 68). Then the energy is $\geqslant 0$ and is zero only when $\mu_1 - \mu_2 \equiv 0$ (the principle of energy). The space $\mathfrak{G}$ of these signed measures of finite energy with mutual energy as scalar product is pre-Hilbert, and the subset $\mathfrak{G}^+$ of the nonnegative measures is complete. As an application, the balayage of a measure μ in $\mathfrak{G}^+$ relative to a compact K is given by the projection in $\mathfrak{G}$ of μ on the set of measures of $\mathfrak{G}^+$ supported by K. It is now known that a polar set has μ-measure 0 for (μ, μ) finite; the previous definitions can be directly extended to the case of a Green space and the Green potentials. The principle of energy with its consequences holds. Note that such potentials are BLD functions and also that the energy–norm and the Dirichlet norm are proportional.

Other extensions are possible. Deny even developed a similar theory in $\mathfrak{R}^n$ with Schwartz distributions, a fixed one as kernel N, a variable one T, and the convolution as potential. The energy is defined by $\int \mathfrak{N}|\mathfrak{S}|^2 dx$ where $\mathfrak{N}$ and $\mathfrak{S}$ are the functions given by a Fourier transform (for Jean-Baptiste-Joseph Fourier of France).

Martin boundary. In a Green space Ω, the Green function G and its ratio to the functional value obtained by setting Green function argument y equal to y_0 a fixed point (see 69) may be considered. Here there exists a compact space $\hat{\Omega}$ (Martin space) unique up to homeomorphism, in which Ω is dense and such that all functions $K(x, y)$ of x have continuous *continuations* $K(X, y)$ separating the (Martin) boundary $\Delta = \hat{\Omega} - \Omega$. It is identical to the Euclidean closure when Ω is a ball or a sufficiently regular domain, but generally is quite different. A harmonic function $u > 0$ is called minimal if any other similar but smaller function must be proportional to it. The minimal functions are proportional to certain of the functions of y $K(X, y)$ (where X is in Δ) and the corresponding points are called minimal and form a part Δ_1 of Δ.

Any harmonic function $u \geqslant 0$ has a unique representation as an integral of K integrated with respect to the measure μ_u (see 70) where μ_u is a Radon measure $\geqslant 0$ on Δ supported by Δ_1. A Dirichlet problem may be studied with this boundary and any finite continuous function is resolutive. A set e is said to be thin at $X \in \Delta_1$ if a condition involving K holds (see 71). The fine topology on Ω may be extended to $\Omega \cup \Delta_1$ in a unique way such that the neighbourhoods of any $X \in \Delta_1$ intersect Ω in sets that are the complements of sets in Ω thin at X.

Among many applications of this new tool, Doob's result is important—that for any superharmonic $u > 0$ and any harmonic $h > 0$, u/h has a fine limit at the points of Δ_1 except on a set of u_h measure 0. This implies old results, as, for example, the famous result of P. Fatou that a harmonic function $u > 0$ in a ball has a non-tangential limit almost everywhere on the boundary.

(66) $\|f\| = \sqrt{\int \text{grad}^2 f\, dx}$

(67) $(\mu, \nu) = \int U^\mu\, d\nu$

(68) $\int (U^{\mu_1} - U^{\mu_2})(d\nu_1 - d\nu_2)$

(69) $G(x, y), K(x, y) = \dfrac{G(x, y)}{G(x, y_0)}$

(70) $\int K(X, y)\, d\mu_u(X)$

(71) $R^e_{K(X,y)} \neq K(X, y)$

(72) $\int f\, d\rho^\omega_y$

(73) $\int uf\, d\xi$

AXIOMATIZATIONS AND PROBABILITY THEORY

Because of the vastness and technical difficulties of the modern developments, many studies will not be discussed here.

Local axiomatic theories without basic kernels. On a connected, locally connected, locally compact but noncompact space Ω, on every open set a family of real finite continuous functions (called harmonic) are considered, forming a vector space and satisfying the following axioms, which are basic properties of the classical theory:

1. These functions form a sheaf (that is, harmonicity in an open set is equivalent to local harmonicity).

2. The term regular is applied to any relatively compact open set ω such that any given real finite continuous boundary function f may be continued continuously on ω in a unique way as a harmonic function which is $\geqslant 0$ when $f \geqslant 0$. Then this continuation $H_f^\omega(y)$ may be expressed as an integral (see 72) in which a harmonic measure is $\geqslant 0$ on $\partial\omega$. The second axiom asserts the existence of a base of regular open sets.

3. Any increasing sequence (equivalently directed set) of harmonic functions on a domain tends to $+\infty$ or to a harmonic function.

Note that the quotients by a finite continuous $h > 0$ on Ω define another sheaf satisfying the axioms (whose functions are called h-harmonic and were already used in classical theory). When h is harmonic, the constants are h-harmonic.

A hyperharmonic function is defined as before, by replacing the balls by regular domains. In a domain it is $+\infty$ or finite on a dense set in which case it is called superharmonic, or a potential when the greatest harmonic minorant is 0.

Dirichlet spaces. Inspired by the Hilbert space of BLD functions, one can introduce a locally compact set Ω, countable union of compact sets, a measure $\xi > 0$ on Ω and a vector space H of locally ξ-integrable complex functions f on Ω, with a norm making H a Hilbert space and decreasing when f is changed into $T \circ f$ for any contraction T of the complex plane preserving 0.

A potential generated by a bounded f with compact support is defined as the point P_f of H such that the scalar product (u, P_f) is equal to an integral (see 73) for any u of H. The closure of the set of these $f \geqslant 0$ is the set of "pure potentials." With some density conditions on the so-called Dirichlet space H, a rich theory may be developed with equilibrium and balayage problems.

Hunt kernels. On a locally compact space Ω, with countable base, one can consider a family of measures

$\geqslant 0$, μ_x depending on $x \in \Omega$; that is, a function depending on x (see 74), a Borel function of x for any Borel set e relatively compact, and a measure in e for any $x \in \Omega$. For any Borel function $f \geqslant 0$ one defines Vf in either of three ways (see 75). For example, for a specific form of a family of measures (see 76) in $\mathfrak{R}^3$, Vf is the Newtonian potential of the measure given by its density f.

The product V_0W of two kernels V and W is defined as the kernel $\int V(x, dy)\, W(y, e)$ and a family V_t ($t \geqslant 0$) can be defined as a semi-group (see 77). Now, under some restrictions on V, there exists a unique semi-group of kernels $\{P_t\}$ such that an integral restriction holds (see 78) where P_t has particular properties: P_t is sub-Markovian (that is, less than or equal to 1), P_0 is the identity kernel, $P_t f \to f$ uniformly as $t \to 0$ (f, finite, continuous and tending to 0 at infinity), $P_t f$ is finite continuous and tends to 0 at infinity when f is finite continuous with compact support.

A Borel function $f \geqslant 0$ is said to be excessive if $P_t f \leqslant f$ for any t and $P_t f$ tends to f as $t \to 0$. In the previous local axiomatics, the hyperharmonic functions greater than or equal to 0 are, roughly speaking, the excessive functions corresponding to the semi-group of a suitable Hunt kernel, and, in the most interesting cases of Dirichlet spaces, the real potentials are also excessive for a Hunt kernel.

(74) $V(x, e) = \mu_x(e)$

(75) $Vf = \int f\, d\mu_x,\ \int f(y) V(x, dy),\ \int V(x, dy) f(y)$

(76) $V(x, e) = \int_e \frac{1}{|x-y|}\, dy$

(77) $V_{t+s} = V_t \cdot V_s$

(78) $Vf = \int_0^\infty P_t f\, dt$

(79) $\int_\alpha y dP = \int_\alpha x dP$

(80) $P^y(\{X_0 = y\}) = 1,\ P^y(\alpha)$

(81) $X_r^{-1}(e), \qquad r \leqslant s, \qquad e \in \mathfrak{B}$

(82) $P^y(X_{s+t} \in \mathfrak{B} \mid \beta_s) = P^{X_s(x)}(X_t \in \mathfrak{B})$

(83) $P_t(y, e) = P^y(X_t \in e)$

Relations with probability theory. For a long time it was known that in R^2 or R^3, for a Brownian motion (named after a Scottish botanist, Robert Brown, who investigated the random motion of a microscopic particle in suspension) starting from a point y_0 in a domain ω, the probability that the trajectory hits for the first time a Borel set e of $\partial\omega$ (formerly supposed to be smooth) is exactly the harmonic measure of e at x_0. This inspired Doob, who, among other things, compared submartingales and subharmonic functions. Briefly the probabilistic interpretation of Hunt kernel V is as follows. In an abstract space A with a σ-field of sets, $\mathfrak{A}$, a probability measure P on $\mathfrak{A}$ is considered (positive abstract measure with total 1) with a function $X(x)$ (called the random variable), which maps A onto another similar space (B, $\mathfrak{B}$) and is $\mathfrak{A}$-measurable (*i.e.*, $X^{-1}(e) \in \mathfrak{A}$ for any $e \in \mathfrak{B}$). In case X is real ($B = R'$; $\mathfrak{B}$, Borel field), the expectation of X is $\int X dP$; the conditional expectation relative to a sub-σ-field $\mathfrak{A}'$ is any random variable Y, but $\mathfrak{A}'$-measurable such that the indefinite integral over α of y with respect to the measure P equals the indefinite integral over α of x with respect to the measure P (for any $\alpha \in \mathfrak{A}$; see 79). Y is unique in a certain sense and Y is denoted $E(X/\mathfrak{A}')$. In case X is the characteristic function of a set α_0, Y becomes the so-called conditional probability $P(\alpha_0/\mathfrak{A}')$.

Now a rather general case of Markov process is given by considering a family $X_t(x)$ ($t > 0$) and a probability measure P^y on $\mathfrak{A}$ depending on a point $y \in B$, and satisfying two conditions (see 80).

If $\mathfrak{A}_s$ denotes the σ-field generated by the sets induced inversely through x_r for certain values of the subscript (see 81) a last condition is expressed in terms of conditional probabilities (see 82) for any $\beta \in \mathfrak{B}$, any two s, t, and almost everywhere on A. Roughly speaking, that means the trajectory $X_t(x)$, t variable, in B, issued from y, does not depend after the time s on the positions of the particle before the time s.

With a slight adaptation, given a semi-group P_t of the Hunt theory, there exists an above (weak) Markov process such that a transition semigroup is defined (see 83).

There are now available probabilistic notions, as, for example, trajectories the use of which gives among many things an interpretation of thinness and polar sets (on B identified to the space Ω of the local axiomatics). There are other important results relating reduced functions (using excessive functions) to hitting times, etc. (Ed.)

BIBLIOGRAPHY. The reader interested in the foundations of the subject and the Cauchy theory may consult some of the following works: LARS V. AHLFORS, *Complex Analysis*, 2nd ed. (1966); EDWARD T. COPSON, *An Introduction to the Theory of Functions of a Complex Variable* (1935); EINAR HILLE, *Analytic Function Theory*, 2 vol. (1959–62); KONRAD KNOPP, *Funktionentheorie*, 5th ed., 2 vol. (1937–41; Eng. trans., *Theory of Functions*, 2 vol., 1945–47); ZEEV NEHARI, *Conformal Mapping* (1952); EDWARD C. TITCHMARSH, *The Theory of Functions*, 2nd ed. (1939); EDMUND T. WHITTAKER and GEORGE N. WATSON, *A Course of Modern Analysis*, 4th ed. (1927).

The following are references on elliptic functions and related topics: FRANK BOWMAN, *Introduction to Elliptic Functions, With Applications* (1854, reprinted 1953); PAUL F. BYRD and MORRIS D. FRIEDMAN, *Handbook of Elliptic Integrals for Engineers and Physicists* (1954); ALAN FLETCHER, J.C.P. MILLER, and LOUIS ROSENHEAD, *An Index of Mathematical Tables* (1946).

The following works deal with the theory of functions of several complex variables and related subjects: ROBERT C. GUNNING and HUGO ROSSI, *Analytic Functions of Several Complex Variables* (1965), containing an extensive bibliography; LARS HORMANDER, *An Introduction to Complex Analysis in Several Variables* (1966); RAGHAVAN NARASIMHAN, *Introduction to the Theory of Analytic Spaces* (1966); HENRI CARTAN, *Séminaire de l'École Normale Supérieure* (1952–53, 1953–54, 1957–58, 1960–61); K. KODAIRA and D.C. SPENCER, "On Deformations of Complex Analytic Structures, pt. I," *Ann. Math.*, 67:328–466 (1958); KIYOSHI OKA, *Sur les fonctions analytiques de plusieurs variables* (1961); and CARL L. SIEGEL, "Symplectic Geometry," *Am. J. Math.*, 65:1–86 (1943), are among the classics in the field.

The following are references on classical potential theory: WOLFGANG J. STERNBERG and TURNER L. SMITH, *The Theory of Potential and Spherical Harmonics*, 2nd ed. (1946), is an elementary work; OLIVER D. KELLOGG, *Foundations of Potential Theory* (1929, reprinted 1953), is more advanced. GRIFFITH C. EVANS, *The Logarithmic Potential, Discontinuous Dirichlet and Neumann Problems* (1927); and LESTER L. HELMS, *Introduction to Potential Theory* (1969), use modern integration theory. MASAJI TSUJI, *Potential Theory in Modern Function Theory* (1959), deals with applications. MARCEL BRELOT, *Eléments de la théorie classique du potentiel*, 4th ed. (1969), also deals with the classical theory. For modern developments in potential theory, see MARCEL BRELOT, *Lectures on Potential Theory*, 2nd ed. rev. (1967), and *On Topologies and Boundaries in Potential Theory* (1971); PAUL A. MEYER, *Probabilities and Potentials* (1966); HEINZ BAUER, *Harmonische Raume und ihre Potentialtheorie* (1966); and ROBERT M. BLUMENTHAL and RONALD K. GETOOR, *Markov Processes and Potential Theory* (1968).

(K.Iw./R.Na./Ed.)

Analysis, Fourier

Fourier analysis, one of the most important tools of mathematics and mathematical physics, arose originally from problems involving vibrating cords as studied by the Swiss mathematicians Leonhard Euler and Daniel Bernoulli in the 18th century. The subject was named for Jean-Baptiste-Joseph Fourier, a French mathematician

and physicist, whose work on the representation of mathematical functions as sums of trigonometric functions was published in his *Théorie analytique de la chaleur* (*The Analytical Theory of Heat*, 1878), in 1822. Before considering the theory of Fourier series, and the more general theory of integral transforms and harmonic analysis, a summary of the theory of series will be given.

The parts and subparts of this article are as follows:

Series
- Examples
- Series with complex terms
- Double sequences and series
- Series of functions
- Power series
- Divergent series
- History of series

Fourier series
- Convergence and divergence of Fourier series
- Summability of Fourier series
- Conjugate series and functions
- Parseval's formula
- Uniqueness of representation by trigonometric series
- Trigonometric interpolation
- Orthogonal series
- Fourier integrals
- Multiple Fourier series

Harmonic analysis and integral transforms
- Convolution
- The Fourier transform
- Examples of Fourier transforms
- Variants of the Fourier transform
- Convolution and the Fourier transform
- Applications to partial differential equations
- Generalizations of the Fourier transform

Representations of groups and algebras
- Historical development
- Groups
- Group representations
- Irreducible representations
- Topological groups and Lie groups

SERIES

Sequences of numbers

The notion of a series has a close connection with that of a sequence; a set of numbers arranged in order. The simplest example of a sequence is that of the positive integers $1, 2, 3, \cdots, n, \cdots$. A sequence may be finite or infinite, depending on whether the number of terms is limited or unlimited. An infinite sequence of numbers exists if every positive integer $1, 2, 3, \cdots, n, \cdots$ is assigned a number; in what follows the expression sequence means an infinite sequence of numbers.

Some examples of sequences are as follows: a sequence of squares of positive integers (see Box, sequence 1), an alternating sequence of the integers one and zero (see 2), an alternating sequence of integers (alternating means that the sign of each term is opposite to that of the preceding term; see 3), and a sequence of fractions each with numerator one and with denominators ascending in integer value (see 4).

The numbers constituting a sequence are called the terms of the sequence (such as the first, the second, $\cdots$, the nth, $\cdots$). It is customary to denote a sequence by using a single letter with a subscript (called the index) to indicate the position of the term; for example $a_1, a_2, a_3, a_4, \cdots, a_n, \cdots$. An abbreviated notation $\{a_n\}$ is also used. It is assumed here that the terms are real numbers (natural numbers, their ratios, negatives and limits); the case of sequences of complex numbers is considered later.

A series is the indicated sum of a sequence of numbers; for example, $a_1 + a_2 + a_3 + \cdots + a_n + \cdots$. A series can also be finite or infinite depending on the sequence. It is customary to denote a series by using the Greek letter sigma to indicate summation. For example, $a_1 + a_2 + a_3 + \cdots + a_n + \cdots$ is written $\sum_{n=1}^{\infty} a_n$, or simply Σa_n.

The properties of series are related to the properties of sequences. If every term in a sequence is less than or equal to the following term, the sequence is said to be increasing or (nondecreasing). In other words if, in $\{a_n\}$, a_n is less than or equal to a_{n+1} for every n, then $\{a_n\}$ is increasing. If each term is less than the following one (that is, if $a_n < a_{n+1}$), the sequence is strictly increasing. On the other hand, if for a sequence $\{a_n\}$, a_n is greater than or equal to a_{n+1}, the sequence is decreasing (or non-increasing); if $a_n > a_{n+1}$, it is strictly decreasing.

(1) $1^2, 2^2, 3^2, \cdots, n^2, \cdots$

(2) $1, 0, 1, 0, 1, \cdots$

(3) $1, -2, 3, -4, 5, -6, \cdots$

(4) $1, 1/2, 1/3, \cdots, 1/n, \cdots$

Note: (1), (2), (3), and (4) are examples of sequences.

(5) $\lim_{n\to\infty} a_n = a$ or, $a_n \to a$

(6) $\lim_{n\to\infty} a_n = +\infty$ or, $a_n \to +\infty$

(7) $a_n + b_n \to a + b$, $a_n b_n \to ab$, $a_n/b_n \to a/b$

Sequences that are either increasing or decreasing are called monotone (monotonically increasing or monotonically decreasing, as the case may be; see 1 and 4). Some sequences are not monotone (see 2 and 3). If all the terms of the sequence a_n are less than a given number M (in symbols, $a_n < M$ for some M and all n), the sequence is said to be bounded above. If all a_n exceed a certain number, $\{a_n\}$ is said to be bounded below (see 1). Sequences bounded both above and below are simply called bounded (see 2). Some sequences are not bounded either above or below (see 3).

Convergence and divergence

A sequence $\{a_n\}$ is said to converge (or tend) to the limit a if the terms a_n approach a indefinitely as n increases (see 5). A precise definition is as follows: a_n tends to the limit a if, given any positive number, symbolized by the Greek letter epsilon, ε (no matter how small), the magnitude of the difference, $|a_n - a|$, is less than ε for all n large enough. If $\{a_n\}$ tends to a limit, the sequence is said to be convergent; otherwise it is divergent. The following type of divergence is important: $\{a_n\}$ is said to diverge to infinity, or the limit of a_n as n approaches infinity is positive infinity (see 6), if for any number M (no matter how large) a_n is greater than M for all n large enough; divergence of a sequence to negative infinity is defined in a similar way.

Of the sequences mentioned above, one (see 4) converges to the limit 0; others (see 1 and 2) diverge. Whether a given sequence converges and, if so, toward what limit, are not always easy to decide. The following facts, however, can be helpful.

I. If $a_n \to a$ and $b_n \to b$, then the limit of the sum of the product and the ratio of a_n and b_n is the sum of the product and ratio of a and b, respectively (see 7), the last relation requiring, however, the additional assumption that b is not equal ($\neq$) to 0.

II. Every convergent sequence is necessarily bounded; an example (see 2) shows that the converse is not true: a bounded sequence need not converge.

III. If a sequence is increasing and bounded above, or decreasing and bounded below, it converges.

IV. If a sequence converges, then its terms, coming arbitrarily close to the limit, must also come arbitrarily close to each other, provided the indices are large enough. The converse of this is also true, though less easy to prove, and the result is the following important theorem of the 19th-century French mathematician Augustin-Louis Cauchy: a sequence $\{a_n\}$ converges if and only if it has the following property: for any positive number ε (no matter how small) $|a_m - a_n| < \varepsilon$ for all m and n large enough.

Starting with a sequence $a_1, a_2, \cdots, a_n, \cdots$ a new sequence $s_1, s_2, \cdots, s_n, \cdots$ is formed by the following rule:

a sequence of partial sums is generally labelled s_n and defined as a sum of the first n of its terms (see 8). If $\{s_n\}$ converges to limit s, then it is said that the infinite series, an infinite sum of a_k (see 9), converges to sum s. The numbers s_n are called the partial sums of the series Σa_n, and the relation $s_n \to s$ is also written $\Sigma a_n = s$. The numbers a_n are the terms of the series Σa_n.

For any sequence $b_1, b_2, \cdots, b_n, \cdots$, there is always an associated series that is a sum of differences constructed of b_k and with typical term b_n minus b_{n-1} (see 10), the partial sums of which are $b_1, b_2, b_3, \cdots, b_n, \cdots$. Hence, every series is represented by a sequence (that of its partial sums) and, conversely, every sequence by a series. It follows that every statement about series can be given in a form bearing on sequences, and conversely (but usually, in individual cases, one form may be preferable to the other). For example, Cauchy's theorem stated above and pertaining to sequences assumes, in the case of series, the following form:

V. A series Σa_n converges if, and only if, given any positive number ε, $|\, a_{m+1} + a_{m+2} + \cdots + a_n \,| < \varepsilon$ for all m large enough and any n greater than m. In particular (taking $n = m + 1$), the terms of a convergent series must necessarily tend to 0.

Tests for convergence

Examples. If a is any number and q is any number of absolute value less than 1, the series $a + aq + aq^2 + \cdots + aq^{n-1} + \cdots = \Sigma aq^{n-1}$ (called a geometric series) converges, and its sum is $a/(1 - q)$.

The series $\Sigma 1/n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots$, called the harmonic series, diverges.

If the terms of a series are positive numbers, the series is called positive. The partial sums of a positive series form a strictly increasing sequence and, in view of proposition III, a positive series converges if its partial sums are bounded above. If the partial sums of a positive series are not bounded above, the series diverges to $+\infty$. Proposition III also shows that:

VI. If Σa_n and Σb_n are positive series, and if $a_n \geqslant b_n$ for all n, then the convergence of Σa_n implies convergence of Σb_n and, equivalently, the divergence of Σb_n implies that of Σa_n. This result, which is usually called the comparison test, is very useful because the knowledge of convergence or divergence of some positive series can be used to obtain information about other series. For example, from the fact that the geometric series Σaq^{n-1} converges if a is positive and q is positive and less than 1, it is possible to deduce the following corollaries for positive series:

VII. If $a_{n+1}/a_n \to l$, then, if $l < 1$, the series Σa_n converges; if $l > 1$, the series diverges. This is known as the ratio test.

VIII. If $a_n^{1/n} \to l$, and if $l < 1$, the series Σa_n converges; if $l > 1$, the series diverges. This is known as the root test.

If the limit l in VII or VIII exists but is equal to 1, the test fails (the case $l = 1$ can occur for both convergent and divergent series) and other tests are sought. There are many of them, but the most useful is the following one based on the notion of the integral:

IX. If $f(x)$ is a function defined for x positive, itself positive and steadily decreasing to 0 as x increases indefinitely, then the series $\Sigma f(n) = f(1) + f(2) + f(3) + \cdots$ converges or diverges according as the $\int_1^\infty f(x)dx$ is finite or not. This is known as the integral test.

When applied to the function $f(x) = x^{-k}$, in which k is a fixed positive constant, the test shows that the series that is a sum of reciprocals of kth powers of positive integers (see 11) converges if $k > 1$ and diverges in all other cases. The result is not deducible from either the ratio or root test, because for the series $\sum_{n=1}^{\infty} n^{-k}$, l is equal to 1 both in VII and VIII.

Of interest among series that are not positive are alternating series; this name is given to series $a_1 - a_2 + a_3 - a_4 + \cdots$ in which the numbers a_n themselves are of constant sign. It can be shown that such a series converges if $\{a_n\}$ decreases monotonically to 0. For example, the series $1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots$ converges.

(8) $\begin{cases} s_1 = a_1, \quad s_2 = a_1 + a_2, \quad s_3 = a_1 + a_2 + a_3, \quad \cdots \\ s_n = a_1 + a_2 + \cdots + a_n, \quad \cdots \end{cases}$

(9) $a_1 + a_2 + a_3 + \cdots + a_n + \cdots$

(10) $b_1 + (b_2 - b_1) + (b_3 - b_2) + \cdots + (b_n - b_{n-1}) + \cdots$

(11) $1^{-k} + 2^{-k} + 3^{-k} + \cdots + n^{-k} + \cdots = \sum_{n=1}^{\infty} n^{-k}$

(12)
$$\begin{cases} 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \cdots = \ln 2 \\ 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots = \frac{1}{4}\pi \text{ (Leibniz' series)} \\ 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \cdots = \frac{1}{6}\pi^2 \\ 1 + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \frac{1}{9^2} + \cdots = \frac{1}{8}\pi^2 \\ 1 + \frac{1}{1} + \frac{1}{1\times 2} + \frac{1}{1\times 2\times 3} + \frac{1}{1\times 2\times 3\times 4} + \cdots \\ \quad = e = 2.7182818\cdots \end{cases}$$

These series converge to known values as shown.

A series Σa_n is said to converge absolutely if $\Sigma\, |\, a_k \,|$ converges. Cauchy's theorem for series shows that if a series converges absolutely, then it converges in the ordinary sense; but the example of the convergent alternating series $1 - \frac{1}{2} + \frac{1}{3} - \cdots$ shows that a series may converge without converging absolutely. It follows that absolutely convergent series form a special class among all convergent series. Because series $\Sigma\, |\, a_k \,|$ have positive or zero terms, the problem of whether a given series converges absolutely can be solved by applying one of the tests for the convergence of positive series. In many cases the simplest way of proving that a certain series converges is by showing that the series converges absolutely, and this shows the importance of positive series for the general theory of series. Absolutely convergent series have a number of properties that are not shared by all convergent series. The most important of these properties is the fact that if a series converges absolutely, then arbitrary changes in the order of the terms of the series affect neither the convergence nor the sum of the series.

Absolute convergence

The situation is completely different for series that converge but do not do so absolutely (such series are called conditionally convergent); by a suitable change of the order of the terms, such a series can be made to converge to any prescribed sum or can even be made divergent. This is a theorem of the 19th-century German mathematician Bernhard Riemann. Hence, unlike the case of absolutely convergent series, a conditionally convergent series has no intrinsic sum and everything depends on the order of the terms.

The number of important series that occurs in mathematics is enormous. A few basic convergent series are given in five examples of series the sums of which are known and are expressed in terms of the natural logarithm and the numbers π and e (see 12).

Series with complex terms. The convergence of complex sequences and series (that is, sequences and series the terms of which are complex numbers) is formally defined in exactly the same way as in the case of real numbers, the absolute value $|\, z \,|$ of a complex number $z = x + iy$, in which i is $\sqrt{-1}$, being $\sqrt{(x^2 + y^2)}$. This definition leads to the following important conclusion: if $c_n = a_n + ib_n$, the convergence of Σc_n is equivalent to the simultaneous convergence of Σa_n and Σb_n; and if $\Sigma a_n = s$, $\Sigma b_n = t$, then $\Sigma c_n = s + it$. Hence, the study of complex series reduces to that of real series. As in the case of real series, the absolute convergence of Σc_n is defined as the convergence of $\Sigma\, |\, c_n \,| = \Sigma \sqrt{(a_n^2 + b_n^2)}$.

$$(13)\qquad \begin{cases} s_{1,1}, s_{1,2}, \cdots, s_{1,n}, \cdots \\ s_{2,1}, s_{2,2}, \cdots, s_{2,n}, \cdots \\ \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \\ s_{m,1}, s_{m,2}, \cdots, s_{m,n}, \cdots \\ \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \end{cases}$$

$$(14)\qquad \begin{cases} a_{1,1} + a_{1,2} + \cdots + a_{1,n} + \cdots \\ a_{2,1} + a_{2,2} + \cdots + a_{2,n} + \cdots \\ \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \\ a_{m,1} + a_{m,2} + \cdots + a_{m,n} + \cdots \\ \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \end{cases}$$

$$(15)\qquad \int_a^b f(x)\,dx = \sum \int_a^b u_n(x)\,dx$$

Because the convergence of $\Sigma\sqrt{(a_n^2 + b_n^2)}$ means the same thing as the simultaneous convergence of $\Sigma \mid a_n \mid$ and $\Sigma \mid b_n \mid$, the absolute convergence of $\Sigma(a_n + ib_n)$ is equivalent to the absolute convergence of both $\Sigma \mid a_n \mid$ and $\Sigma \mid b_n \mid$; it therefore implies the ordinary convergence of Σa_n and Σb_n, and so also of $\Sigma(a_n + ib_n)$. Among complex series, absolutely convergent series play the same important role as among real series; in particular, it is the case that the convergence and the sum of an absolutely convergent complex series are independent of the order of the terms.

Double sequences and series. A double sequence is defined by assigning to every pair m, n of positive integers a number $\{s_{m,n}\}$. A double sequence can thus be written as a double array of numbers: a sequence of sequences with typical terms s_m and s_n (see 13). The indices m, n of a term can be interpreted as the coordinates of a point in a plane. The notion of convergence of a double sequence is not so immediate and can be introduced in many ways (which are not equivalent). The following definition is most commonly used: $\{s_{m,n}\}$ is said to converge to limit l if for any positive ε the quantity $\mid s_{m,n} - l \mid < \varepsilon$ provided both m and n are large enough. With this definition, changing the terms in any finite number of rows and columns in the table of (13) above (the number of changes may be infinite) does not affect the convergence or limit of $\{s_{m,n}\}$.

The definition of convergence of a double series (see 14) can also be introduced in various ways; the series itself will be written as $\sum_{m=1}^{\infty} \sum_{n=1}^{\infty} a_{m,n}$ or simply $\Sigma\Sigma a_{m,n}$. Three

Convergence of double series

definitions follow, only the first of which has close connection with the just defined notion of convergence of double sequences (at any rate, connection between double series and sequences is less intimate than in the case of single series and sequences):

(a) $\Sigma\Sigma a_{m,n}$ is said to converge to sum s, if the double sequence $\{s_{m,n}\}$ tends to s, in the sense just described in which $s_{m,n}$ denotes the sum of the terms $a_{j,k}$ with $j \leqslant m$, $k \leqslant n$; such terms $a_{j,k}$ fill a rectangle in the upper left corner of the double series table of (14) and the sums $s_{m,n}$ are correspondingly called rectangular partial sums of the series; the kind of convergence that has just been introduced is called the rectangular convergence of the series;

(b) the series $\Sigma\Sigma a_{m,n}$ is said to converge to sum s, if S_R tends to s as R tends to $+\infty$, in which S_R is the sum of the terms $a_{m,n}$ with $m^2 + n^2 \leqslant R^2$; indices m, n satisfying this inequality are coordinates of points situated in the circle with the centre at the origin and radius R, so that it is natural to call the S_R the circular partial sums, and the kind of convergence they generate circular convergence;

(c) $\Sigma\Sigma a_{m,n}$ is said to converge to sum s if the terms in each column of the table form a convergent series and if $A_1 + A_2 + \cdots + A_n + \cdots = s$, in which A_n is the sum of terms in the nth column of the table; this kind of summation is called summation by columns, and, by interchanging the roles of columns and rows, leads to a parallel notion of convergence by rows.

These definitions of convergence are not equivalent and a series can converge according to one definition and not according to another. The problem of additional conditions under which the definitions are equivalent is not very difficult, but the conditions themselves are not of a simple nature and most (though not all) important series $\Sigma\Sigma a_{m,n}$ in analysis are absolutely convergent (that is, $\Sigma\Sigma \mid a_{m,n} \mid$ converges). For such series all reasonable definitions of convergence are equivalent and the sum of the series is independent of the definition. Definitions and results can be extended to triple, quadruple, etc., series.

Series of functions. If $\Sigma u_n(x)$ is a series the terms $u_1(x), u_2(x), \cdots, u_n(x), \cdots$ of which are functions defined in an interval of the real variable x, and if $\Sigma u_n(x)$ converges at each point of the interval, then the series is said to converge pointwise in the interval; $f(x)$ denotes the sum, and $S_n(x)$ the partial sums of the series. Pointwise convergence means that for any positive ε, there is an index n_0, depending in general on ε and x, in symbols $n_0 = n_0(\varepsilon, x)$, such that $\mid f(x) - S_n(x) \mid < \varepsilon$ for $n > n_0$. If for each $\varepsilon > 0$ an n_0 can be found independent of x, the series is said to converge uniformly in the interval considered; geometrically this means that for each $n > n_0$ the graph of the curve $y = s_n(x)$ and that of the curve $y = f(x)$ differ by less than ε in the whole interval under consideration. Pointwise convergence in general does not preserve properties of the terms of the series. For example, if the functions $u_n(x)$ are continuous and $\Sigma u_n(x)$ converges pointwise, the sum $\Sigma u_n(x)$ need not be continuous. A simple example is provided by the series $(1 - x) + x(1 - x) + x^2(1 - x) + \cdots$, which converges at each point of the interval $0 \leqslant x \leqslant 1$ and the sum of which is equal to 0 for $x = 1$ and equal to 1 at the remaining points of the interval. Similarly, if the $u_n(x)$ are integrable over an interval and the series $\Sigma u_n(x)$ converges pointwise, the sum of $\Sigma u_n(x)$ need not be integrable. The significance of the notion of uniform convergence is that the sum of a uniformly convergent series of functions inherits many important properties of the terms—in particular their continuity and integrability. Not all properties, however, are preserved by uniform convergence; for example, the sum of a uniformly convergent series of differentiable functions need not be differentiable. As a matter of fact, examples of continuous and nowhere differentiable functions are usually given as sums of uniformly convergent series of differentiable functions. The three theorems that now follow about uniformly convergent series are particularly useful.

Uniformly convergent series; Weierstrass' M-test

X. If the functions $u_n(x)$ are continuous in an interval, and if the series $\Sigma u_n(x)$ converges uniformly in the interval, the sum of the series is continuous in the interval.

XI. If the $u_n(x)$ are integrable over a finite interval (a, b), and if $\Sigma u_n(x)$ converges uniformly in (a, b), then the sum $f(x)$ of $\Sigma u_n(x)$ is integrable over (a, b) and a definite integral of the function f is written as a sum of integrals of functions u_n (see 15).

XII. If each term of the series $\Sigma u_n(x)$ has a continuous derivative in a finite interval (a, b), if the series $\Sigma u'_n(x)$ of derivatives converges uniformly in (a, b), and if the series $\Sigma u_n(c)$ converges for one point c in (a, b), then the series $\Sigma u_n(x)$ also converges uniformly in (a, b), its sum $f(x)$ is differentiable, and $f'(x) = \Sigma u'_n(x)$.

There are no special tests for the pointwise convergence of series because pointwise convergence simply means ordinary convergence at each point x separately. There are tests for the uniform convergence of a series of functions, among which the following, usually called the Weierstrass' M-test, is particularly useful:

XIII. If there are positive constants $M_1, M_2, \cdots, M_n, \cdots$ such that the series ΣM_n converges and if $\mid u_n(x) \mid \leqslant M_n$ for each n and all x of an interval, then the series $\Sigma u_n(x)$ converges uniformly in the interval.

Only the uniform convergence of series has been considered. There are parallel definitions and results for the

uniform convergence of sequences. One can also consider the uniform convergence of double series and sequences. Special types of series of functions are particularly important for various branches of analysis (as in the Fourier series described later). The convergence of series $\Sigma u_n(z)$ the terms of which depend on a complex variable $z = x + iy$ may be considered. The most important among them is the case of power series.

Circle of convergence of power series

Power series. Power series are of the form with coefficients a_n (see 16), in which the a_n are constants (in general, complex) and $z = x + iy$ is a complex variable; in the symbol $\Sigma a_n z^n$ the index n ranges through the values $0, 1, 2, \cdots, n, \cdots$. Each power series has a circle of convergence; this is a circle (disk) with the centre at the point $z = 0$ and the following properties: the series converges at each point interior to the circle and diverges at each point exterior to the circle. As to the behaviour on the circumference of the circle of convergence, $\Sigma a_n z^n$ may converge at some points, diverge at others. The intersection of the circle of convergence with the real axis is an interval $(-R, R)$; and the series $\Sigma a_n x^n$, a power series of the real variable x, converges in the interior of this interval and diverges in the exterior. The radius R of the circle of convergence is called the radius of convergence of the power series. The circle of convergence can degenerate to a point; each power series $\Sigma a_n z^n$ necessarily converges at the point $z = 0$ but may converge at no other point: the series $\Sigma n! z^n$ is an example of this ($n!$—factorial-n—is the product $n(n-1)(n-2)\cdots 2, 1$). The circle of convergence may also cover the whole complex plane, as in the series $\Sigma z^n/n!$. Finally, the radius of convergence may have any prescribed positive value R because the radius of convergence of the geometric series $\Sigma(z/R)^n$ is precisely R.

At each point interior to the circle of convergence, the power series not only converges but converges absolutely. What is very important is that in each circle concentric with the circle of convergence but of smaller radius, the power series converges uniformly. Series of the form $a_0 + a_1(z-b) + a_2(z-b)^2 + \cdots = \Sigma a_n(z-b)^n$, in which b is a fixed complex number, are called power series with centre b. Their study does not require the introduction of new concepts because by replacing $z - b$ with a new variable Z the series is reduced to the previously discussed case $\Sigma a_n Z^n$. It follows that $\Sigma a_n(z-b)^n$ has also its circle of convergence, but this time the circle has its centre at the point b. Functions representable by power series have very important properties. Listed below are a few basic power series, considering for simplicity only real values of z. Six examples of expansions of functions of x into power series form may be given (see 17).

Divergent series. The series $1 - 1 + 1 - 1 + \cdots$ has partial sums $1, 0, 1, 0, \cdots$ that do not tend to any limit; therefore, according to the definition, the series is divergent. On the other hand, the average value of the partial sums $1, 0, 1, 0, \cdots$ is ½, a fact that made some mathematicians as early as the 17th century adopt the point of view that though the series, strictly speaking, is divergent it somehow has the sum ½. Such ideas, though for a long time unclear and confused, were finally clarified and led to a logically coherent and fruitful theory of divergent series and sequences. A few facts are presented here. The theories of divergent series and sequences being parallel, sequences are considered first.

If a sequence $\{s_n\} = s_0, s_1, \cdots, s_n, \cdots$ is defined (in this context it will be convenient to include 0 as a value of n), a definition of convergence of $\{s_n\}$ has already been formulated. This definition, which may be called classical, is not the only one possible; alternate definitions could be adopted. The following one is very intuitive. The averages of the successive terms of $\{s_n\}$ are denoted $\sigma_0, \sigma_1, \sigma_2, \cdots, \sigma_n, \cdots$, so that a sequence of arithmetic averages can be constructed from the s_k (see 18). It can be said that $\{s_n\}$ converges on the average to a limit s if the sequence $\{\sigma_n\}$ converges to the same s in the classical sense. It can be shown that any sequence $\{s_n\}$ that converges in the classical sense also converges on the average, and that the limits are the same in both cases. Some

(16) $$a_0 + a_1 z + a_2 z^2 + \cdots + a_n z^n + \cdots = \Sigma a_n z^n$$

(17)
$$\begin{cases} (1+x)^\alpha = 1 + \alpha x + \dfrac{\alpha(\alpha-1)}{1\times 2}x^2 + \dfrac{\alpha(\alpha-1)(\alpha-2)}{1\times 2\times 3}x^3 + \\ \qquad + \cdots \quad \text{(Newton's binomial series; } R = 1) \\ \ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \cdots \qquad (R=1) \\ \arctan x = x - \frac{1}{3}x^3 + \frac{1}{5}x^5 - \cdots \qquad (R=1) \\ e^x = 1 + \dfrac{x}{1} + \dfrac{x^2}{1\times 2} + \dfrac{x^3}{1\times 2\times 3} + \cdots \qquad (R=\infty) \\ \cos x = 1 - \dfrac{x^2}{1\times 2} + \dfrac{x^4}{1\times 2\times 3\times 4} - \\ \qquad - \dfrac{x^6}{1\times 2\times 3\times 4\times 5\times 6} + \cdots \qquad (R=\infty) \\ \sin x = x - \dfrac{x^3}{1\times 2\times 3} + \dfrac{x^5}{1\times 2\times 3\times 4\times 5} - \cdots \\ \qquad (R=\infty) \end{cases}$$

(18)
$$\begin{cases} \sigma_0 = \dfrac{s_0}{1}, \quad \sigma_1 = \dfrac{s_0+s_1}{2}, \quad \sigma_2 = \dfrac{s_0+s_1+s_2}{3}, \quad \cdots \\ \sigma_n = \dfrac{s_0+s_1+\cdots+s_n}{n+1}, \quad \cdots \end{cases}$$

(19)
$$\begin{cases} a_{0,0}, a_{0,1}, a_{0,2}, \cdots, a_{0,n}, \cdots \\ a_{1,0}, a_{1,1}, a_{1,2}, \cdots, a_{1,n}, \cdots \\ a_{2,0}, a_{2,1}, a_{2,2}, \cdots, a_{2,n}, \cdots \\ \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \\ a_{m,0}, a_{m,1}, a_{m,2}, \cdots, a_{m,n}, \cdots \\ \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \end{cases}$$

(20) $$\sigma_m = a_{m,0}s_0 + a_{m,1}s_1 + a_{m,2}s_2 + \cdots + a_{m,n}s_n + \cdots$$

Convergence on the average: method of first arithmetic mean

sequences do not converge in the classical sense but converge on the average; the sequence $\{s_n\} = 1, 0, 1, 0, \cdots$ is an instance in point because here $\{\sigma_n\} = 1, ½, ⅔, ½, ⅗, ½, \cdots$ tends to ½. Hence, in adopting convergence on the average as a definition of convergence of a sequence, conflict with the old definition is avoided on the one hand, and, on the other, sequences previously considered divergent (and useless) are made convergent. Instead of saying that $\{s_n\}$ converges on the average, it is now customary to say that $\{s_n\}$ is summable by the method of the first arithmetic mean. Not every sequence is thus summable. For example, if $s_n = 1, -2, +3, -4, +5, \cdots$, then σ_n does not converge in the classical sense; but it is easy to prove that σ_n is summable by the method of the first arithmetic mean to the limit 0, so that it may be said that $\{s_n\}$ is summable to limit 0 by the method of the second arithmetic mean. The third, fourth, etc., arithmetic means could be similarly introduced. It is, however, preferable to adopt a more general point of view.

The averages considered in the definition of $\{\sigma_n\}$ are not the only conceivable ones, since weighted averages may be introduced. The general scheme is as follows: Denote by M the fixed matrix of numbers for a double array of symbols with typical term $a_{m,n}$ displayed in rectangular pattern (see 19). Given any sequence $\{s_n\}$, a new sequence σ_m may be introduced defined by the formulas in which a weighted sum of s_n has the typical coefficient $a_{m,n}$ (see 20), in which $m = 0, 1, 2, \cdots$. Then σ_m can be considered as a weighted average of the numbers $s_0, s_1, \cdots$ formed with weights situated in the mth row of the matrix. If $\{\sigma_m\}$ converges in the classical sense to limit s, it may be said that $\{s_n\}$ is summable by the matrix M, or simply summable M, to limit s; s is then the generalized limit of $\{s_n\}$, associated with the

matrix M. It is easily seen that the method of the first arithmetic mean is given by the matrix that is a sequence of sequences composed of a single 1, two one-halves, three one-thirds, and so forth (see 21). Classical convergence corresponds to the matrix whose terms on the main diagonal are all 1 ($a_{n,n} = 1$ for all n) and the remaining terms are 0, because then $\sigma_n = s_n$ for all n.

This definition of summability of a sequence introduced by the 20th-century German mathematician Otto Toeplitz is extremely general, though it could be generalized still further, and raises a very large number of problems that are the subject of the general theory. A few such problems are considered below.

The first question concerns the conditions under which this definition is a genuine generalization of the classical convergence. The answer is that a necessary and sufficient condition that every convergent sequence should also be summable M, and to the same limit, is that (1) the terms of M should tend to 0 in each column separately; (2) the sum of the absolute values of the terms in each row should stay below some constant; (3) the algebraic sum of elements in each row should tend to 1 as the index of the row increases. Many results about summable sequences are corollaries of this theorem. In the special case in which the elements of the matrix are nonnegative, condition (2) is essentially a consequence of (3) and can be dropped.

It is said that a method M' is stronger than M if every sequence summable M is summable M', but not conversely, and if the limits in both cases are the same. It may then be asked in what case a given method M' is stronger than M. Theorems answering this question are called Abelian, after the 19th-century Norwegian mathematician Niels Henrik Abel. If M' is stronger than M, the question may be asked under what additional conditions the sequence summability by M' implies summability by M; results of this type are called Tauberian theorems, after the 20th-century Austrian mathematician Alfred Tauber. The most interesting theorems of Abelian and Tauberian character pertain to specific and important methods of summability.

Summability of convergent sequences; Abelian, Tauberian theorems

Given a series $a_0 + a_1 + \cdots + a_n + \cdots$, it may be said that it is summable M to limit s. To distinguish between series and sequences, some mathematicians say that a series is summable while a sequence is limitable, but no confusion arises if the word summable is used in both cases. Certain forms of summability, however, are more adapted to series than sequences, and vice versa. This situation, in particular, applies to Abel's method of summation, which defines the sum of Σa_n as the limit of $\Sigma a_n r^n$ for r tending to 1 through values less than 1. Abel's method is stronger than convergence, and even stronger than the method of the first arithmetic mean (Abelian theorems). A series that is Abel summable need not converge; for example, the divergent series $1 - 1 + 1 - \cdots$ is Abel summable to sum ½ because $1 - r + r^2 - r^3 + \cdots = 1/(1 + r)$ tends to ½ as $r \to 1$. If, however, $a_n \leqslant 1/n$ for all n, and Σa_n is Abel summable, then Σa_n converges (a Tauberian theorem).

The method of the first arithmetic mean and that of Abel are among the most important ones. Their main applications are to the study of the behaviour of power series on the circumference of their circles of convergence and to the theory of Fourier series. There is a great variety of methods of considerable intrinsic interest and importance in applications. The method of arithmetic means of various orders (hinted at above) has developed into a wide and elegant theory with far-reaching applications. The method of Borel assigns to a sequence $s_0, s_1, \cdots, s_n, \cdots$ the sum that is a limit as x approaches infinity of an expression involving an exponential and a specific power series in x (see 22) and can sum power series outside their circles of convergence. There is also an extension of the theory to divergent integrals. The importance of the theory of divergent series lies, however, not only in individual results but also in the general point of view upon the notion of convergence, a point of view that has had a great impact upon the development of mathematical analysis.

(21)
$$\begin{cases} 1, & 0, & 0, & 0, & \cdots \\ \frac{1}{2}, & \frac{1}{2}, & 0, & 0, & \cdots \\ \frac{1}{3}, & \frac{1}{3}, & \frac{1}{3}, & 0, & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{cases}$$

(22)
$$\lim_{x \to +\infty} \left\{ e^{-x} \sum_{n=0}^{\infty} \frac{s_n}{n!} x^n \right\}$$

(23)
$$\frac{1}{2} a_0 + (a_1 \cos x + b_1 \sin x) + \\ + (a_2 \cos 2x + b_2 \sin 2x) + \cdots$$

History of series. The origin of the modern theory of series precedes that of the calculus and can be associated in the 17th century with the names of Bonaventura Cavalieri (Italian), John Wallis (English), James Gregory (Scottish), and others. In the same century, the discovery of the calculus by Gottfried Wilhelm Leibniz and Sir Isaac Newton gave a very strong impulse to the theory; and the 18th century witnessed considerable progress in the work of Jakob and Daniel Bernoulli (Swiss), Jean Le Rond d'Alembert (French), Leonhard Euler (Swiss), and others. Much of this work was of purely formal character, and mathematical rigour was not the primary concern of the investigators.

Origin and development of the theory

Though Newton and Leibniz avoided divergent series, formal and uninhibited use of them brought such mathematical rewards that the temptation was too much to withstand, and divergent series became—mostly through the work of Euler—an accepted tool of investigation. Mathematical instinct of the great mathematicians of the 18th century prevented them, however, from making mistakes, and their results, properly interpreted, can now be proved rigorously. Rigorous foundations for the theory of infinite series were laid in the first half of the 19th century by the French mathematician Augustin-Louis Cauchy and Abel; their definitions and proved results can be found in all textbooks of calculus. Their attitude toward divergent series, however, was purely negative. (Abel wrote in 1828: "Divergent series are the invention of the devil, and it is shameful to base on them any demonstration whatsoever.") The respect commanded by the work of Cauchy and Abel banished divergent series from mathematics for over half a century. Their work on the foundations of the theory of series was continued by others in the 19th century, in particular by Peter Gustav Lejeune Dirichlet, Riemann, and Karl Weierstrass (German). By the last decade of the 19th century the effects of Abel's anathema began to disappear, and the work of Ernesto Cesàro (Italian), Émile Borel (French), Leopold Fejér (Hungarian), and others showed that the theory of divergent series can be put on a rigorous basis and is a source of great progress in mathematics. In the 20th century, the work of the British mathematicians Godfrey Harold Hardy and John Edensor Littlewood brought major breakthroughs and extended the field in depth. Important progress in Tauberian theorems was made by Norbert Wiener (U.S.).

FOURIER SERIES

Fourier series are a particular case of the general class of series of functions discussed above.

An infinite series of the form of a sum composed of a constant term and weighted trigonometric functions with angles x, $2x$, and so forth (see 23) is called a trigonometric series. The quantities $a_0, a_1, b_1, a_2, b_2, \cdots$ are constants and x takes values between $-\infty$ and $+\infty$. If the series converges for all values of x, and $f(x)$ denotes the sum of the series, then $f(x)$ is periodic with a period 2π, because replacing x by $x + 2\pi$ does not change the series. If both sides of $f(x) = \frac{1}{2} a_0 + (a_1 \cos x + b_1 \sin x) + \cdots$ are multiplied by $\cos nx$ or $\sin nx$ and the result is integrated over the interval $(0, 2\pi)$, then two integral ex-

An infinite series of trigonometric functions

pressions for the a_n and b_n involving the function f and a trigonometric function are obtained (see 24); this argument is purely formal (it is justified if the series converges uniformly, or even in some more general case). It suggests, however, an interesting problem. If for any periodic function $f(x)$ with period 2π the numbers a_n, b_n are computed by means of the integral formulas to form a series (see 23), does the series so defined represent the function $f(x)$? The numbers obtained from the formulas (24) are called the Fourier coefficients of $f(x)$ and the series (23) is the Fourier series of $f(x)$. The fact that, subject to suitable conditions on $f(x)$, this problem admits of an affirmative answer is one of the main achievements of the theory.

Historical development of Fourier series

Although Fourier series originated in mathematical physics, their influence upon the development of analysis has been almost as great. Very often they appear there in a more general form, as orthogonal series (see below), and provide a unifying link in such mathematical theories as differential and integral equations and analytic functions. Especially important are trigonometric (and, in particular, Fourier) series for the theory of analytic functions because for $z = e^{ix}$ the real part of the power series, which is a power series in z with complex coefficients (see 25), is the series already obtained (see 23). Thus trigonometric series are real parts of power series, and so form a bridge between real and complex functions. Aside from the value of trigonometric series as a method of investigation, they have played a great role in the historic development and clarification of various mathematical concepts, some of them quite abstract. A few examples may suffice.

Already in the days of its inception in the 18th century, the theory of Fourier series stirred controversies in connection with the notion of mathematical function. The general attitude at that time was to call $f(x)$ a function if $f(x)$ could be represented by a single analytic expression such as a polynomial, a power series, or a trigonometric series. If the graph of $f(x)$ were arbitrary—*e.g.*, a polygonal line—$f(x)$ would not have been accepted as a function. So it came as a shock to many when the discovery of Fourier series showed that many such arbitrary graphs could be represented by trigonometric series and so should be treated as functions. It took a long time before the matter was clarified, and it was not merely a coincidence that the now universally adopted definition was first formulated in 1837 in a memoir of Dirichlet, devoted to the theory of Fourier series. (Incidentally, this memoir contains the first rigorous proofs of the representation of functions by their Fourier series. The work of Fourier himself and of his predecessors was rather formal.) Another instance of application of trigonometric series to the theory of functions is due to Weierstrass, who in 1861 gave the first example of a continuous function without a derivative at any point. Weierstrass' function was given in the form of a trigonometric series, and it exemplified the general fact that trigonometric series are particularly useful in the construction of functions with various peculiarities. Another much more important example of the influence of Fourier series may be found in the history of the notion of integral. On account of previous formulas (see 24), this notion is a prerequisite to that of Fourier series. It is therefore interesting that the classical definition of integral due to Riemann (usually given in textbooks of calculus) was for the first time expounded in 1854 in his fundamental paper *Über die Darstellbarkeit einer Funktion durch eine trigonometrische Reihe* ("Concerning the Representation of Functions by Trigonometric Series"). The ideas contained in that paper influenced the work of the German mathematician Georg Cantor concerning the uniqueness of the representation of a function by a trigonometric series. Cantor had to consider various sets of points that would not interfere with the uniqueness of representation. Starting from simplest cases he tried to master the structure of such sets—sets of uniqueness (see below). He was not completely successful, and the problem was still open in the mid-1960s, but his investigations led him to the general theory of sets, and in particular of point sets, one of the greatest mathematical discoveries of the 19th century. His results have permeated and changed the aspect of many mathematical theories. The theories of measure and of integral developed later by Émile Borel and Henri-Léon Lebesgue in France, so fundamental in mathematical analysis, are based on Cantor's work on the theory of sets.

Cantor's contribution

$$(24)\qquad \begin{cases} a_n = \dfrac{1}{\pi}\displaystyle\int_0^{2\pi} f(x)\cos nx\,dx \\ b_n = \dfrac{1}{\pi}\displaystyle\int_0^{2\pi} f(x)\sin nx\,dx \end{cases}$$

$$(25)\qquad \frac{1}{2}a_0 + (a_1 - ib_1)z + (a_2 - ib_2)z^2 + \cdots$$

$$(26)\qquad s_n(x) = \frac{1}{\pi}\int_0^{2\pi} f(x+t)\frac{\sin(n+\frac{1}{2})t}{2\sin\frac{1}{2}t}\,dt$$

$$(27)\qquad \int_0^{\pi}\frac{|f(x+t)+f(x-t)-2f(x)|}{t}\,dt$$

Particularly important is the work of Lebesgue, who put Fourier series on a new basis through his theory of integral. Though more general definitions of integral are possible, and though some of them are of interest for Fourier series, Lebesgue's definition is the most appropriate and is almost universally adopted in research. In this connection a French mathematician, Arnaud Denjoy, showed with a definition of integral more general than that of Lebesgue that an everywhere-convergent trigonometric series is always the Fourier series of its sum. Important for Lebesgue's theory is the notion of set of measure zero. A point set S situated, for example, on a straight line is said to be of measure zero if S can be covered by a finite or infinite system of intervals of total length arbitrarily small. In Lebesgue's theory, sets of measure zero are considered as negligible, and if a certain property holds for all points except those forming a set of measure zero, that property is said to hold almost everywhere.

Convergence and divergence of Fourier series. If $s_n(x)$ denotes the sum of the first $n+1$ terms of the Fourier series (23), using the integral formulas (24), s_n as a function of x is a definite integral from zero to 2π with integrand formed with the function f and certain sine function expressions (see 26), a formula basic for Fourier series. It may be proved that as n increases indefinitely $s_n(x)$ tends to $f(x)$, provided the function $f(x)$ satisfies certain conditions. For example, if the graph of the function $f(x)$ in the interval $0 \leqslant x \leqslant 2\pi$ has only a finite number of maxima and minima (Dirichlet's condition), the Fourier series of $f(x)$ converges to $f(x)$ at every point of continuity of the function. At points x at which such a function is discontinuous, the Fourier series also converges, its sum being $[f(x+0)+f(x-0)]/2$ in which $f(x+0)$ represents the limit of the function at the point x from the right, and $f(x-0)$ a similar limit from the left. Another condition ensuring the convergence of the Fourier series to sum $f(x)$ is the convergence of a definite integral from zero to π with integrand formed in terms of the function f and variable x (the Dini condition; see 27). It certainly is satisfied at every point at which the function $f(x)$ is differentiable. These and all other known conditions for convergence are sufficient conditions only. Conditions that would be both necessary and sufficient (and would not be trivial tautologies) are not known. That mere continuity of the function is not sufficient to ensure the convergence of the Fourier series was shown by the German Paul David Gustave du Bois-Reymond in 1872. He constructed a continuous function for which the Fourier series diverges at some points. Whether the Fourier series of a continuous function must converge at least almost everywhere, that is, everywhere except possibly for a set

Conditions for convergence

(28) $$\sigma_n(x)=\frac{s_0(x)+s_1(x)+\cdots+s_n(x)}{n+1}$$

(29) $$\sigma_n(x)=\frac{1}{\pi(n+1)}\int_0^{2\pi} f(x+t)\frac{\sin^2 \frac{1}{2}(n+1)t}{2\sin^2 \frac{1}{2}t}\,dt$$

(30) $$\tfrac{1}{2}a_0+(a_1\cos x+b_1\sin x)r+(a_2\cos 2x+ +b_2\sin 2x)r^2+\cdots$$

(31) $$\frac{1}{2\pi}\int_0^{2\pi} f(x+t)\frac{1-r^2}{1-2r\cos t+r^2}\,dt$$

(32) $$(a_1\sin x-b_1\cos x)+(a_2\sin 2x-b_2\cos 2x)+\cdots$$

(33) $$-\frac{1}{\pi}\int_0^{\pi}\frac{f(x+t)-f(x-t)}{2\tan \frac{1}{2}t}\,dt, \text{defined as } \lim_{\epsilon\to 0}\left\{-\frac{1}{\pi}\int_{\epsilon}^{\pi}\right\}$$

(34) $$\frac{1}{\pi}\int_0^{2\pi} f^2(x)\,dx=\tfrac{1}{2}a_0^2+(a_1^2+b_1^2)+(a_2^2+b_2^2)+\cdots$$

of points of measure zero, was for a long time an unsolved and tantalizing problem. In 1966 it was shown by the Swedish mathematician Lennart Axel Edvard Carleson that it is actually so, a major achievement of mathematical analysis. More generally, Carleson showed that the Fourier series of any integrable (in the Lebesgue sense) function $f(x)$ the square $f^2(x)$ of which is also integrable must necessarily converge almost everywhere. Using his method the U.S. mathematician Richard Hunt proved in 1967 that the same conclusion holds if $|f(x)|^p$ is integrable for some p strictly greater than 1. That the result fails for $p = 1$ was shown in 1926 by the Soviet mathematician A.N. Kolmogorov, who constructed an integrable function the Fourier series of which diverges at each point. These results basically settle the problem of convergence almost everywhere of Fourier series.

Summability of Fourier series. The fact that Fourier series of continuous functions need not converge everywhere endangered the whole theory of the representation of functions by their Fourier series. The situation was salvaged by the Hungarian mathematician Leopold Fejér in 1900, who showed that the Fourier series of a continuous function $f(x)$ is summable to $f(x)$ by the method of the arithmetic mean discussed under *Divergent series*. More precisely, if $s_n(x)$ has the same meaning as above, the expression in which σ_n as a function of x is an arithmetic average of terms expressed as $s_n(x)$ (see 28) tends to $f(x)$ at every point of continuity of the function. Following this, Lebesgue proved that for every integrable function $f(x)$ the expression $\sigma_n(x)$ tends to $f(x)$ almost everywhere. This again put the theory on a firm basis, and at the same time showed that summability rather than convergence is important for Fourier series. Using formulas already given (see 26, 28) leads to an expression in which σ_n as a function of x is an integral from zero to 2π with integrand expressed in terms of f and certain sine functions (see 29), but there is one essential difference. The factor $\sin^2 \frac{1}{2}(n+1)t/2(n+1)\sin^2 \frac{1}{2}t$, called Fejér's kernel, is nonnegative, whereas the factor $\sin(n+\frac{1}{2})t/2\sin \frac{1}{2}t$, Dirichlet's kernel, is of variable sign. This positive character of Fejér's kernel is responsible for many interesting properties of the expressions $\sigma_n(x)$, and in particular for the fact that they represent $f(x)$ so much better than the $s_n(x)$ do.

Fejér's, Dirichlet's kernels; Abel's method; method of arithmetic mean

Besides the method of the arithmetic mean, many other methods of summability have been applied to Fourier series. The most important of them is the method of Abel that defines the sum of Fourier series (see 23) as the limit of the expression that is a power series in r with typical coefficient expressed in terms of cosine and sine (see 30), in which r tends to 1 through values less than 1. The last expression may be written as a definite integral from zero to 2π with integrand composed of the function f and a function involving cosine and the variable r (see 31) and is called the Poisson integral of $f(x)$. As r approaches 1, it tends to $f(x)$ at every point of continuity of the function (or almost everywhere, if $f(x)$ is merely integrable).

Conjugate series and functions. For $z = e^{ix}$ the real part of a power series (see 25) is the trigonometric series (see 23). The imaginary part is a trigonometric series, a sum of terms with typical member composed of the sine and the cosine (see 32), called the conjugate of the series (23). Suppose that there is the Fourier series of a function $f(x)$ (see 23); the conjugate series is then almost everywhere summable by the method of the arithmetic mean, or by the method of Abel, the generalized sum being equal to the definite integral from zero to π with integrand constructed of the function f and the tangent (see 33).

This integral is called the function conjugate to $f(x)$ and will be denoted by $\tilde{f}(x)$. The fact that for every integrable function $f(x)$ the integral (see 33) exists almost everywhere is a very remarkable property that is far from obvious even in the case in which $f(x)$ is continuous. Even in this special case, the integral (33) need not exist everywhere; it is a curious fact that continuous functions may have certain metric properties almost everywhere, without having them everywhere. Conjugate series are a link between trigonometric series and power series. They are also of importance for the theory of Fourier series because there seems to be a close relation between the behaviour of the partial sums of Fourier series and the behaviour of certain conjugate functions. As a matter of fact, the proofs of the theorems of Carleson and Hunt (see above) strongly use refined results on conjugate functions. The function conjugate to an integrable function $f(x)$ need not be integrable. Thus the series conjugate to the Fourier series of $f(x)$ need not be a Fourier series.

Parseval's formula. Suppose that one has the Fourier series of $f(x)$ (see 23). If in the formula $f(x) = \frac{1}{2}a_0 + (a_1\cos x + b_1\sin x) + \cdots$ both sides are squared and the result is integrated over the interval $0 \leqslant x \leqslant 2\pi$, then a definite integral from zero to 2π of the square of the function f is equal to an infinite sum with typical term composed of the square of a_k plus the square of b_k (see 34), called Parseval's formula. The above argument was formal, but rigorous proofs are available to show that the formula is valid for any function $f(x)$ such that $f^2(x)$ is integrable. Thus the Fourier coefficients $a_0, a_1, b_1, \cdots$ of any such function have the property that $\Sigma(a_n^2 + b_n^2)$ is finite. That the converse is true was proved by the mathematicians Frigyes Riesz (Hungarian) and Ernst Sigismund Fischer (Austrian-born) in 1907. The Riesz-Fischer theorem, which is one of the great achievements of the Lebesgue theory of integration, asserts that, given any sequence of numbers $a_0, a_1, b_1, a_2, b_2, \cdots$ such that the series $\Sigma(a_n^2 + b_n^2)$ converges, there is always a function $f(x)$ having the numbers $a_0, a_1, b_1, \cdots$ as Fourier coefficients and such that $f^2(x)$ is integrable. This function is unique. Thus, the theorems of Parseval and of Riesz-Fischer give a very simple characterization of functions with integrable squares in terms of their Fourier coefficients. There is no other class of function that has a characterization of comparable simplicity and completeness.

The Riesz-Fischer theorem

Uniqueness of representation by trigonometric series. Can a function be represented by two different and everywhere-convergent trigonometric series? By taking the difference of two such series, the problem may be restated as follows. Can a trigonometric series not identically zero (*i.e.*, a series the coefficients a_n, b_n of which are not all zero) converge everywhere to zero? That this is impossible was proved by Cantor. His result is the simplest theorem on the uniqueness of the representation of a function by a trigonometric series. A point set S situated in the interval $0 \leqslant x \leqslant 2\pi$ is called a set of uniqueness (set U, for brevity) if every trigonometric series convergent to zero outside S is identically zero. Otherwise, the set S will be called a set of multiplicity

(set M), and there will be trigonometric series convergent to zero outside S, but not identically zero. Generalizing the result just stated, Cantor proved that every set S consisting of a finite number of points, and even certain infinite sets, are sets U. It may be shown that every set U must be of measure zero, but it is remarkable that among sets of measure zero are found sets U as well as sets M. The problem of characterizing sets U and M in structural terms was still unsolved in the mid-1970s. One of its implications is that from the point of view of trigonometric series the notion of set of measure zero is too general, and that special classes of such sets may be of particular importance.

Trigonometric polynomials

Trigonometric interpolation. A finite sum, a sum from $k = 1$ to n of weighted trigonometric terms (see 35), is called a trigonometric polynomial of order n (or less, if both a_n and b_n are zero). A trigonometric polynomial of order n has $2n + 1$ coefficients $a_0, a_1, b_1, \cdots, a_n, b_n$, and it can be shown that if $2n + 1$ distinct points (see 36) are fixed in the interval $0 \leqslant x < 2\pi$, there is always a trigonometric polynomial of order n or less that takes prescribed values at these points; moreover, this polynomial is unique. Consider now a function $f(x)$ defined on the interval $(0, 2\pi)$ and take the trigonometric polynomial $T(x)$ of order n that coincides with $f(x)$ at certain points (see 36). Such a polynomial is called an interpolating polynomial for $f(x)$. Suppose now that n increases indefinitely; does the polynomial $T(x)$ approach the function $f(x)$ at all points? This must depend on the geometric configuration of the points (see 36) and the answer is not known in full generality. If, however, the points (see 36) are assumed distributed equally to form an arithmetic progression with difference $2\pi/(2n + 1)$, and if the function $f(x)$ satisfies some smoothness conditions (*e.g.*, is differentiable), then the answer is affirmative, and the whole problem has points in common with the theory of Fourier series.

(35) $$T(x) = \tfrac{1}{2}a_0 + \sum_{k=1}^{n} (a_k \cos kx + b_k \sin kx)$$

(36) $$x_0, x_1, x_2, \cdots, x_{2n}$$

(37) $$\int_a^b \phi_m \phi_n \, dx = 0, \quad \text{if } m \neq n; \qquad \int_a^b \phi_n^2 \, dx = \lambda_n > 0$$

(38) $$c_n = \frac{1}{\lambda_n} \int_a^b f(x)\phi_n(x) \, dx$$

(39) $$\int_a^b \phi_m \bar{\phi}_n \, dx = 0, \quad \text{if } m \neq n, \int_a^b |\phi_n|^2 \, dx = \lambda_n > 0$$

(40) $$c_n = \frac{1}{\lambda_n} \int_a^b f(x)\, \bar{\phi}_n(x) \, dx$$

(41) $$\sum_{n=-\infty}^{+\infty} c_n e^{inx}, \text{ in which } c_n = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-inx} \, dx$$

(42) $$c_0 + \sum_{n=1}^{\infty} (c_n e^{inx} + c_{-n} e^{-inx})$$

(43) $$\int_0^{\infty} \{a(u) \cos ux + b(u) \sin ux\} \, du$$

(44) $$\begin{cases} a(u) = \dfrac{1}{\pi} \displaystyle\int_{-\infty}^{+\infty} f(t) \cos ut \, dt \\ b(u) = \dfrac{1}{\pi} \displaystyle\int_{-\infty}^{+\infty} f(t) \sin ut \, dt \end{cases}$$

(45) $$\sum_{m=1}^{\infty} \sum_{n=1}^{\infty} c_{m,n} \phi_m(x) \psi_n(y)$$

Orthogonal series. Fourier series as defined above are a special class of more general series. Suppose that in an interval $a \leqslant x \leqslant b$ there is a fixed set of functions $\phi_1(x)$, $\phi_2(x), \cdots$ with the integrals involving the product of the functions $\phi_m \phi_n$, being zero if m is not equal to n and being a constant, λ_n, if m equals n (see 37) for all m and n. Such sets of functions are called orthogonal. If a function $f(x)$ has a representation $f(x) = c_1\phi_1(x) + c_2\phi_2(x) + \cdots$, then, multiplying both sides by $\phi_n(x)$, integrating over x in the interval (a, b), and using orthogonality (see 37) formally leads to the equation in which c_n is a definite integral from a to b involving the function f and the function ϕ_n (see 38) for all n. Conversely, given any function $f(x)$, the latter formulas may be used to construct the series $c_1\phi_1(x) + c_2\phi_2(x) + \cdots$, and it may be asked whether this series actually represents $f(x)$. This last series is called the Fourier series of $f(x)$ with respect to the orthogonal system $\phi_1, \phi_2, \cdots$. Trigonometric Fourier series are obtained if for $\phi_1, \phi_2, \cdots$ the system 1, cos x, sin x, cos $2x$, sin $2x$, $\cdots$ is taken, and for (a, b) the interval $(0, 2\pi)$. Many other orthogonal systems are of considerable importance for mathematics.

In the definition (see 37) of orthogonality, it was tacitly assumed that the functions were real-valued. If the functions $\phi_1, \phi_2 \cdots$ are complex-valued, the definition (37) must be slightly modified, and certain requirements must be met concerning two integrals involving ϕ_m and ϕ_n (see 39), in which $\bar{\phi}_n$ is the conjugate (in the sense of complex numbers) to ϕ_n. The Fourier series is, as before, $c_1\phi_1(x) + c_2\phi_2(x) + \cdots$, but now the Fourier coefficients are c_n as a definite integral from a to b of the function f and a weighted term composed of ϕ_n (see 40). The most important special case is that of the system e^{inx} ($n = 0, \pm 1, \pm 2, \cdots$) that is readily seen to be orthogonal over $(0, 2\pi)$, with all $\lambda_n = 2\pi$. This system is called the complex trigonometric system. The Fourier series of a function $f(x)$ defined in the interval $(0, 2\pi)$ is a sum with integer index ranging from minus infinity to plus infinity with summand of exponential type and coefficient c_n, the coefficient determined by an integral (see 41). The latter series is called the complex Fourier series of $f(x)$ and is meant as a sum including a constant term and terms of exponential type and negative exponential type (see 42). It can be shown that this series is nothing else but the Fourier series (23) with coefficients given by the familiar formula (24). Hence, except for notation, complex Fourier series are the same as ordinary trigonometric Fourier series, but the complex form sometimes simplifies proofs and leads to more elegant formulations.

The complex Fourier series

Fourier integrals. Fourier series are used for the representation of periodic functions. For the study of a nonperiodic function $f(x)$, the Fourier integral of $f(x)$ is used. It has the form of an integral from zero to infinity with integrand composed of functions a and b and functions cosine and sine (see 43), in which $a(u)$ and $b(u)$ are defined by the formulas in which a as a function of u is expressed as an integral involving f and the cosine, b as a function of u is expressed as an integral involving f and the sine (see 44). The integral (43) is analogous to the series (23), the only difference being that instead of summation there is integration. The functions $a(u)$ and $b(u)$ are analogues of Fourier coefficients and are called the Fourier transforms of $f(x)$. Under very general conditions, the integral (43) actually represents $f(x)$, the proofs being similar to the corresponding proofs in the case of Fourier series.

Multiple Fourier series. If $\phi_1(x), \phi_2(x), \cdots$ is an orthogonal system over an interval $a \leqslant x \leqslant b$, and if $\psi_1(y), \psi_2(y), \cdots$ is another orthogonal system over an interval $c \leqslant y \leqslant d$, then it may be shown that the double system $\phi_m(x)\psi_n(y)$ (in which $m, n = 1, 2, \cdots$) is orthogonal over the rectangle $a \leqslant x \leqslant b$, $c \leqslant y \leqslant d$ of the plane, and a function $f(x, y)$ defined in this rectangle can be developed into a double Fourier series, a double sum with summand composed of a product of a constant, the function ϕ_m, and the function ψ_n, two variables x and y being involved (see 45). The most interesting case occurs when both systems ϕ_m and ψ_n are complex trigonometric systems. The Fourier series is then a double sum with summand com-

posed of a coefficient with double index and exponential with complex argument depending upon x and y (see 46). Similarly, triple Fourier series can be considered, and so on. The problems concerning the behaviour of multiple Fourier series are usually much more difficult than those of ordinary Fourier series.

The integral (43) can also be written in the form of a definite integral from minus infinity to plus infinity with integrand of exponential complex type and coefficient determined by an integral again of exponential type (see 47), which is an analogue of a complex Fourier series. Multiple Fourier integrals are also considered (analogues of multiple Fourier series) and have become a very important tool in various branches of analysis.

HARMONIC ANALYSIS AND INTEGRAL TRANSFORMS

It usually happens in mathematical analysis that certain given functions are to be operated on, and other functions or numbers are thus obtained. Often, the operations used involve integration, and here is found a very general class of operators, the so-called integral transforms. The simplest integral transform is the operation of integrating. If f is a function defined on the interval $[a, b]$, a new function F on $[a, b]$ can be defined (see 48). A somewhat more general example arises if f is multiplied by another function $h(y)$ before integrating: F is expressed as a definite integral from a to b with integrand f multiplied by h, each being functions of the variable of integration (see 49).

The following situation already embraces quite a number of important integral transforms: if the function $f(y)$ is given in the interval $a \leqslant y \leqslant b$, and $K(x, y)$ is a function of two variables, defined when $a \leqslant y \leqslant b$ and $c \leqslant x \leqslant d$, then Φ as a function of x is an integral from a to b with integrand a product of terms including K with two arguments and the function f with a single argument (see 50). This is an integral transform of $f(y)$ by means of the function K, which is called the kernel. [Here K is regarded as fixed, and the integral transform (50) is regarded as an operation that can be performed on any one of a large class of functions f, yielding for each appropriate f, a new function Φ.] The operation F as a function of f equal to an indefinite integral of the function f (see 51) is merely a special case of the integral transform (50), in which $K(x, y) = 1$ if $y \leqslant x$ and $K(x, y) = 0$ if $y > x$. By considering a variety of kernels K, a number of important classical integral transforms are obtained, some of which will be discussed below.

Function K, the kernel

(46) $$\sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} c_{m,n} e^{i(mx+ny)}$$

(47) $$\int_{-\infty}^{+\infty} c(u) e^{ixu}\, du, \text{ in which } c(u) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} f(t) e^{-iut}\, dt$$

(48) $$F(x) = \int_a^x f(y)\, dy$$

(49) $$F(x) = \int_a^x h(y) f(y)\, dy$$

(50) $$\Phi(x) = \int_a^b K(x, y) f(y)\, dy, \qquad c \leqslant x \leqslant d$$

(51) $$F(x) = \int_a^x f(y)\, dy$$

(52) $$K(x, y) = k(x - y)$$

(53) $$\Phi(x) = \int_{-\infty}^{\infty} k(x - y) f(y)\, dy$$

(54) $$\Phi = k * f$$

(55) $$\hat{f}(x) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} e^{-ixy} f(y)\, dy$$

(56) $$K(x, y) = (2\pi)^{-1/2} e^{-ixy}$$

(57) $$|e^{-ixy}| = 1$$

The definitions given above may be generalized in various ways, one of which is to use functions f of several variables (see below). In many important special cases of the integral transform (50), the intervals $[a, b]$ and $[c, d]$ are replaced by the whole real line $(-\infty, \infty)$, or a half-line, such as $[a, \infty)$.

At this stage in the discussion, two technical points should be mentioned. First of all, there may be a restriction to real-valued functions $f(y)$ and $K(x, y)$, or more generally to functions $f(y)$ taking complex values, $f(x) = u(x) + iv(x)$, and likewise for the kernel K. It is preferable to adopt the latter point of view, especially because operations performed on real-valued functions often lead to complex-valued functions anyway. A more delicate point is the notion of integral to be employed. In many cases, especially in practical applications, it is enough to deal with the classical Riemann integral, or with the improper Riemann integral, for example, when functions are integrated over infinite intervals. In most instances, however, coherence and generality are best achieved by introducing the notion of integral due to Lebesgue (see ANALYSIS, REAL: *Measure and integral calculus*). To help the reader not familiar with Lebesgue's theory, this point will not be stressed, and only in the cases in which the mathematical results depend most strongly on Lebesgue measure or integration will the fact be indicated.

Convolution. Before passing to various special cases, one very important type of integral transform will be indicated. If $f(y)$ is a function defined in $(-\infty, \infty)$, and $k(x)$ is another (fixed) function defined in $(-\infty, \infty)$, then the kernel K as a function of x and y equal to a function of the difference of the arguments (see 52) may be considered. In this case the transform (50) takes the form in which Φ as a function of x is a definite integral over the entire real line of the product of the function k and the function f (see 53). This integral (assuming it converges) is called the convolution of the functions f and k (see 54). The centred asterisk symbol for convolution resembles that of the centred dot for multiplication; and there is actually a close connection between the two notions, as will be indicated below. The convolution is an integral transform occurring often in analysis. Usually, after fixing the kernel k, a class of functions f is considered, and the effect of the transform investigated, in order to determine which properties of functions f are preserved and which properties changed, and in what way. Below are considered a number of important special cases.

The Fourier transform. As a second example of an integral transform, the Fourier transform is now introduced, initially of a function of a single variable. If $f(y)$ is a given function defined in the interval $(-\infty, \infty)$, $\hat{f}$ is a definite integral over the entire real line in which the integrand is composed of a product of a negative complex exponential and the function f (see 55), in which x is a real variable and $\hat{f}$ is called the Fourier transform of f. The Fourier transform is an integral transform of the form already discussed (see 50), with kernel of the form in which K as a function of two variables is of exponential type with negative complex unit multiplied by the product of the variables in the argument (see 56). In view of the fact that the absolute value of the negative complex exponential is unity (see 57), the integral defining $\hat{f}(x)$ converges for all values of x and represents a continuous function of x, provided f is absolutely integrable over $(-\infty, \infty)$. In this case it can also be shown that $\hat{f}(x)$ tends to zero as x tends to $\pm\infty$. The Fourier transform also exists in some cases in which f is not necessarily absolutely integrable on $(-\infty, \infty)$. Thus $\hat{f}$ is an integral

transform that makes sense for a large class of functions f; the study of this transform and its generalizations is called harmonic analysis. It should be added that certain variants of the definition (55) are sometimes useful and frequently occur in the literature. Thus the Fourier transform is sometimes defined by the equation in which $\hat{f}$ is a definite integral extended from minus to plus infinity of the complex negative exponential in which the variables appear as products (see 58), or by omission of the minus sign in the exponent. But adherence here is to the definition (55).

Fourier transforms and inversions

The main questions to be studied are: What are the properties of $\hat{f}$ in terms of those of f, or of f in terms of $\hat{f}$; and also, can f be found if $\hat{f}$ is known (the problem of inversion of the Fourier transform)? Typical properties of interest are the size and the smoothness of a function. The former question is considered first.

The simplest way of measuring the size of a function is by giving its bounds, $A \leqslant f(x) \leqslant B$. (Upper and lower bounds may be given if f is real-valued; bounds for $|f|$ may also be considered if f is complex-valued.) Other estimates, however, especially in the case of unbounded functions, can also be used. If p is a fixed positive number, $L^p[a, b]$ represents the class of functions f on $[a, b]$ for which the definite integral from a to b of the pth power of the absolute value of f is finite (see 59). In this case, the pth root of the integral from a to b of the pth power of the absolute value of f (see 60) may serve as a measure of the size of f. It will be called the pth norm, or L^p-norm, of f, denoted by $\| f \|_p$. (For the sake of economy, $[a, b]$ are not displayed in the symbol for the norm of f.) The functions f and g may be considered close to each other if $\| f - g \|_p$ is small. If $\| f - g \|_p$ is regarded as a measure of the distance from f to g, it is natural to expect that the pth norm of the difference between f and g is less than or equal to the sum of two such norms involving a third function h (the triangle inequality; see 61) for any three functions f, g, and h. Indeed, this inequality holds, provided $p \geqslant 1$. In the vast majority of uses of L^p in analysis, it is customary to consider only the case $p \geqslant 1$; this will be the assumption below.

It would probably be beneficial to acquaint the reader further with $L^p[a, b]$ (see 62) by means of some simple examples. On a finite interval, say $0 < x < 1$, every bounded function f belongs automatically to L^p, because the integral from 0 to 1 of the pth absolute value of f is less than or equal to a constant, which is less than infinity (see 63), if M is an upper bound for $|f|$. In addition, certain unbounded functions also belong to L^p. If $f_\delta(x) = x^{-\delta}$, with $0 < x < 1$, and $\delta > 0$ fixed, then each f_δ is unbounded; as x tends to zero, $f_\delta(x)$ tends to infinity. If f_{δ_1} is compared with f_{δ_2} where say $\delta_1 > \delta_2$, however, then it is found that $f_{\delta_1}(x)$ goes to infinity as x tends to zero more rapidly than does $f_{\delta_2}(x)$. Now it may be shown that f_δ belongs to L^p for $p < 1/\delta$, but not for $p \geqslant 1/\delta$. Thus, an unbounded function may belong to L^p, but if so, it may not tend to infinity too rapidly. As p increases, more and more functions are ruled out as being too large for L^p, so that fewer functions belong to L^p.

Use of L^2 functions

Having considered the L^p-classes in general, an important special case can now be mentioned, namely, $p = 2$. The class L^2 is of fundamental importance in analysis. To understand the special significance of L^2, it is possible to begin with the formula for the distance between two functions f and g in the L^2 sense for the range a to b, the squared norm identified with the difference of two functions expressed as a square root of the integral of the square of the function (see 64). This equation is very similar to the Pythagoras formula $\| X - Y \| = (\sum_{j=1}^{n} | x_j - y_j |^2)^{1/2}$ for the distance between two points (see 65) in n-dimensional Euclidean space. This similarity suggests a way of understanding $L^2[a, b]$: functions f, g, etc., may be regarded as points that lie in a space $H = L^2[a, b]$, and the geometrical properties of H may be studied almost as if H were merely Euclidean space. Because of the similarity of the distance formulas noted above, virtually all of classical Euclidean and coordinate (analytic) geometry can be carried out in the setting of $L^2[a, b]$. Furthermore, the geometric insights thus gained into the structure of L^2 can be applied to numerous concrete problems throughout analysis. For other L^p-classes ($p \neq 2$), the present state of geometrical understanding is relatively crude, and consequently many simple theorems on L^2 functions become more difficult when generalized to L^p. As a result, it is preferable to use L^2 as much as possible.

In the above, the notion of the Fourier transform was introduced. For which functions f does the definition (55) of f make sense? A great deal of work has been done on this problem, and a few highlights can be mentioned briefly. Suppose f belongs to $L^p(-\infty, \infty)$. If $p = 1$, the integral in the definition of f (see 55) exists as an absolutely convergent integral for each point x. The situation is somewhat different if p is greater than 1. Restriction may be made initially to $p = 2$, which is in many ways the central case of the theory.

To define the Fourier transform of an L^2 function, consider the partial integral in which $\hat{f}_M$ as a function of x is a definite integral from minus to plus M in which the integrand is a product of a negative exponential and the function f (see 66). (This integral is absolutely convergent for all finite M.) It can be shown that $\hat{f}_M$ belongs to $L^2(-\infty, \infty)$ for each fixed M, and that as M tends to infinity, the $\hat{f}_M$ tend to a limit in an appropriate sense. One way to express precisely the convergence of the $\hat{f}_M$ to a limit is as follows: There is one (and only one) function $\hat{f}$ in $L^2(-\infty, \infty)$ for which the distance $\| \hat{f}_M - \hat{f} \|_2$ tends to zero as M tends to infinity; the function $\hat{f}$ is called the Fourier transform of f. Thus, the Fourier transform makes sense for all L^2 functions.

The action of the Fourier transform on $L^2(-\infty, \infty)$ has been exhaustively studied by making use of the geometrical properties of $L^2(-\infty, \infty)$ already alluded to above. The main point of the L^2-theory is the celebrated Plancherel formula, which may be stated thus: a function f in $L^2(-\infty, \infty)$ is considered. By what has been said above, $\hat{f}$ also belongs to L^2. By means of the Plancherel formula, it is possible to predict exactly the size of $\hat{f}$ in terms of that of f. The formula states that two integrals are identified, each extended over the entire real line and each with integrands of the type absolute value of a function squared, one function being $\hat{f}$ and the other f (see 67). Thus f and $\hat{f}$ have the same L^2-norms. A closely re-

Plancherel formula

(58) $$\hat{f}(x) = \int_{-\infty}^{+\infty} e^{-2\pi i x y}\, dy$$

(59) $$\int_a^b |f(x)|^p\, dx < \infty$$

(60) $$\left(\int_a^b |f(x)|^p\, dx \right)^{\frac{1}{p}}$$

(61) $$\|f - g\|_p \leqslant \|f - h\|_p + \|h - g\|_p$$

(62) $L^p[a, b]$, the set of functions such that $\|f\|_p < \infty$

(63) $$\int_0^1 |f(x)|^p\, dx \leqslant \int_0^1 M^p\, dx = M^p < \infty$$

(64) $$L^2[a, b] : \|f - g\|_2 = \left(\int_a^b |f(x) - g(x)|^2\, dx \right)^{\frac{1}{2}}$$

(65) $X = (x_1, x_2, \cdots, x_n)$ and $Y = (y_1, y_2, \cdots, y_n)$

(66) $$\hat{f}_M(x) = (2\pi)^{-\frac{1}{2}} \int_{-M}^{M} e^{-ixy} f(y)\, dy$$

(67) $$\int_{-\infty}^{\infty} |\hat{f}(x)|^2\, dx = \int_{-\infty}^{\infty} |f(x)|^2\, dx$$

lated theorem asserts that an arbitrary function g in $L^2(-\infty, \infty)$ can be written as the Fourier transform of some function f in L^2, $\hat{f} = g$, and that there is only one f that serves. It will be explained presently how to calculate f for any given g. These results determine completely the behaviour of the Fourier transform on $L^2(-\infty, \infty)$.

The solution to the problem of inversion of the Fourier transform will now be stated; *i.e.*, an explicit formula will be given for a function f in terms of its Fourier transform $\hat{f}$. Define the function $\check{g}(x)$ such that $\check{g}$ is definite integral from minus infinity to plus infinity of a product of a complex exponential and the function g (see 68). This operation differs from the Fourier transform only in the negative sign in the exponential, and is called the inverse Fourier transform. Just as for the Fourier transform, the integral in the definition of $\check{g}$ (see 68) is absolutely convergent if g is absolutely integrable, while if g belongs to $L^2(-\infty, \infty)$ partial integrals (see 69) may be formed and the limit obtained in the sense described above in connection with f_M. Just as was seen above, the inverse Fourier transform sends L^2 functions to L^2 functions and satisfies an equation in which two integrals are equal, each extended over the entire real line, each having as integrand an absolute value of a function squared, one function being g and the other $\check{g}$ (see 70).

The connection between g and the problem of inverting the Fourier transform is as follows: Every function f in $L^2(-\infty, \infty)$ is the inverse Fourier transform of its Fourier transform $\hat{f}$. In other words, written formally, $\hat{f}$ as a function of x is a definite integral extended over the entire real line of a product in which factors are a complex exponential and the function f (see 71), in which the non-absolutely convergent integral is defined by the procedure explained above. This formula, which is called the Fourier inversion formula, solves the problem of inverting the Fourier transform and thus completes coverage of the relationship between an L^2 function and its Fourier transform.

Harmonic oscillations

For each y, the expression $e^{ixy} = \cos(xy) + i \sin(xy)$ represents a harmonic oscillation in x. Hence, for each y, the integrand in the inverse Fourier transform (see 71) is a harmonic oscillation in x, and so the formula for the inverse (71) gives a decomposition of f into a sum of harmonic oscillations. For this reason the theory of Fourier transforms is often called harmonic analysis.

A few words should be said about the Fourier transform in other classes L^p. If f belongs to L^p in which $1 < p < 2$, then f can be decomposed into a sum of two functions $f = f_1 + f_2$, in which f_1 is in L^1 and f_2 is in L^2. For example, f_1 may be set equal to f wherever $|f| \geqslant 1$ and equal to zero elsewhere; and similarly f_2 set equal to f in which $|f| < 1$ and equal to zero elsewhere. Both f_1 and f_2 are in L^p, but because clearly f_1 in absolute value and f_2 in absolute value squared are less than or equal to corresponding pth powers of absolute values (see 72); it follows that f_1 is in L^1, f_2 is in L^2, and therefore the Fourier transforms f_1 and f_2 are defined; $\hat{f}$ may be defined as the sum $\hat{f}_1 + \hat{f}_2$. It can be shown that the new definition of $\hat{f}$ does not conflict with the earlier definitions of $\hat{f}$ for f in L^1 or L^2, nor does it depend on the particular way of splitting f into a sum $f_1 + f_2$. Thus the Fourier transform of an L^p function exists if $1 \leqslant p \leqslant 2$. The argument given above, however, presupposes that $p \leqslant 2$, and fails for $p > 2$. In fact, the Fourier transform f of an L^p function (in which $p > 2$) in general cannot be given by a function, but only by a generalized function or distribution (see ANALYSIS, FUNCTIONAL: *Generalized functions, or theory of distributions*). There will be no further discussion of the problem of interpreting $\hat{f}$ for f in L^p when $p > 2$, apart from the remark that even then in some cases $\hat{f}$ will be a function.

A point of interest is the size of $\hat{f}$ if f is in L^p, in which $1 \leqslant p \leqslant 2$. Given p, the conjugate index p' can be defined by $1/p + 1/p' = 1$. It is notable that if $1 < p < 2$, then $2 < p' < \infty$. An important result of the theory of Fourier integrals states that if f belongs to L^p, in which $1 < p \leqslant 2$, then f belongs to $L^{p'}$ and the inequality in which the p' norm of f is less than or equal to the pth norm of f (the inequality of Hausdorff-Young; see 73)

$$\check{g}(x) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} e^{ixy}\, g(y)\, dy \tag{68}$$

$$\check{g}_M(x) = (2\pi)^{-\frac{1}{2}} \int_{-M}^{M} e^{ixy}\, g(y)\, dy \tag{69}$$

$$\int_{-\infty}^{\infty} |\check{g}(x)|^2\, dx = \int_{-\infty}^{\infty} |g(x)|^2\, dx \tag{70}$$

$$f(x) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} e^{ixy} \hat{f}(y)\, dy \tag{71}$$

$$|f_1| \leqslant |f_1|^p \quad \text{and} \quad |f_2|^2 \leqslant |f_2|^p \tag{72}$$

$$\|\hat{f}\|_{p'} \leqslant \|f\|_p \tag{73}$$

$$\widehat{(f')}(x) = ix\hat{f}(x) \tag{74}$$

$$\widehat{(f^{(k)})}(x) = (ix)^k \hat{f}(x) \tag{75}$$

$$x = (x_1, x_2, \cdots, x_n) \tag{76}$$

$$y = (y_1, y_2, \cdots, y_n) \tag{77}$$

holds. Also the Fourier inversion formula (71), suitably interpreted, holds for f in L^p. Only in the case $p = 2$ does $p = p'$, and so there is a complete symmetry between functions and their Fourier transforms. This is one more reason for the importance of L^2.

Consideration will now be given to the effect of smoothness properties of a function f upon its Fourier transform $\hat{f}$. If f has a continuous derivative f', and the latter has a Fourier transform $(\hat{f'})$, then, in the definition (see 55) of $(\hat{f'})$, it is possible to integrate by parts formally, with the assumption that the boundary terms tend to zero at infinity. Thus is obtained the transform of the derivative of f evaluated as a function of x equal to ix times the transform of f as a function of x (see 74) for all x in $(-\infty, \infty)$. The formal manipulations can be rigorously justified in a large number of cases. Assuming that f has a kth derivative $f^{(k)}$, and repeating the argument, it is found that the transform of the kth derivative of f as a function of x is the kth power of the product of i and x times the transform of f as a function of x (see 75). If $f^{(k)}$ is absolutely integrable over $(-\infty, \infty)$, then because $(\widehat{f^{(k)}})$ must be bounded, it can be seen that $\hat{f}(x)$ must tend to zero at least as rapidly as $|x|^{-k}$ as x tends to infinity. Thus, differentiability of a function f is reflected in the rapid decrease of its Fourier transform $\hat{f}(x)$ as x tends to infinity. This relationship may be understood intuitively as follows: As was already observed, the Fourier inversion formula (see 71) exhibits a function f as a sum of harmonic oscillations e^{ixy}. For large y, the exponential e^{ixy} oscillates very rapidly as a function of x. If $f(x)$ itself is smooth and does not oscillate rapidly, then that component of f that does consist of very rapid oscillations must be extremely small. Thus, $\hat{f}(y)$ must go rapidly to zero, as y tends to infinity.

Extension of the Fourier transform to functions of several variables

The notion of Fourier transform that was initially considered for functions of a single variable can be extended to functions of several variables. If x denotes a point in the n-dimensional Euclidean space R^n, then x can be treated as an n-tuple of real numbers (the coordinates of x; see 76). Correspondingly, $f(x)$ may be written for a function of n real variables $f(x_1, \cdots, x_n)$. Similarly, the vector y has n components y_k (see 77) for another point of Euclidean space. It should be noted that in the context of n-dimensional space, x_j, y_k, etc., will represent real numbers, while x, y, etc. (without indices), will denote points of Euclidean space; $x + y$ is defined as the n-tuple in which a vector has n components, each of which is a

sum of two terms x_k and y_k (see 78), and αx as the n-tuple in which a vector has n components, each of which has as a factor α (see 79), which is a real number; $x \cdot y$ will be used for the expression that is a sum of products of components, a typical summand being $x_k y_k$ (see 80), and $|x|$ for the square root of the sum of squares of components equal to the square root of the scalar product of the vector x by itself (see 81). (In geometric terms, $|x|$ is simply the distance from the point x to the origin.) Given a function $f(y)$ defined in R^n, the Fourier transform of f will be called the function in which the transform of f as a function of x is an integral over n-dimensional Euclidean space with complex exponential in the integrand involving a scalar product and with integrand including as a factor $f(y)$ (see 82), with dy representing $dy_1 dy_2 \cdots dy_n$. For $n = 1$, this definition coincides with that previously given. Just as in the one-variable case, the integral (82), suitably interpreted, is well defined, and the Fourier transform has all the various properties mentioned above for $n = 1$. In particular, the Plancherel formula holds and the Fourier inversion formula is obtained, in which $f(x)$ is an integral over n-dimensional Euclidean space with the integrand containing as factors the transform of f and the complex exponential involving a scalar product (see 83).

Analogous to the previous formula (75) above is the key equation that is the transform of the partial derivative of f with respect to one of the variables equal to a product of i times the variable times the transform of f (see 84), which will later be applied to partial differential equations.

Examples of Fourier transforms. A few examples of Fourier transforms will be given, largely restricted to the one-variable case. (1) If $f(x) = 1$ for x belonging to a finite interval $[a, b]$ and $f(x) = 0$ for x outside this interval, then the transform of f is proportional to a difference of exponentials times the reciprocal of i times x (see 85). (2) The function $f(x)$ is called a triangular function if $f = 0$ outside an interval $[a, b]$, $f = 1$ at the centre point $c = \frac{1}{2}(a + b)$, and is linear in the intervals $[a, c]$ and $[c, b]$. Then the transform of f is proportional to a product of three functions of x, one being algebraic, one being exponential, and one being trigonometric (see 86), in which $d = b - a$ is the length of $[a, b]$. (3) The expression in which f as a function of x is a negative exponential with the square of x over 2 appearing in the argument (see 87) is a function basic in the calculus of probability (see PROBABILITY, THEORY OF). Then the transform of f is the same as f, namely, the negative exponential with the square of x over 2 in the argument (see 88), and so f is its own Fourier transform. Slightly more generally, by setting f_δ as a function of x proportional to a negative exponential in which the reciprocal of the square of δ appears in the argument and in which the reciprocal of δ is the proportionality factor (see 89), then the transform of f_δ is a negative exponential in which the square of δ is a factor with the square of x in the argument (see 90); here can be seen an instance of an important general principle: As f becomes more concentrated, $\hat{f}$ spreads out. In the present example, the graph of f_δ is a bell-shaped curve of width δ, while the graph of $\hat{f}_\delta$ has width $1/\delta$. (4) By selecting suitable f, a large number of functions important in analysis are obtained as $\hat{f}$. Consider only one example. Set f as a function of x equal to the difference of 1 and x raised to the $\alpha - \frac{1}{2}$ power (see 91) if $-1 < x < 1$, and $f(x) = 0$ otherwise. Then J_α as a function of x includes the transform of f as a factor and x raised to the α power as a factor (see 92), which is called the Bessel function of order α. (Here c_α is a constant the value of which is of no present concern.) This function occurs often in mathematical physics and elsewhere. (5) Finally, one example in higher dimensions will be given. If $k(x)$ is defined on R^n by an equation such that $k(x)$ is the absolute value of x raised to the $\alpha - n$ power (see 93), in which $0 < \alpha < n$, and the length of x equals the square root of the sum of squares of the components (see 94), then $\hat{k}(x) = c_\alpha |x|^{-\alpha}$, in which c_α is a certain positive constant. Use will be made of the example later.

(78) $(x_1 + y_1, x_2 + y_2, \cdots, x_n + y_n)$

(79) $(\alpha x_1, \alpha x_2, \cdots, \alpha x_n)$

(80) $x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$

(81) $(x_1^2 + x_2^2 + \cdots + x_n^2)^{\frac{1}{2}} = (x \cdot x)^{\frac{1}{2}}$

(82) $\hat{f}(x) = (2\pi)^{-\frac{n}{2}} \int_{R^n} e^{-ix \cdot y} f(y)\, dy$

(83) $f(x) = (2\pi)^{-\frac{n}{2}} \int_{R^n} e^{ix \cdot y} \hat{f}(y)\, dy$

(84) $\widehat{\frac{\partial}{\partial x_j} f}(x) = i x_j \hat{f}(x)$

(85) $\hat{f}(x) = (2\pi)^{-\frac{1}{2}} (ix)^{-1} [e^{ibx} - e^{iax}]$

(86) $\hat{f}(x) = (2\pi)^{-\frac{1}{2}}\, 2x^{-2} e^{icx} \left[\left(\cos \frac{d}{2} x \right) - 1 \right]$

(87) $f(x) = e^{-\frac{x^2}{2}}$

(88) $\hat{f}(x) = \exp(-\frac{1}{2}x^2) = f(x)$

(89) $f_\delta(x) = \delta^{-1} \exp\left(-\frac{1}{2} \frac{x^2}{\delta^2} \right)$

(90) $\hat{f}_\delta(x) = \exp\left(-\frac{1}{2} \delta^2 x^2 \right)$

(91) $f(x) = (1 - x)^{\left(\alpha - \frac{1}{2}\right)}$

(92) $J_\alpha(x) = c_\alpha x^\alpha \hat{f}(x)$

(93) $k(x) = |x|^{\alpha - n}$

(94) $|x| = (x_1^2 + \cdots + x_n^2)^{1/2}$

(95) $\mathfrak{L}f(x) = \int_0^\infty e^{-xy} f(y)\, dy$

(96) $\hat{f}(x) = (2\pi)^{-\frac{1}{2}}\, \mathfrak{L}(ix)$

Variants of the Fourier transform. In addition to the Fourier transform, harmonic analysis deals with other closely related integral transforms. It is sufficient here to list only a few. (1) the Laplace transform: If $f(y)$ is a function defined on the interval $(0, \infty)$, the Laplace transform of f, denoted by $\mathfrak{L}f$, is given by an equation such that the Laplace transform of f as a function of x is a definite integral from zero to infinity in which the negative exponential appears as a factor in the integrand and the product of x and y constitutes the argument in the exponential, f being the remaining factor in the integrand (see 95), for all real or complex values of x for which the defining integral converges. It may be observed that extension of the domain of f to the whole real line by setting $f(x) = 0$ for $x < 0$ gives formally the equation in which the transform of f is proportional to the Laplace transform of i times x (see 96), for all real values of x. Thus, the Laplace transform is a natural extension of the Fourier transform, and for this reason the former is sometimes called the Fourier-Laplace transform. This integral transform is rather useful in studying ordinary differential equations and also in the theory of analytic functions (see DIFFERENTIAL EQUATIONS: *Linear*

The Laplace transform

$$(97)\qquad (f * g)(x)=\int_{R^n} f(x-y)g(y)\,dy$$

$$(98)\qquad \int_{R^n} f(z)g(x-z)\,dz$$

$$(99)\qquad (\alpha f_1+\beta f_2) * g=\alpha(f_1 * g)+\beta(f_2 * g)$$

$$(100)\qquad \widehat{f * g}=\hat{f}\cdot\hat{g}$$

$$(101)\qquad \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1}\left(\frac{\partial}{\partial x_2}\right)^{\alpha_2}\cdots\left(\frac{\partial}{\partial x_n}\right)^{\alpha_n}(f * g)$$
$$=\left[\left(\frac{\partial}{\partial x_1}\right)^{\alpha_1}\left(\frac{\partial}{\partial x_2}\right)^{\alpha_2}\cdots\left(\frac{\partial}{\partial x_n}\right)^{\alpha_n} f\right] * g$$
$$=f * \left[\left(\frac{\partial}{\partial x_1}\right)^{\alpha_1}\left(\frac{\partial}{\partial x_2}\right)^{\alpha_2}\cdots\left(\frac{\partial}{\partial x_n}\right)^{\alpha_n} g\right]$$

$$(102)\qquad \frac{1}{r}=\frac{1}{p}+\frac{1}{q}-1$$

$$(103)\qquad \|h\|_r \leqslant \|f\|_p\|g\|_q$$

$$(104)\qquad \int_{R^n} k(x)dx=1$$

$$(105)\qquad k_\varepsilon(x)=\varepsilon^{-n}k\left(\frac{x}{\varepsilon}\right)=\varepsilon^{-n}k\left(\frac{x_1}{\varepsilon},\frac{x_2}{\varepsilon},\cdots,\frac{x_n}{\varepsilon}\right)$$

$$(106)\qquad \int_{R^n} k_\varepsilon(x)dx=\int_{R^n} k(x)dx=1$$

$$(107)\qquad \frac{\partial^2 u}{\partial x_1^2}+\frac{\partial^2 u}{\partial x_2^2}+\cdots+\frac{\partial^2 u}{\partial x_n^2}=-f$$

differential equations; ANALYSIS, FUNCTIONAL). (2) The Hankel transform of a function f defined in the interval $(0, \infty)$ is given by $H_\alpha f(x)=\int_0^\infty J_\alpha(xy)\cdot(xy)^{1/2}f(y)dy$, in which J denotes the Bessel function defined in the last section. This transform occurs in the study of those functions and equations in Euclidean space that depend only on the distance from the origin. (3) The Mellin transform of a function on $(0, \infty)$ is defined as $Mf(x)=\int_0^\infty x^{it}f(t)dt$ for all real values of x. It will be seen below that the Mellin transform arises from the multiplicative structure of the real line in the same way that the Fourier transform comes from the additive structure.

Convolution and the Fourier transform. The notion of convolution is now extended to functions of n real variables. If $x=(x_1, x_2, \cdots, x_n)$ is an n-tuple of real numbers and $f(x)$, $g(x)$ are functions of x, then $f * g$ as a function of x is an integral over the n-dimensional Euclidean space of a product of f and g, arguments $x-y$ and y appearing in the functions of the integrand (see 97), in which $dy=dy_1, dy_2\cdots dy_n$; such an expression is called the convolution of f and g. When $n=1$ this reduces to the previous definition. By making a change of variable $z=x-y$, $dz=dy$, this integral may be written in the form of an integral over n-dimensional Euclidean space of the product of f and g, the variables z and $x-z$ serving as arguments, respectively (see 98), so that $f * g = g * f$.

It is easily shown that a linear combination of functions f_1 and f_2 convoluted with g equals a linear combination of convolutions (see 99). The connection between convolutions and Fourier transforms lies deeper. If f and g both have Fourier transforms $\hat{f}$ and $\hat{g}$, respectively, then a simple formal argument shows that a transform of the convolution is the product of the transforms (see 100), so that in terms of the Fourier transform, convolution corresponds to multiplication. In view of the analogy between Fourier transform and inverse Fourier transform, the result could also be interpreted as follows: to multiplication of functions corresponds the convolution of their Fourier transforms. And so, convolution actually is a kind of multiplication. Clearly, the properties of $f * g$ depend on those of both f and g. The theory of convolutions shows that the "good" properties of both f and g (integrability over R^n, boundedness, continuity, differentiability, etc.) are inherited by their convolution. For instance, the formula in which a general differential operator composed of powers of partial derivatives operating upon a convolution is equal to a convolution in which one function remains unmodified and the other is modified with the equivalent differential operation (see 101) holds for differentiation. Now if f belongs to L^p, in which $p>1$, and g belongs to L^q, then it can be shown that $h=f * g$ belongs to L^r, in which r is defined by a relationship in terms of equality between reciprocals r, p, and q (see 102; provided r is positive), and hence r is greater than q, and the inequality in which the rth norm of h is less than or equal to the product of the pth norm of f and the qth norm of g (see 103) holds.

"Good" functions

Convolutions are useful in giving approximations to general functions by good functions. This approximation is based on the following idea: For any kernel $k(x)=k(x_1, x_2, \cdots, x_n)$ of, say, a total integral over n-dimensional Euclidean space of k equal to 1 (see 104), and given any positive number ε, a new kernel $k_\epsilon(x)$ can be formed by the equation in which k_ϵ as a function of x is a product of the negative nth power of ε and k as a function of n arguments, the typical one being x_k divided by ε (see 105). For any f, the convolution of f with k_ϵ can also be defined. The graph of $k_\epsilon(x)$ is that of $k(x)$, contracted horizontally by the ratio ε, and expanded vertically by the ratio ε^{-n}, so that the integral over n-dimensional Euclidean space of k_ϵ is 1 (see 106) for all ε. It is assumed for the moment that $k(x)=0$ outside the sphere of radius 1 centred at the origin. Then $k_\epsilon(x)=0$ outside the sphere of radius ε centred at the origin. On the other hand, if f happens to be continuous at the point x_0, then $f(x)$ is practically constant on a ball of radius ε centred at x_0, and so it is not difficult to see that $f_\epsilon(x_0)$ tends to $f(x_0)$ as ε tends to zero. For this reason, $f_\epsilon=k_\epsilon * f$ is called an approximate identity. The functions f_ϵ provide good approximations to f under very general circumstances. For example, if k is absolutely integrable and f is in L^p, then likewise f_ϵ is in L^p, and $\|f-f_\epsilon\|_p$ tends to zero as ε tends to zero; if f is uniformly continuous, then f_ϵ tends uniformly to f as ε tends to zero. It may be noted that if k is a good function —say α times differentiable—then k_ϵ is likewise good, and so f_ϵ is good. Thus the convolution $k_\epsilon * f$ gives an approximation to f by good functions.

Applications to partial differential equations. The applications of convolutions and Fourier transforms will now be illustrated with an example from partial differential equations and potential theory. If a distribution of electrical charges throughout Euclidean space R^n $(n \geqslant 3)$ has a density of charge at the point x given by a function $f(x)$, it can be shown that the electrical potential $u(x)$ arising from the charge distribution f satisfies an expression in which the sum of second-order partial derivatives of u taken with respect to each of the n arguments is equal to the negative value of f (see 107).

Traditionally, the left-hand side (see 107) is called the Laplacian of u and denoted by Δu or $\nabla^2 u$. The equation $\Delta u=0$ is called Laplace's equation after the 18th–19th-century French mathematician and physicist Pierre-Simon Laplace, and its solutions are called harmonic functions. The inhomogeneous Laplace equation (107) is sometimes called Poisson's equation after another 18th–19th-century Frenchman, Siméon-Denis Poisson. Now given any f (subject to some technical conditions that are not specified) the equation (107) can be solved by using the Fourier transform. To do so, the effect of the Fourier transform on Δu must be examined. From two consecutive applications of an earlier formula (84), it can be seen that the transform of the second-

The inhomogeneous Laplace equation

order partial derivative of u with respect to one variable all considered as a function of x is a product of the negative square of the variable times the transform of f (see 108), and therefore the transform of Δu as a function of x is the negative of the squared absolute value of x times the transform of u (see 109). Consequently, the Fourier transform changes (107) to the equation in which the absolute square of x times the transform of u as a function of x is equal to the transform of f as a function of x (see 110), which has the solution that the transform of u as a function of x is equal to the square of the reciprocal of the absolute value of x multiplied by the transform of f considered as a function of x (see 111). In principle, this formula (111) solves the problem completely. For, given any well-behaved function f, the function f may be computed from its definition (see 82), $\hat{u}$ found from the above formula (111), and then passage from $\hat{u}$ back to u made by means of the Fourier inversion formula. This procedure, however, is cumbersome in practice, and in order to solve the problem (107) more explicitly, the solution (111) will be combined with a formula involving convolution. If k as a function of x is equal to a constant times the $2-n$ power of the absolute value of x (see 112), in which c is a constant the exact value of which is not of concern here, $k(x)$ is called the Newtonian potential. Some calculations show that the transform of k as a function of x is the -2 power of the absolute value of x (see 113; see above *Examples of Fourier transforms*). It can be seen that the transform of the convolution of k and f is by calculation the transform of u (see 114) by virtue of the formula (111) for $\hat{u}$. Consequently, $u = k\ f$; that is, u as a function of x is proportional to an integral over n-dimensional Euclidean space in which the integrand includes as a factor f and as a factor the reciprocal of the absolute difference between x and y raised to the $n-2$ power (see 115).

Thus an explicit solution has just been found to the inhomogeneous Laplace equation.

Certain formulas (see 111 and 115) form the starting point for many further investigations in harmonic analysis. Some formulas (see 111, 115) may be generalized to define an interesting family of integral transforms, the fractional integrals. If α is any positive number, the transform of $I^\alpha f$ as a function of x is a product, one term of which is the negative α power of the absolute value of x, one term of which is the transform of f, considered as a function of x (see 116), or, what is in fact the same thing, the product $I^\alpha f$ considered as a function of x is equal to a constant times an integral over n-dimensional Euclidean space in which the integrand contains as a factor f and as a factor the reciprocal of the $n-\alpha$ power of the absolute difference between x and y (see 117). It follows that $(\widehat{I^\alpha f})$ tends to zero at infinity more rapidly than f. Because rapid decrease of a Fourier transform $\hat{g}$ corresponds (more or less) to smoothness of the function g, it can be seen that I^α transforms rough functions f to smooth functions $I^\alpha f$. It may be said that a function g on R^n belongs to the class $L_\alpha^p(R^n)$ if g arises by applying I^α to an L^p function f: $I^\alpha f = g$. The classes $L_\alpha^p(R^n)$, called Sobolev spaces, are used by analysists as a powerful tool to measure the smoothness of functions, and they occur often in harmonic analysis and partial differential equations. In the special case in which $n = 1$, $1 < p < \infty$ and $\alpha =$ integer k, a function f belongs to L_α^p if, and only if, its kth derivative $f^{(k)}$ is in L^p. When α is not an integer, the Sobolev spaces give meaning to such statements as "the function f is 3½ times differentiable."

Sobolev spaces

The explicit solution to the inhomogeneous Laplace equation (see 115) may be considered from the point of view of smoothness. Because u was obtained from a formal calculation as the solution to a second-order partial differential equation (see 107), it must be established that u actually has second derivatives; for only then will Δu be meaningful. Formally, it is possible to differentiate (115) twice under the integral sign to obtain the second partial derivative of u with respect to one of the variables equal to an integral over n-dimensional Euclidean space of a product, one factor being a ratio of Ω_j to

(108) $$\widehat{\frac{\partial^2 u}{\partial x_j^2}}(x) = -x_j^2 \hat{f}(x) \quad \text{for } j = 1, 2, \cdots, n$$

(109) $$\widehat{\Delta u}(x) = -(x_1^2 + x_2^2 + \cdots + x_n^2)\hat{u}(x) = -|x|^2 \hat{u}(x)$$

(110) $$|x|^2 \hat{u}(x) = \hat{f}(x)$$

(111) $$\hat{u}(x) = |x|^{-2} \hat{f}(x)$$

(112) $$k(x) = c|x|^{2-n}$$

(113) $$\hat{k}(x) = |x|^{-2}$$

(114) $$\widehat{k * f}(x) = |x|^{-2} \hat{f}(x) = \hat{u}(x)$$

(115) $$u(x) = c \int_{R^n} \frac{f(y)}{|x-y|^{n-2}}\, dy$$

(116) $$\widehat{I^\alpha f}(x) = |x|^{-\alpha} \hat{f}(x) \quad \text{for all } x \text{ in } R^n$$

(117) $$I^\alpha f(x) = c_\alpha \int_{R^n} \frac{f(y)}{|x-y|^{n-\alpha}}\, dy \quad \text{for all } x$$

(118) $$\frac{\partial^2}{\partial x_j^2} u(x) = \int_{R^n} \frac{\Omega_j(y)}{|y|^n} f(x-y)\, dy$$

(119) $$\Omega_j(y) = c(1 - n|y|^{-2} y_j^2)$$

(120) $$Tf(x) = \int_{R^n} \frac{\Omega(y)}{|y|^n} f(x-y)\, dy$$

(121) $$\begin{cases} \Omega(\lambda y_1, \lambda y_2, \cdots, \lambda y_n) = \Omega(y_1, y_2, \cdots, y_n) \\ \text{for all } (y_1, y_2, \cdots, y_n) \text{ in } R^n. \end{cases}$$

(122) $$T_\varepsilon f(x) = \int_{|y|>\varepsilon} \frac{\Omega(y)}{|y|^n} f(x-y)\, dy$$

the nth absolute power of y, one factor being f in which the argument is the difference between x and y (see 118), and also Ω_j as a function of y equal to a constant times a difference involving n, y, and y_j (see 119). The integral (118), however, is not absolutely convergent. Roughly speaking, showing that u in the explicit solution (115) is an honest solution to the differential equation (107) is equivalent to the problem of defining what is meant by a certain integral (see 118) and investigating its properties. This leads at once to consideration of singular integral operators; *i.e.*, integral transforms T of the form in which the transform of f is equal to an integral over n-dimensional Euclidean space in which two factors appear in the integrand, one a ratio of Ω to the nth absolute power of y, one factor the function f with argument $x - y$ (see 120). Here Ω is regarded as a fixed function and is assumed to share with Ω_j the following essential properties: (1) $\Omega(y)$ depends only on the direction of y and not on its magnitude. That is, the function Ω with λ placed as a factor in each argument is equal to the function Ω for all n arguments as components of a vector in n-dimensional Euclidean space (see 121) and any positive number λ. (2) $\Omega(y)$ has average value zero over the unit sphere $|y| = 1$ in R^n.

Singular integral operators

Again, the problem arises that a certain integral (see 120) is not absolutely convergent. The theory of singular integrals, however, shows that the T_ε operating upon f is an integral over a region in which the absolute value of y is greater than ε, factors in the integral being the ratio of Ω to the nth absolute value of y as well as the factor composed of f with argument $x - y$ (see 122), tending to a limit at almost every point x in R^n, as ε tends to zero. This limit may be denoted by $Tf(x)$, and in

fact T is a bounded operator on all the L^p classes in which $1 < p < \infty$. These results give a powerful hold on the smoothness of u (see 115) and more generally yield powerful insights into the important class of elliptic partial differential equations (see DIFFERENTIAL EQUATIONS: *Partial differential equations*).

Mention should be made of a particular singular integral operator of fundamental importance in harmonic analysis and the theory of analytic functions. That operator is the Hilbert transform T defined on functions of a single variable such that T operating upon f is an integral over the entire real line with integrand composed of a ratio of f to y, the argument in f being $x - y$ (see 123). This transform is of a specific form (see 120) with $\Omega(y) = +1$ if y is positive and $\Omega(y) = -1$ if y is negative. It is not possible to present an adequate discussion of the Hilbert transform here because it is beyond the scope of this article; at least the two main properties of this basic operator, however, are mentioned.

(1) If $f(x)$ is a function of period 2π, defined on $(-\infty, \infty)$, and $S_N f(x)$ denotes the Nth partial sum of the Fourier series of f: S_N operating upon f is a sum extending over integers from $-N$ to $+N$ of constants times exponentials (see 124), in which f as a function of x is an infinite sum extending from minus to plus infinity of constants times exponentials (see 125), and then S_N operating upon f is a difference of products, factors and products being complex exponentials and transforms T operating upon products of exponentials and f (see 126; the Dirichlet formula), in which T denotes the Hilbert transform, named for the German mathematician David Hilbert.

(2) If $f(x)$ is a real-valued function of one variable, then the complex-valued function $f(x) + iTf(x)$, defined initially for real values of x, extends to an analytic function of $z = x + iy$ defined in the half-plane $y \geqslant 0$.

Via (1) and (2), the Hilbert transform provides a link between harmonic analysis and the theory of analytic functions.

In addition to the inhomogeneous Laplace equation, there are a number of other partial differential equations arising from mathematical analysis and physics that can be solved by methods that are similar to those outlined above (see further DIFFERENTIAL EQUATIONS: *Partial differential equations*).

Generalizations of the Fourier transform. Much of the highly developed theory of Fourier integrals, which has just been sketched in the setting of Euclidean space, can actually be carried out in a wider context. The theory of various generalizations of the Fourier transform is known as abstract harmonic analysis. In addition to providing new mathematical results, abstract harmonic analysis sharpens an understanding of classical Fourier analysis by elucidating the essential unity that exists among the many classical integral transforms: Fourier integrals, Fourier series, Mellin transforms, and expansions in spherical harmonics, to cite a few examples. The present exposition of abstract harmonic analysis will concentrate on the part of the subject that is easiest to understand, namely the study of Fourier transforms on Abelian groups (see below), and attention will now be turned to this topic. It is best to begin by viewing the classical theory from a proper perspective.

Abstract harmonic analysis

From its very definition (see 82) it can be seen that the Fourier transform on R^n is built up from the functions of such a form as χ as a function of y equal to a complex exponential containing a vector product of x and y in the product (see 127). (Here x is thought of as a fixed point in R^n and $\chi_x(y)$ regarded as a function of y.) Indeed, the Fourier transform (82) may be rewritten in a form such that the transform of f as a function of x is an integral over n-dimensional Euclidean space of a product of χ_x and f (see 128). The essential property of χ_x that is used in harmonic analysis is the multiplicative law of the exponential: χ_x operating upon the sum of y_1 and y_2 is a product of terms, χ operating upon y_1 and y_2, respectively (see 129). Now if G is any set on which there is defined some notion of the sum $y_1 + y_2$ of two elements y_1 and y_2 of G, and also some notion of integral, then the key formulas (128) and (129) are valid for complex-valued functions χ_x and f defined on G, and thus harmonic analysis can be carried out on G. More precisely, the following definitions can be made:

Abelian groups

An Abelian group is a set G, together with an operation $+$, that combines any two elements x and y of G into a third element $x + y$ of G, and that satisfies axioms 1–4 below.

1. $(x + y) + z = x + (y + z)$ for any three elements x, y, and z in G (the associative law; see 130).
2. $x + y = y + x$ for any two elements x and y in G (the commutative law).
3. There is an element called 0 in G, with the property that $x + 0 = x$ for every element x of G (existence of an identity element).
4. Corresponding to any given element x in G, there is an element called $(-x)$ in G, which satisfies the equation $x + (-x) = 0$ (existence of an inverse element.).

A locally compact Abelian group is an Abelian group in which, very roughly, statements such as "x is close to y" or "f is a continuous function," or various similar remarks involving the limit concept can be made. (The precise definition is omitted, because it is technical and requires some background in general topology. All the Abelian groups considered in this article, however, will be locally compact.) If G is a locally compact Abelian group, a character χ of G is a continuous complex-valued function defined on G, which has the properties: (a) $\chi(y)$ has modulus one for every element y of G; (b) χ operating upon $y_1 + y_2$ is equal to the product of χ operating upon y_1 and χ operating upon y_2 (see 131) for every two elements y_1 and y_2 of G.

A significant portion of harmonic analysis can be carried out in the setting of locally compact Abelian groups. Three classical examples that will be used later are now considered. (1) $G = R^n$ is an Abelian group under the sum between two n component vectors that is a vector of n components, a typical one being $x_k + y_k$ (see 132), with identity element $(0, 0, \cdots, 0)$; the inverse of the vector composed of n components is the vector composed of n negatives of the same components (see 133). The char-

$$(123) \qquad Tf(x) = c\int_{-\infty}^{\infty} \frac{f(x-y)}{y}\,dy$$

$$(124) \qquad S_N f(x) = \sum_{n=-N}^{N} c_n e^{inx}$$

$$(125) \qquad f(x) \sim \sum_{-\infty}^{\infty} c_n e^{inx}$$

$$(126) \qquad S_N f = e^{iNx}\,T(e^{-iNy} f(y)) - e^{-iNx}\,T(e^{iNy} f(y))$$

$$(127) \qquad \chi_x(y) = e^{ix \cdot y}$$

$$(128) \qquad \hat{f}(x) = (2\pi)^{-\frac{n}{2}} \int_{R^n} \chi_x(y) f(y)\,dy$$

$$(129) \qquad \chi_x(y_1 + y_2) = \chi_x(y_1) \cdot \chi_x(y_2)$$

$$(130) \qquad (x + y) + z = x + (y + z) \quad \text{for any three elements } x, y, \text{ and } z \text{ in } G$$

$$(131) \qquad \chi(y_1 + y_2) = \chi(y_1) \cdot \chi(y_2)$$

$$(132) \qquad (x_1, x_2, \cdots, x_n) + (y_1, y_2, \cdots, y_n) = (x_1 + y_1, x_2 + y_2, \cdots, x_n + y_n)$$

$$(133) \qquad (x_1, x_2, \cdots, x_n) \text{ is } (-x_1, -x_2, \cdots, -x_n)$$

acters of R^n are seen to be precisely the functions $\chi_x(y) = \exp(ix \cdot y)$. Thus, characters play a basic role in harmonic analysis on R^n. (2) $G = R^+$, the set of all positive real numbers is not an Abelian group under ordinary addition. (For instance, axiom 3 is violated, for the only number a with the property that $x + a = x$ for all x in R^+ is zero, and zero does not belong to R^+.) If the symbol $+$ is allowed to stand for multiplication of real numbers, however, then R^+ becomes a locally compact Abelian group. (For instance the identity element demanded by axiom 3 is one, and the inverse of a given element x of R^+ is $1/x$.) The characters of R^+ are functions $\chi_t(x) = x^{it}$, in which t is any fixed real number. It should be possible to do harmonic analysis on R^+ in terms of the functions $\chi_t(x)$, and as will be seen below, this is the case. (3) The circle, denoted by S^1, may be made into a group by regarding it as the set of all complex numbers z of modulus $|z| = 1$, with the operation $+$ as ordinary multiplication of complex numbers. Then S^1 becomes a locally compact Abelian group. Via the correspondence $z = e^{i\theta}$, the circle S^1 may also be thought of as consisting of all angles θ in the interval $[0, 2\pi]$. The characters on S^1 are precisely the functions in which χ_n as a function of θ is a complex exponential with the product $n\theta$ in the argument, n being any integer (see 134). According to the classical theory of Fourier series (see above *Fourier series*), any integrable function $f(\theta)$ defined on $[0, 2\pi]$ may be expanded in an infinite series of a form such that f as a function of θ is an infinite sum with indices extending from minus infinity to plus infinity and summands having factors a_n and either a factor of exponential form or of the form χ_n, arguments being θ (see 135), in which a_n being a transform of n is an integral from zero to 2π of the product of f with either a negative exponential or the complex conjugate of χ (see 136). Here again, characters appear as the fundamentals of harmonic analysis on the group $G = S^1$.

Returning to the general theory of Abelian groups: If χ_1 and χ_2 are two characters of a locally compact Abelian group G, a new character, denoted $\chi_1 + \chi_2$, may be defined by the formula $(\chi_1 + \chi_2)(y) = \chi_1(y) \cdot \chi_2(y)$ for every element y in G. Under this "addition" operation, the set $\hat{G}$ of all the characters of G becomes a locally compact Abelian group, called the dual group or character group of G with $\chi_0(x)$ identically equal to 1 as the identity element, and $(-\chi)(y) = 1/\chi(y)$ as the inverse to a given character χ. For the three classical examples above, the dual groups G are as follows: R^n is its own dual group, R^1 is the dual group of R^+, and the set Z of integers $0, \pm 1, \pm 2, \cdots$ with the operation of addition is the dual group of S^1. In each case, the Fourier transform f of a function f on G is a function on the dual group G.

In order to define the Fourier transform and study harmonic analysis on a generally locally compact Abelian group G, it is desirable to have a notion of the integral of a real- or complex-valued function defined on G. Initially only continuous functions f that are equal to zero outside of some bounded—*i.e.*, compact—subset of G will be considered. The class of all such functions is denoted by $C_c(G)$. In the known case, $G = R^n$, $I(f)$ equal to an integral over n-dimensional Euclidean space of f (see 137), defined for functions in $C_c(G)$, has the following key properties: (a) I as a function of a linear combination of f and g is a linear combination of I as a function of f and I as a function of g (see 138) for any functions f and g in $C_c(R^n)$ and any two numbers a and b. (b) If $f(x) \geqslant 0$ for all x, then $I(f) \geqslant 0$; in fact, $I(f) > 0$ unless $f(x)$ is identically zero. (c) For a given function f and an element x_0 of G, the translate f_{x_0} of f by x_0 is defined by the equation in which f_{x_0} as a function of x is equal to f as a function of the sum x and x_0 (see 139). (Thus on R^n, the graph of f_{x_0} is obtained from that of f simply by sliding horizontally.) Then for arbitrary f and x_0, the equation is $I(f_{x_0}) = I(f)$ (translation-invariance of the integral).

Moreover, it is not hard to prove that the only way of associating a number $I(f)$ to each f in $C_c(R^n)$ in such a way that (a), (b), and (c) are satisfied is to set $I(f)$ equal to a constant times an integral over n-dimensional Euclidean space of f (see 140) for some positive constant k. Thus, except for an insignificant constant factor, properties (a), (b), and (c) uniquely characterize the Riemann integral on R^n. In the general case of a locally compact Abelian group G, a method of associating a number $I(f)$ to each function f in $C_c(G)$ is an integral if properties (a), (b), and (c) hold. It can be shown that on any locally compact Abelian group G, there is one and (essentially) only one integral on G; it is called the Haar integral on G, and $I(f)$ is written $\int_G f$, by analogy with the case $G = R^n$. Once the Haar integral is defined for functions in $C_c(G)$, it may be extended to a large class of (possibly discontinuous) functions on G, just as the Riemann integral on R^n may be extended to the Lebesgue integral. (See ANALYSIS, REAL: *Integration and measure.*) Thus, in particular, L^p classes are definable on any locally compact Abelian group.

The Haar integral

Haar measure can be illustrated by considering the same three classical examples. On $G = R^n$, the Haar integral is just $I(f)$ equal to an integral over n-dimensional Euclidean space of the function f (see 141), as was seen above. On $G = S^1 = [0, 2\pi]$, the Haar integral is given by $I(f)$ as an integral from zero to 2π of f (see 142). Finally, on $G = R^+$, the Haar integral is I as a function of f equal to an integral over the positive real line of f divided by its argument (see 143).

The basic notions of the character group and Haar integral having been defined, the basic results of harmonic analysis on a locally compact Abelian group G can be stated. For any absolutely integrable function f on G, the Fourier transform f of f is defined to be the following function on the character group $\hat{G}$: the transform of f as a function of χ is a constant times the integral over G of the product of χ and f (see 144). If f is a square-integrable function on G, then the integral (144) need not converge absolutely, but can be interpreted by a procedure analogous to that used above to define (82) in the case $G = R^n$. Moreover, several of the major results of harmonic analysis on R^n are also valid on G. For ex-

(134) $$\chi_n(\theta) = e^{in\theta}, \quad \text{in which } n = 0, \pm 1, \pm 2, \cdots.$$

(135) $$f(\theta) \sim \sum_{n=-\infty}^{\infty} a_n e^{in\theta} = \sum_{-\infty}^{\infty} a_n \chi_n(\theta)$$

(136) $$a_n = \hat{f}(n) = (2\pi)^{-1} \int_0^{2\pi} e^{-in\theta} f(\theta)\, d(\theta) = (2\pi)^{-1} \int_0^{2\pi} \overline{\chi_n(\theta)} f(\theta)\, d\theta$$

(137) $$I(f) = \int_{R^n} f(x)\, dx$$

(138) $$I(af + bg) = aI(f) + bI(g)$$

(139) $$f_{x_0}(x) = f(x + x_0)$$

(140) $$I(f) = k \int_{R^n} f(x)\, dx$$

(141) $$I(f) = \int_{R^n} f(x)\, dx$$

(142) $$I(f) = \int_0^{2\pi} f(\theta)\, d\theta$$

(143) $$I(f) = \int_0^{\infty} f(t) \frac{dt}{t}$$

(144) $$\hat{f}(\chi) = c \int_G \chi(y) f(y) \quad \text{for all } \chi \text{ in } \hat{G}$$

$$\int_G |f(x)|^2 = \int_{\hat{G}} |\hat{f}(\chi)|^2 \tag{145}$$

$$f(x) = c\int_{\hat{G}} \overline{\chi(x)}\hat{f}(\chi) \tag{146}$$

$$\hat{f}(t) = \int_0^\infty x^{it} f(x)\frac{dx}{x} \tag{147}$$

$$\text{for } R^+, \quad f(x) = c\int_{-\infty}^{\infty} x^{it}\hat{f}(t)\,dt \tag{148}$$

$$\hat{f}(n) = (2\pi)^{-1}\int_0^{2\pi} e^{inx} f(x)\,dx \tag{149}$$

$$f(x) = \sum_{n=-\infty}^{\infty} \hat{f}(n)e^{-inx} \tag{150}$$

$$(2\pi)^{-1}\int_0^{2\pi} |f(x)|^2\,dx = \sum_{n=-\infty}^{\infty} |\hat{f}(n)|^2 \tag{151}$$

ample, there is the Plancherel formula: two integrals are equal, one extended over G, one over $\hat{G}$, the integrand of the first being the absolute square of f, the integrand of the second being the absolute square of f (see 145), and there is the Fourier inversion formula: f as a function of x is a constant times the integral over $\hat{G}$ of a product of terms, one being the complex conjugate of χ, the other being $\hat{f}$ (see 146); both hold for suitable functions f on G. (Integration over $\hat{G}$ in formulas [145] and [146] makes sense, because $\hat{G}$ is a locally compact Abelian group, and thus has its own Haar integral. The bar over $\chi(x)$ [see 146] denotes complex conjugation.) Convolutions and approximate identities can be defined on G, and their relationship to the Fourier transform is much the same as in the Euclidean case $G = R^n$. On a general locally compact Abelian group, however, there is no notion of differentiation, so that the rich interplay between harmonic analysis and partial differential equations in the Euclidean case is lost.

Fourier transform and inversion

The meaning of the Fourier transform (144) and inverse Fourier transform (146) may be illustrated by the three classical examples $G = R^n$, $G = R^+$, and $G = [0, 2\pi]$. On R^n, the Fourier transform and inverse Fourier transform are precisely the usual expressions (see 82 and 83). The Fourier transform (144) on R^+ is given by an expression in which f is an integral extended over the positive real axis with integrand composed of three terms, one of which is x to a complex exponent, one of which is f, one of which is the reciprocal of x (see 147), and thus is none other than the Mellin transform defined earlier. The inversion formula (146) becomes an expression in which f as a function of x is a constant times an integral extended over the entire real line, one factor in the integrand being x raised to a complex exponent, the other factor being f (see 148; the Mellin inversion formula). Finally, on $G = [0, 2\pi]$, the Fourier transform becomes an expression in which f as a function of n is proportional to the integral from zero to 2π of the product of an exponential times f, nx being the argument in the exponential (see 149). In other words, the Fourier transform on G simply associates to every function f on $[0, 2\pi]$ its sequence of Fourier coefficients. The inversion formula (146) asserts that f as a function of x is a sum extending over integers from minus infinity to plus infinity with summand composed of the product f times an exponential, nx appearing in the exponential (see 150), and the Plancherel formula (see 145) specializes to Parseval's relation, the expression in which an integral and a sum are equal, the integral extending from zero to 2π and applying to the integrand absolute square of f, the sum extending over integers from minus infinity to plus infinity with summand being the absolute square of f (see 151). Thus the basic elementary results in the classical theories of Fourier series and integrals are part of the larger whole of harmonic analysis on locally compact Abelian groups.

Finally, it is noted that highly significant results are known about Fourier analysis on non-Abelian groups; *i.e.*, roughly, mathematical structures satisfying axioms 1, 3, and 4 above but not axiom 2.

REPRESENTATIONS OF GROUPS AND ALGEBRAS

Since their formal introduction in the early 19th century, groups have been one of the principal objects of mathematical attention (see ALGEBRAIC STRUCTURES). Their widespread and profound applications to such physical subjects as crystallography, quantum mechanics, and hydrodynamics and to such other mathematical regimes as number theory, harmonic analysis, and geometry have demonstrated their importance.

The main general technique for studying groups is the method of group representation. Technically, a representation of a group is a homomorphism of it into another group, most commonly, into the group of invertible linear transformations (or matrices) on some linear space. Less technically, this method amounts to comparing the given group with better known examples of groups. Representation theory plays a significant part in all branches of mathematics in various ways; it also is important in theoretical physics, particularly in quantum theory.

Historical development. The general representation theory for finite groups was developed by the mathematicians Ferdinand Georg Frobenius (German), Issai Schur (German), and William Burnside (British) at the end of the 19th and the beginning of the 20th century. Nothing of the deeper theory of finite groups developed by such later workers as Émile Artin (French) and Richard Brauer (German-born U.S.) has been discussed here. The extension of the finite group theory to compact groups is due in part to Schur, who developed the representation theory of the rotation group in those terms. The general extension to compact groups of the completeness and orthogonality of the finite-dimensional irreducible unitary representations is known as the Peter-Weyl theorem (published in 1927). Lie groups are named after Lie, who discovered their fundamental properties in the last decades of the 19th century. A deeper analysis of their properties and, in particular, the classification of the semisimple groups were carried out by Élie-Joseph Cartan (French) in the early part of the 20th century. The general finite-dimensional representation theory for semisimple Lie groups was developed principally by Weyl in the mid-1920s as were the connections between group theory and quantum mechanics. The general representation theory for topological groups started with the discovery by Alfréd Haar (Hungarian) in the early 1930s of an invariant measure on locally compact groups. The existence of a separating family of irreducible unitary representations for locally compact groups was established by the Soviet mathematicians Izrail M. Gelfand and Dmitry A. Raykov in the early 1940s. The decomposition of unitary representations of locally compact groups into irreducible and factor unitary representations was carried out by Friederich I. Mautner (U.S.) at the end of the 1940s. This decomposition was based on the powerful theory of operator algebras developed by John von Neumann (U.S.) in collaboration with F.J. Murray (U.S.) in the 1930s and early 1940s. In particular, the theory of factors is their creation. A general duality and function transform theory based on the Murray–von Neumann operator algebras was developed by Irving E. Segal (U.S.) in the early 1950s. George W. Mackey (U.S.) created a forceful technique for the analysis of representations of groups with type I representations by generalizing the Frobenius theory of induced representations. The detailed analysis of the infinite-dimensional representations of the semisimple Lie groups is the work of Gelfand, Mark Aronovich Naymark (Soviet), and Harish-Chandra. The fact that these groups have type I representations is a consequence of Harish-Chandra's results. Jacques Dixmier (French) established this same fact for

the so-called nilpotent Lie groups. Examples of Lie groups with representations not of type I are known.

Groups. To sketch briefly, a group is a set on which a multiplication satisfying certain specific properties is defined. Corresponding to each pair of elements a, b (in a given order) is another element (written ab), and this law of combination has the properties: (1) $a(bc) = (ab)c$; (2) there is an element e, called the group identity element, such that $ae = ea = a$; and (3) for each a there is an element a^{-1}, called the inverse of a such that $aa^{-1} = a^{-1}a = e$. If in addition $ab = ba$ for each a and b, the group is said to be Abelian (also commutative). A homomorphism is a mapping (function, correspondence) f of one group G into another group H such that $f(ab) = f(a)f(b)$ for each a and b in G; that is, f preserves multiplication. The positive and negative integers with their usual addition provide an example of an infinite Abelian group (here, addition plays the role of group multiplication, 0 is the group identity and, for example, -5 is the group inverse to 5). The two integers 0 and 1 with the usual addition as group multiplication (except that 1 is to combine with 1 to give 0) provide an example of a finite Abelian group, as does the set of two integers $+1$ and -1 with the usual multiplication of integers. The mapping of an integer into 0 if it is even and into 1 if it is odd provides an example of a homomorphism of the group of integers onto the first two-element group. The function which assigns 0 to $+1$ and 1 to -1 is an example of an isomorphism (*i.e.*, one-to-one homomorphism) of the second two-element group onto the first. Structurally, as far as group theoretical properties are concerned, the existence of this isomorphism indicates that both two-element groups are the same—they are different (isomorphic) representations of the same abstract group, so to speak.

Group representations. The process of group representation may be likened to measuring some physical object with a ruler. By comparing the object to be studied with the known object (the ruler), special information is obtained. Spreading the object out completely along the ruler might be thought of as an isomorphic representation, but such a procedure may not be possible. For example, a polyhedron would have to be studied by applying a series of partial measurements and combining them (corresponding to combining information from homomorphic representations). It would be hoped that sufficiently many measurements could be made to give detailed information—corresponding to the case of sufficiently many (a separating family of) representations so that each pair of distinct elements is carried into a pair of distinct elements by at least one of the representations. An instance of combining information is afforded by the simple fact that a group with a separating family of representations in Abelian groups is itself Abelian (ab and ba are identified by all such representations). It would help little to make measurements with an unmarked straightedge; in the same way, the representations would be expected to take place in groups with some discernible structure. On the other hand, if the ruler is too specialized, for example, circular, it is available for only a limited type of measurement; by analogy, the class of groups in which the representations take place should remain broad enough to have general application. The so-called linear groups, groups of invertible linear transformations on a linear vector space, are well suited to this task. Representations in such groups often are referred to as linear representations, or simply as representations if confusion is unlikely. In particular, representations by unitary operators on Hilbert spaces, called unitary representations, and representations by operators on finite-dimensional linear spaces, called finite-dimensional representations, have proved very useful.

Group algebra

In applying the technique of representations to the study of groups, the allied method of associating an algebra, the so-called group algebra, with the group has a key function. (An algebra is a linear space with a multiplication satisfying certain specific conditions.) For finite groups and the complex group algebra, this amounts to associating with the group sums of formal complex multiples of the group elements and multiplying two such sums by distributing products in the usual way, commuting numbers past group elements, and multiplying group elements by means of the group multiplication.

A linear representation of the group can be extended to the group algebra by assigning to each sum of multiples of group elements the same sum of multiples of the operators (matrices) corresponding to those group elements under the given representation. This extension is a representation of the group algebra, a mapping into the algebra of operators on the representation space that preserves both products and sums. The analysis of the algebra representation is somewhat simpler than that of the group representation and yields much information about the group representation. The process of extension can be reversed, for the group elements are found in a natural way among those of the group algebra; a representation of the group gives rise to a representation of the group.

There are several ways of constructing new representations from one or more given representations. The first, restriction of a representation, starts with a representation f of a group G (or algebra) and a subspace V of the representation space that is mapped into itself by each of the representing operators, a so-called reducing or invariant subspace. The restricted representation assigns to each g in G the operator $f_V(g)$ on V that is $f(g)$ restricted (in its action) to V. If V is not 0 or the full space, the representation f is said to be reducible. If no such proper invariant subspace exists, the representation is said to be irreducible. (In the case of representations on Hilbert spaces, subspace is understood to mean closed subspace.) Representations of finite groups enjoy the important property of complete reducibility. If V is a reducing subspace, there is a complementary space V' that is also invariant; V and V' have only 0 in common and each vector in the representation space is the sum of one from V and one from V'.

The representation f is said to be the direct sum of f_V and $f_{V'}$. This description of direct sum from an internal viewpoint has an external counterpart. If f and f' are representations of G on spaces V and V', respectively, a representation $f \oplus f'$, called the direct sum of f and f', on the direct sum $V \oplus V'$ of V and V' (*i.e.*, the linear space of pairs of elements, one from V and one from V') is defined by assigning to each group element g the operator on $V \oplus V'$ that transforms a pair (v, v') into the pair $([f(g)](v), [f'(g)](v'))$. In the case of reducing complementary spaces V and V' for a representation f, it is an easy matter to see that $f_V \oplus f_{V'}$ is f when the full space is viewed as $V \oplus V'$. In a similar manner, direct sums of more than two representations can be defined. For finite groups, the process of decomposition of a representation may be continued on each of the restricted representations until a full reduction into irreducible representations is effected; the representation is a direct sum of irreducible representations and such a decomposition is unique.

The tensor product

A more complicated procedure for constructing representations from families of representations, the tensor product, stems from the tensor product of linear spaces (the space of sums of multiples of formal products of basis elements one from each space in a given order—the dimension is the product of the dimensions of the individual spaces as contrasted with that of the direct sum which is the sum of the dimensions). If f and f' are representations of G on spaces V and V', respectively, then $f \otimes f'$, their tensor product, assigns to each g in G the operator $f(g) \otimes f'(g)$ that transforms one of the generating elements $v \otimes v'$ of the tensor product $V \otimes V'$ of V and V' into $[f(g)](v) \otimes [f'(g)](v')$. In particular, the square of a representation consists of tensoring it with itself (similarly for higher powers).

Irreducible representations. A problem of some importance and difficulty is that of describing the irreducible representations that appear in the decomposition of the higher powers of irreducible representations of specific groups, because these are the basic representations not only of the general theory but also of the fundamental

physical situation to which the theory applies. Speaking of the possible occurrence of a representation that is reducible in the description of a general scheme for quantum mechanics, the 20th-century German-born U.S. mathematician Hermann Weyl, who was one of the founders of modern representation theory, wrote, "Nature could hardly be expected to indulge in such a superfluous luxury."

In dealing with representations of a given group, there is clearly no additional information to be obtained by passing from one representation to a second the space of which can be identified with that of the first in such a way that the operators corresponding to each group element under both representations act in the same way on the spaces (relative to the identification). In this case, the representations are said to be equivalent. (Technically, if f and f' are the equivalent representations of G on spaces V and V', there is a linear isomorphism P of V onto V' such that $Pf(g)P^{-1} = f'(g)$ for each g in G.) When a representation f of a group is expressed as the direct sum of irreducible representations and precisely n of these are equivalent to a given irreducible representation f', then f' is said to occur in f with multiplicity n. If the group G has m elements the set of complex-valued functions on G is a linear space of dimension m (addition and scalar multiplication performed pointwise). A natural representation of G on this space is available—to the function a is assigned the function a_g the value of which at g' is $a(g^{-1}g')$, and $f(g)$ is taken to be the transformation of the function space so defined. Then f is a representation of G, called the left regular representation of G. In a similar manner, there is the reflected situation of the right regular representation of G. The crucial property of the regular representations is that each contains every irreducible representation of G in its direct sum decomposition (with multiplicity equal to the dimension of the representation space—as a consequence of the reflection situation of the right and left regular representations). The regular representation is an isomorphic one, from which it is concluded that the group G has a separating family of irreducible representations. If f is a representation of the finite group G as operators on a Hilbert space, then the function that assigns to a pair of vectors x, y in the space the number $\Sigma_g(f(g)x, f(g)y)$ is again an inner product equivalent to the original one and relative to which each $f(g)$ is a unitary operator (orthogonal transformation, in the real case), so that each such representation of G is equivalent to a unitary representation of G. If f and f' are a pair of inequivalent (irreducible) unitary representations of G, then the sum over g of a product of inner products is zero, one of the inner products being composed of f', x', and y' (see 152), in which x, y and x', y' are orthonormal pairs of vectors in their respective spaces (the same holds when $f' = f$ and x', y', x, y is an orthonormal set such that not both $x = x'$ and $y = y'$); while $\Sigma_g \mid (f(g)x, y \mid^2$ is the number of elements in G, called the order of G, divided by the dimension of the representation. These are the so-called orthogonality relations. The sum of the diagonal entries of the matrix corresponding to an operator on and a basis for some finite-dimensional space is called the trace of the operator. It is independent of the basis chosen. If f is a representation of the finite group G on a finite-dimensional space, the function that assigns to each element of G the trace of its representing operator is called the character of f. A critical result of the theory states that two such representations are equivalent if, and only if, their characters are identical; the study of such representations is reduced to the study of the characters of G. The character of a direct sum of representations is the sum of their characters and that of a tensor product is their product. The character of the n-dimensional identity representation, which assigns the identity operator on n-dimensional space to each group element, has the constant function n as its character, and that of the regular representation is 0 at each group element except the identity in which it takes the value equal to the order of the group; this equality value is easily checked on the basis consisting of functions that are 1 at some group element and 0 at all others. The characters of the irreducible representations are precisely those characters k such that $\Sigma_g \mid k(g) \mid^2$ is the order of G.

Schur's lemma

Schur's lemma, which states that the only operators commuting with a unitary irreducible representation are the scalar multiples of the identity operator, is a key result in the development of representation theory. (Because each representation of a finite group is equivalent to a unitary representation, Schur's lemma covers all such representations.) Applying this fact to Abelian groups, it follows that each irreducible unitary representation of such a group is 1-dimensional; each group element is mapped into a complex multiple (having absolute value 1) of the identity operator. The function on the group that assigns to each element its corresponding complex number is the character of the representation. There is, of course, no need to distinguish between the representation and its character in this situation. The complex conjugate of such a character is again a character as is the product of two such characters. With this multiplication, the characters form a group—the character or dual group (to the original group). If G is an Abelian group and G' its dual, then each g in G gives rise to a character of G' that assigns to an element of G' its value at g. The duality theorem for Abelian groups states that the mapping of G just described is an isomorphism of G onto the character group of G'; roughly speaking, G is the dual group of its dual group. A function a on G may be viewed as the element $\Sigma_g a(g)g$ in the group algebra of G, and each character g' of G has the extension to this group algebra that assigns to $\Sigma a(g)g$ the number $\Sigma a(g)g'(g)$. Denoting this number by $a'(g')$, the transformation that maps a onto a' is an isomorphism of the group algebra of G onto the function algebra of G', the linear space of functions on G' with pointwise multiplication. For finite Abelian groups, this function mapping is the counterpart of the Fourier transform, which is so important in mathematical analysis (see above *Harmonic analysis and integral transforms*).

$$(152) \qquad \Sigma_g(f(g)x, y)(f'(g^{-1})x', y') = 0$$

The symmetric group on n letters, S_n, is the set of all one-one transformations of n objects (letters) onto itself (that is, the set of all permutations or arrangements of the n objects) with iteration of transformations (or arrangements) as the group multiplication. Thus S_2 can be viewed as the permutations of the numbers 1, 2: one permutation (the identity) leaving 1 and 2 fixed and the other interchanging 1 and 2 (the square of this last is the identity). It follows that S_2 has 2 elements; S_3 has $1 \cdot 2 \cdot 3$ $(= 6)$ elements, and quite generally, S_n has $1 \cdot 2 \cdot \cdots \cdot n$ $(= n!)$ elements. The natural combinatorial character of the symmetric groups results in their dominance of representation theory for finite groups—wherever physical quantities exhibit combinatorial symmetries (for example, in the quantum mechanics of a system of several identical electrons) the symmetric groups make their appearance. In the theory of groups, the importance of symmetric groups is obvious from the fact that each finite group is isomorphic with a subgroup of some symmetric group. In fact, if G has n elements, the mapping that assigns to each element g of G the permutation of the n elements of G effected by (left) multiplying each by g is an isomorphism of G with a subgroup of S_n. From the general theory, one sees that the study of the representations of S_n can be carried out by a detailed analysis of its group algebra and, in particular, of the regular representation.

Symmetric groups

For arbitrary finite groups, the number of inequivalent irreducible representations is the linear dimension of the centre of the group algebra (those elements of the group algebra that commute with all others); and this is seen to be the number of distinct conjugate classes of group elements. (Two group elements a and b are con-

jugate if there is some group element g such that $a = gbg^{-1}$. A conjugate class in a group is the set of conjugates of some fixed element.) For symmetric groups each conjugate class can be associated with a certain symmetry pattern. Each permutation on n letters permutes various disjoint subsets of the letters cyclically among themselves; *i.e.*, placing the letters of one subset along the rim of a wheel at equal intervals and in the proper order, the permutation affects these letters as would a rotation of the wheel carrying one into the next. Two permutations are conjugate precisely when the subsets of one can be matched with those of the other so that matching subsets have the same number of elements. Thus there are as many inequivalent irreducible representations of S_n as there are ways of expressing n as a sum of nonnegative integers (independent of order). The subsets corresponding to a partition of n constitute a particular symmetry pattern. Relative to these, a generalized process of symmetrization and alternation is defined that gives rise to special elements of the group algebra, the so-called Young symmetrizers, in terms of which the irreducible representation corresponding to the pattern can be described.

If a particular pattern has $m_1, m_2, \cdots, m_k$ elements in the respective subsets (listed in decreasing order of size), then the dimension of the space of the corresponding irreducible representation is the quotient of $s(n!)$ by $(n_1!)\,(n_2!)\cdots(n_k!)$, in which s is the product of all the differences $n_j - n_h$ with h less than j and $n_j = m_j + k - j$. Applied to S_5, for example, these results assert the existence of seven inequivalent irreducible representations corresponding to the partitions 5, $4+1$, $3+2$, $3+1+1$, $2+2+1$, $2+1+1+1$, and $1+1+1+1+1$, with dimensions 1, 4, 5, 6, 5, 4, and 1, respectively. It is no accident that the sums of the squares of these dimensions (120) is the order of S_5—it is a consequence of the fact that the regular representation takes place on 120-dimensional space, and each irreducible representation occurs in it with multiplicity equal to its dimension.

Topological groups and Lie groups. The applications of the theory of group representations to mathematical analysis and to physics entail, for the most part, the description of the representations of infinite groups that have a topological structure related to their group structure—the so-called topological groups. (The topological structure is a mathematical formulation of the concepts of nearness. The topological groups are those in which nearness is so defined that the product of an element and the inverse of another element is near to the group identity provided both elements are themselves near the identity.)

The most important and the most intensively studied of these groups are the Lie groups, named after the 19th-century Norwegian mathematician Sophus Lie. These are the topological groups in which the topology near each point is like that of n-dimensional Euclidean space for some fixed n. A more general class of topological groups, the locally compact groups, have received some attention. These include the discrete groups (those in which no group element is infinitely near the others—every group can be given such a topology), the Lie groups, and those that can be constructed from these two classes by certain group-theoretic processes. The various Euclidean spaces with vector addition as group multiplication and the complex numbers of modulus 1 (the circle group) with the usual multiplication of complex numbers each with its usual notion of nearness are examples of Abelian Lie groups. The invertible operators on n-dimensional Euclidean space, with nearness described in terms of nearness of the images of a specific vector and group multiplication the multiplication of operators, is an example of a non-Abelian Lie group, the general linear group of dimension n.

Haar measure of locally compact groups

The crucial property of locally compact groups is the existence of a measure on them, Haar measure (a notion of volume for subsets), which assigns the same number (measure) to a subset of the group and to the subset obtained from it by left multiplying each element of the subset by a fixed group element (each left translate of the set). Those groups with finite total measure constitute an especially important subclass, the compact groups. The circle group and the subgroup of the general linear group consisting of those operators that preserve the lengths of vectors, the orthogonal group (unitary group in the complex case), are examples of compact Lie groups. The finite groups are compact (discrete) groups.

Several different possibilities for group algebras and regular representations present themselves in the case of locally compact groups, each shedding some light on the representation theory of these groups. The set of finite sums of complex multiples of group elements is a direct generalization of the group algebra for a finite group, but it takes no account of the group topology. The integrable functions on the group with multiplication like that in the finite case (each function thought of as the sum of the products of each group element by the function value at that group element), except that integration replaces ordinary summation, is an effective generalization that respects the group topology. (This group algebra is the L_1 group algebra, and its multiplication is convolution multiplication of functions.) The space of square integrable functions on G (that is, $L_2(G)$) is a Hilbert space that is available for the various regular representations. To each group element the operator that left translates an L_2 function by g^{-1} (as in the finite group case) is assigned. This operator is unitary, and the algebra of finite sums of multiples of these operators is the (left) regular representation of the finite group algebra. Convolving an L_1 function with L_2 functions associates a bounded operator on $L_2(G)$ with the given L_1 function. This association is the (left) regular representation of the L_1 group algebra of G. The algebra of operators so obtained is not closed in the various relevant topologies on the set of bounded operators, and the various closures of this algebra are reasonable candidates for a group algebra (in particular, the uniform and the weak closures are of special interest).

It is not to be expected that a general unitary representation of G can be extended to the L_1 group algebra (an essentially topological construct) without some further topological restriction. The unitary representations that appear in practice are the strongly continuous representations, those that represent elements near the group identity by unitary operators (strongly) near the identity operator (equivalently: weakly continuous and measurable unitary representations); and these representations are so extendible to the L_1 group algebra. They are often but not always extendible to the uniform closure of this group algebra (in its regular representation on $L_2(G)$); and the question of just when they can be so extended is related to some of the most delicate problems of mathematical analysis. They are rarely extendible to the weakly closed group algebra; that is, to the weak closure of the regular representation of the L_1 group algebra.

The most satisfactory aspects of this general theory are to be found in its application to the classes of compact groups and of Abelian (locally compact) groups. The theory of representations developed for finite groups is valid for continuous representations of compact groups on Hilbert spaces, once summation over the finite group (divided by the group order) is replaced by integration with respect to Haar measure (normalized so that the group has measure 1). The equivalence of such representations with unitary ones, the character theory, direct sum decomposition into irreducible representations, the orthogonality relations, and the existence of a separating family of (finite-dimensional) irreducible unitary representations are valid for compact groups. The critical new fact that must be established in the general case is the finite dimensionality of each irreducible representation. The duality theory for finite Abelian groups carries over intact to the general case. The dual (character) group of a compact Abelian group is discrete (and that of a discrete group is compact). In particular, the character group of the circle is the additive group of integers—the correspondence between (L_2) functions on the circle and those on the integers (described for finite Abelian groups)

is the Fourier series expansion of periodic functions. The additive group of real numbers is its own dual, and the function mapping, in this case, is the important Fourier transform.

The decomposition problem

A foreshadowing of the complexities of the theory of representations for more general locally compact groups is afforded by the decomposition problem for reducible representations of Abelian groups that are locally compact but not compact. Here a direct sum decomposition into irreducible representations (characters) is not always possible as such. Rather, a direct integral decomposition, which bears the same relation to a direct sum decomposition as integration does to summation (ordinary addition), replaces the direct sum decomposition.

The representations of the 3-dimensional rotation group, R_3, supply an excellent and physically significant illustration of the theory for compact groups. The group R_3 consists of linear transformations of (real) 3-dimensional space that preserve the lengths of vectors and have determinant 1. The rotations preserve area on the unit sphere about the origin (best described in terms of 3-dimensional polar coordinates), so that each such rotation induces a unitary transformation on the Hilbert space of square integrable functions over the sphere. The resulting unitary representation decomposes into a direct sum of irreducible, finite-dimensional unitary representations. From a direct analytical study this representation can be shown to decompose into a direct sum of spaces of odd dimension, one for each odd number. This decomposition is effected by means of the spherical harmonic functions on 3-space and their associated surface harmonics on the sphere. The orthogonality relations for the surface harmonics establish the fact that this decomposition yields all the irreducible representations of R_3.

For general locally compact groups it is still possible to establish the existence of a separating family of strongly continuous irreducible unitary representations, and it is still possible to decompose reducible representations into a direct integral of irreducible representations; but much of the value of this fact is lost in the observation that the decomposition can often be performed in many totally distinct ways, so that no irreducible component of one decomposition is equivalent to any component of the others. These developments have given rise to the opinion that the appropriate process is not decomposition into irreducible representations but rather into factor representations. (A unitary representation is said to be a factor representation when the weak closure of the sums of multiples of the representing unitary operators has only multiples of the identity operator in its centre. Such an algebra of operators is said to be a factor.) The factors have been broadly classified into types depending on the nature of a dimension function on the projections in the factor (its discreteness or nondiscreteness, its finiteness or infiniteness). The so-called factors of type I are the most manageable type from the point of view of group representations. These are the factors isomorphic with the algebra of all bounded operators on some Hilbert space. The (direct integral) decomposition of a group representation into factor representations is essentially unique. When the resulting factors are of type I the decomposition of the representation into irreducible components is also essentially unique. The problem of the types of factor representations a specific group has (the so-called type problem) is central to the general theory. In particular, establishing that specific classes of groups have only type I representations is of vital importance to the general theory.

Lie algebra of a group

The detailed investigation of a Lie group and its representations proceeds via the study of its associated Lie algebra. If the Lie group with its associated topology is thought of as a surface in 3-dimensional space (or, more generally, a manifold in n-dimensional space), each vector in the tangent plane (space) at the group identity determines a unique vector at each group element by means of (left) translation by that group element. Relative to (*i.e.*, in the direction of) the invariant vector field so obtained, (differentiable) functions can be differentiated to give another function on the group. If two such vector fields are involved, differentiation is performed with respect to each successively in both orders, and the result of one is subtracted from that of the other; the mapping on functions obtained by this process is the same as that due to differentiation with respect to some third invariant vector field. This third vector field is said to be the Lie product or bracket of the other two. These vector fields, or—what amounts to the same thing because each is determined by a tangent vector at the group identity e—the tangent space at e, with the Lie product is the Lie algebra of the group (it is a nonassociative algebra). Because differentiation is a local process, depending only on the points near a point at which it takes place, the Lie algebra of a group is in reality a construct associated only with that portion of the group near the identity—the so-called local Lie group. Several Lie groups that are distinct globally may have the same local structure and hence the same Lie algebra; for example, the circle group and the group of real numbers both have the real line with the Lie product of each pair of elements 0 as their Lie algebra.

The distinction between Lie groups that have the same Lie algebra is a (global) topological one. Loosely speaking, they differ in the number and relation of the holes in them. A particular one of them covers the other; that is, maps homomorphically onto the other by a mapping that is an isomorphism near the identity. For example, the circle has a hole, the real line does not, and the real line may be wrapped homomorphically around the circle. In the family of Lie groups with a given Lie algebra, there is one that covers all the others—the so-called universal covering group of the others (also called simply connected covering group—the one that has no holes). The group of reals is the universal covering group of the circle. The rotation groups in dimensions 3 and higher are not themselves simply connected, but have a twofold universal covering group; each point in the rotation group is the image of precisely two points in this covering group. These are the spinor groups (Spin(n) for the n-dimensional rotation group R_n).

The Lie algebra of a Lie group may be thought of as a linear approximation to the local Lie group in the same sense that a tangent plane to a surface is the best planar approximation in the neighbourhood of the point of tangency. The multiplicative nature of the local group is made linear (additive) by the transition to the Lie algebra in much the same way that the logarithm converts multiplicative numerical problems into additive ones. In fact, there is a generalized logarithmic mapping (which preserves differentiability) carrying the local Lie group onto a neighbourhood of the origin in the Lie algebra—or, technically more manageable, an exponential mapping of the entire Lie algebra into the group. By means of this mapping, (local) questions about the group can be transferred to questions about the Lie algebra, which are more easily handled because of the added linearity. A (differentiable) homomorphism f of a Lie group G into another H induces a mapping of the functions on H into those on G and consequently of the differentiations with respect to invariant vector fields on G into those on H; *i.e.*, f gives rise to a mapping df of the Lie algebra h of H into the Lie algebra g of G. The mapping df is a homomorphism of g into h; *i.e.*, a linear mapping that preserves the Lie product.

Conversely, a Lie algebra homomorphism stems from a homomorphism of the local Lie group that may or may not be extendible to the full group. It is always extendible if the first group is simply connected. Applied to finite-dimensional representations of G (that is, to a homomorphism of G into the n-dimensional general linear group), this process yields a representation of g; for the Lie algebra of the general linear group can be viewed as the set of all n-dimensional operators with the Lie product of two operators A and B as $AB - BA$. The n-dimensional rotation group—and so Spin(n)—have the skew-symmetric operators of trace 0 (with the same Lie product) as its Lie algebra r. The study of the representations of r uncovers the existence of a representation corresponding to one of the local group that

cannot be extended to all of R_n. The spin representation of Spin(n) can be effected in terms of a construct known as the Clifford algebra. In the physically important case of R_3, Spin(3) can be identified as the group of 2-dimensional unitary operators with determinant 1 and also as the group of unit quaternions; the representation theory, in this case, is related to the physical phenomena that are associated with electron spin.

Clifford algebra

The deeper analysis of the representation theory for Lie groups, the finite-dimensional representations of which are completely reducible (the semisimple Lie groups), involves a profound analysis of the semisimple Lie algebras and the so-called theory of weights associated with their representations. It has been established for such groups that their unitary representations on Hilbert spaces are of type I.

BIBLIOGRAPHY. Elements of the theory of convergent series may be found in any book of calculus; the following are general references on infinite series. A good source is GODFREY H. HARDY, *A Course of Pure Mathematics* (1908, and many later editions). On a more advanced level are TOMLINSON FORT, *Infinite Series* (1930); THOMAS J. BROMWICH, *An Introduction to the Theory of Infinite Series*, 2nd ed. rev. (1926, reprinted 1955); KONRAD KNOPP, *Theorie und Anwendung der unendlichen Reichen*, 4th ed. (1947; Eng. trans., *Theory and Application of Infinite Series*, 2nd ed., 1951), elegantly written and containing a wealth of material; and ISIDORE I. HIRSCHMAN, *Infinite Series* (1962). See also GODFREY H. HARDY, *Divergent Series*, (1949); HARRY R. PITT, *Tauberian Theorems* (1958); KARL ZELLER, *Theorie der Limitierungsverfahren* (1958); and, for summability of integrals and related topics, K. CHANDRASEKHARAN and S. MINAKSHISUNDARAM, *Typical Means* (1952).

The following works deal with Fourier series and related topics. HENRI LEBESGUE, *Leçons sur les séries trigonométriques* (1906), although somewhat outdated, has considerably influenced the development of the theory. LEONIDA TONELLI, *Serie trigonometriche* (1928), deals with double Fourier series; while H.S. CARSLAW, *Introduction to the Theory of Fourier's Theory of Series and Integrals*, 3rd ed. rev. (1930); and DUNHAM JACKSON, *Fourier Series and Orthogonal Polynomials* (1941), contain elementary theory based on Riemann integration. GODFREY H. HARDY and WERNER ROGOSINSKI, *Fourier Series* (1944); ANTONI ZYGMUND, *Trigonometrical Series*, 2nd ed., 2 vol. (1959); SALOMON BOCHNER, *Harmonic Analysis and the Theory of Probability* (1955); EDWARD C. TITCHMARSH, *Introduction to the Theory of Fourier Integrals* (1937); NORBERT WIENER, *The Fourier Integral and Certain of Its Applications* (1933); STEFAN KACZMARZ and HUGO STEINHAUS, *Theorie der Orthogonalreihen* (1935); GABOR SZEGOE, *Orthogonal Polynomials*, rev. ed. (1959); LAURENT SCHWARTZ, *Théorie des distributions*, 2 vol. (1950–51); M.J. LIGHTHILL, *Introduction to Fourier Analysis and Generalised Functions* (1958); RICHARD R. GOLDBERG, *Fourier Transforms* (1961); ROBERT E. EDWARDS, *Fourier Series*, 2 vol. (1967–68); YITZHAK KATZNELSON, *An Introduction to Harmonic Analysis* (1968). The following are of an encyclopaedic nature: A. PLESSNER, "Trigonometrische Reihen," in ERNESTO PASCAL, *Repertorium der höheren Analysis*, 2nd ed., pt. 3 (1929); E. HILB and M. RIESZ, "Neuere Untersuchungen über trigonometrische Riehen," in *Enzyklopädie der mathematischen Wissenschaften* (II C 10, 1923–27); H. BURKHARDT, "Trigonometrische Reihen und Integrale (bis 1850)," *ibid.* (II A 12, 1904–16).

In addition to the references on Fourier series, the reader interested in harmonic analysis and integral transforms may wish to consult some of the following works: NINA BARI, *A Treatise on Trigonometric Series*, 2 vol. (1964; orig. pub. in Russian, 1961), which deals with the classical theory of Fourier series and integrals; ELIAS M. STEIN and GUIDO WEISS, *Introduction to Fourier Analysis on Euclidean Spaces* (1971), which treats the n-dimensional theory extensively and systematically and also has a systematic treatment of convolutions; and ISIDORE I. HIRSCHMAN and DAVID V. WIDDER, *The Convolution Transform* (1955), another good reference on convolutions. For variants of the Fourier transform (such as Hänkel or Mellin transforms), see EDMUND T. WHITTAKER and GEORGE N. WATSON, *A Course of Modern Analysis*, 4th ed. (1927, reprinted 1958); GEORGE N. WATSON, *A Treatise on the Theory of Bessel Functions* (1922); DAVID V. WIDDER, *The Laplace Transform* (1941, reprinted 1963); and THE BATEMAN MANUSCRIPT PROJECT, *Tables of Integral Transforms*, ed. by ARTHUR ERDELYI, 2 vol. (1954). The topic of singular integrals is treated in ELIAS STEIN, *Singular Integrals and Differentiability Properties of Functions* (1970). An exposition of abstract harmonic analysis may be found in WALTER RUDIN, *Fourier Analysis on Groups* (1962). LYNN H. LOOMIS, *An Introduction to Abstract Harmonic Analysis* (1953); and HERMANN WEYL, *The Classical Groups, Their Invariants and Representations* (1946), deal with both abstract harmonic analysis and representations of groups and algebras. Some additional references on representations of groups and algebras are CLAUDE CHEVALLEY, *Theory of Lie Groups* (1946); FRANCIS MURNAGHAN, *The Theory of Group Representations* (1938); LEV PONTRYAGIN, *Topological Groups*, 2nd ed. (1966); ANDRE WEIL, *L'Intégration dans les groupes topologiques et ses applications* (1940); HERMANN WEYL, *The Theory of Groups and Quantum Mechanics* (1949); RICHARD V. KADISON, *Representation Theory for Commutative Topological Algebra* (1951); and GILBERT B. ROBINSON, *Representation Theory of the Symmetric Group* (1961).

(A.Zy./C.L.F./Ed.)

Analysis, Functional

Analysis may be defined as that part of mathematics concerned with smooth abstract objects such as sets of numbers or sets of geometrical points or sets of functions that map numbers into numbers or points into points, and with the processes, called limit processes, that depend on a measure of closeness between numbers or points or functions. Analysis developed initially out of *ad hoc* arguments in which curves, geometric bodies, and physical motions were analyzed by an informal process of decomposition into infinitesimal parts.

This article is divided into the following sections:

I. A broad view of functional analysis
- Early development
- Linear functional analysis; theory of distributions
- Geometric principles; analysis of linear operators
- Hilbert space; spectral analysis
- Linearization as a method; ergodic theory, classical and generalized harmonic analysis
- Functional algebra
- Nonlinear functional analysis

II. Calculus of variations
- Mathematical formulation of variational problems
- Euler's equation
- Some questions involving arcs
- Multiple integral problems
- Morse's theory of critical points

III. Generalized functions, or theory of distributions
- Historical background
- Definitions and basic properties of distributions
- Regularization and convolution of distributions
- Locally convex spaces and their duals
- Fourier transformation of distributions

I. A broad view of functional analysis

EARLY DEVELOPMENT

Limit and approximation processes, the usefulness of which was discovered during this early period of development, came to be employed systematically in the 17th and 18th centuries in the differential and integral calculus of Sir Isaac Newton and Gottfried Wilhelm Leibniz, a German mathematician. Limit processes were applied by their immediate followers and by the important 19th-century French and German schools of analysts to sequences of values of a single function. The development of analysis in the 19th century led, however, to the realization that similar arguments could usefully be applied to whole sequences of functions and, further, that individual functions of interest could be studied to advantage using an indirect technique that established the properties of such a function by considering appropriate sequences of other, and generally simpler, functions that converged to it. This approach was pioneered by Jean-Baptiste-Joseph Fourier, a French mathematician who died in 1830, and it was developed subsequently by many other mathematicians, who improved and generalized the theory of Fourier series, which Fourier's work initiated. Studies of functions using this same method of approximation also played a central role in the theory of functions of a complex variable, the systematic work of the German mathematician Karl Weierstrass in the 19th century being especially notable in this connection. Out of repeated applications of the technique of functional approximation there grew the gradual realization that individual functions ought to be looked at as members of broader collections or spaces

Fourier's contribution

of functions. By focussing attention not on a single function f in isolation but rather on an appropriate collection of functions containing f, and on the special position or defining property of f within this collection, important properties of a particular function f could in many cases be established in a more ready and transparent fashion than was otherwise possible, old proofs could be simplified, and new proofs could be attained. In such studies it was often useful to make use of expressions assigning a value to each member of a space of functions; such expressions were dubbed functionals, this term being coined in analogy with the word function, which designates a rule assigning a value to each member of a collection of numbers. The body of mathematical work growing out of the viewpoint outlined above, in which functions are studied collectively, and in which the older methods of analysis are applied to whole collections of functions, consequently became known as functional analysis.

At the turn of the 20th century, the work of the Italian Vito Volterra and the Swedish mathematician Ivar Fredholm began to underscore the utility of the functional approach, and to make use of this approach in an explicit way. At about the same time, studies undertaken by David Hilbert (German), E. Schmidt (German), Frigyes Riesz (Hungarian), and others served to make plain the analogies between analytic and classical algebraic principles that emerged when collections of functions were regarded as abstract global totalities, and the value of regarding a function as a point in a space of functions became plain. This was recognized clearly by Schmidt, for example, in a 1908 study of systems containing infinitely many equations; Schmidt introduced various geometrical notations and terms, which are still in use, for describing spaces of functions. At approximately the same time the foundations of modern point-set topology, originally the study of geometric properties that remain unchanged under continuous maps, were being laid by Maurice-René Fréchet (French), Felix Hausdorff (German), and Riesz. These topological studies gave an abstract and extremely general form to the basic notions of continuity and differentiability that played a central role in the older analysis, thus allowing these notions to be carried over to spaces of functions. General definitions of the type of space of functions in which the new functional methods could be applied were offered a few years later by the U.S. mathematician Norbert Wiener and by the Polish mathematician Stefan Banach. In 1932 Banach's *Théorie des opérations linéaires* summarized much of the most interesting work in functional analysis up to that time, and inspired much of the work in functional analysis through succeeding decades. The important general class of Banach spaces take their name from him. At about the same time the Hungarian-born U.S. mathematician John von Neumann began a brilliant series of investigations into the functional analysis of Hilbert space (see below *Hilbert space; spectral analysis*), which is that particular space of functions bearing the closest geometrical analogy with ordinary Euclidean space. This work, which gave an abstract, transparent, and greatly generalized form to older analyses of integral equations by Hilbert and Schmidt, soon made contact with other broad areas of algebra and classical analysis, and has led to rich subsequent developments. Von Neumann's work had also a particularly close relation to, and influence on, the mathematical apparatus of quantum physics.

LINEAR FUNCTIONAL ANALYSIS; THEORY OF DISTRIBUTIONS

Linear properties

Many important spaces of functions are linear. A collection of functions possesses this property if the sum of any two functions belongs to the same collection, and if an arbitrary constant multiple of a function in the collection still belongs to the collection. For example, the sum of two continuous functions is continuous, the sum of two bounded functions is bounded, the sum of two smooth functions is smooth, etc., so that continuous functions constitute a linear space of functions, smooth functions another, bounded functions a third, etc. When a space of functions is linear, useful analogies exist between its elements and the vectors of ordinary two- and three-dimensional space. These analogies suggest many useful geometric lines of investigation, and the availability of addition and multiplication operations makes it possible to adapt many of the powerful algebraic procedures of conventional vector algebra or linear algebra to the study of function spaces. To emphasize this analogy, one often speaks of a collection of functions having the property of linearity as a linear space of functions, and of its elements as points or vectors.

(1) $$f = a_1 f_1 + \cdots + a_n f_n$$

(2) $$F(a_1 f_1 + a_2 f_2) = a_1 F(f_1) + a_2 F(f_2)$$

Many significant methods of functional analysis grow out of this analogy. Using terminology adapted from vector algebra, a function f which is the sum of terms of the type $a_k f_k$ (see Box, equation 1), in which $a_1, \cdots, a_n$ are numbers and $f_1, \cdots, f_n$ are functions belonging to the same space as f, is spoken of as a linear combination of $f_1, \cdots, f_n$; the expression $a_1 f_1 + \cdots + a_n f_n$ may be called the expansion of f in terms of $f_1, \cdots, f_n$. The same terms are also used in connection with infinite sums of the form $a_1 f_1 + a_2 f_2 + \cdots$, provided that these sums are convergent (*i.e.*, approach a limit) in an appropriate sense. In studying a problem set within some particular function space, it is often useful to find some particularly simple collection of functions $f_1, f_2, \cdots$ in terms of which every function may be expanded. The observation that such expansions could be useful for analysis goes back to the Swiss mathematicians Daniel Bernoulli and Leonhard Euler in the 18th century, who, in studying the motions of a vibrating string, found it useful to expand certain functions in terms of simpler ones. The British engineer Oliver Heaviside and many other 19th-century mathematicians showed that other similar series expansions were useful for treating a wide range of specific problems. Modern functional analysis affords a comprehensive generalization of all these earlier investigations, and makes plain the relatively simple algebraic-geometric basis on which they rest.

Expansion of functions

Many techniques related to the linear combination of functions are used within functional analysis. Certain important function-associated operations have the property that when applied to the sum of two functions they give a value that is the sum of the values obtained by applying the same operation to the two functions individually. Designating such an operation by the symbol F, this general property may be represented by an algebraic formula (see 2). An operation or functional having this property is said to be linear. Many of the central operations of classical analysis—as, for example, differentiation and integration—are linear: the integral of the sum of two functions in the sum of their integrals, etc. Taking an abstract and general view, functional analysis sees in the collection of all operations satisfying the condition (2) another space of objects, related in a natural way to an initial collection X of functions f. If the operations F in (2) are restricted to have numerical values, one speaks of F as a linear functional on the collection X; if the operations F have values that are themselves functions, one calls F a linear operator on X; if the values of F lie in X, F is called a linear operator in X. The collection of all linear functionals on X, the collection of all linear operators on X, and the collection of all linear operators in X, all form spaces associated with the space X in a natural way. The collection of all linear functionals on X is often called the conjugate space or dual space of X, and it is conventionally designated by the symbol X^*. The conjugate space X^* is important in functional analysis for a variety of reasons. Many operations that are defined for the elements of X carry over in a natural way to the elements of X^*. The study

of X^*, and of various detailed relationships between X and X^*, can yield important information concerning X. Moreover, consideration of the dual space X^* of an artfully chosen collection X may provide the key to the solution of a problem in which X itself plays only a subsidiary role.

Schwartz's theory of distribution

The theory of distributions, introduced in the late 1940s by the French mathematician Laurent Schwartz, gives a famous example of the systematic use of this last observation. Schwartz begins by considering the space of all extremely smooth functions f of a real variable y that are zero for sufficiently large values of the variable y. This collection X_0 of functions is one of the smallest (*i.e.*, most restricted) yet still quite general classes of functions that can be defined. Analytically important operations, such as differentiation and integration, are easily defined for the elements of X_0; the relationship between X_0 and X_0^* allows these operations to be carried over to the very much larger space X_0^*. The elements of X_0^* are called distributions by Schwartz. Because X_0 can be regarded as a subspace of X_0^*, the elements of X_0^* have a natural interpretation as generalized functions. Schwartz's construction makes differentiation, etc., available over a wide range of nonconventional mathematical objects, and in a manner preserving all the formal properties of these operations that are central to traditional analysis. There results a considerable broadening of the differential and integral calculus, including and generalizing many of the formal extensions of calculus developed on *ad hoc* lines for specific application, notably by Heaviside and the British physicist P.A.M. Dirac. Until Schwartz's development of the theory of distributions, most of these schemes of formal calculation lacked rigorous justification and were regarded as slightly illegitimate curiosities. Distribution theory has brought these and even more powerful tools for calculation into general mathematical currency (see below *Generalized functions, or theory of distributions*).

GEOMETRIC PRINCIPLES; ANALYSIS OF LINEAR OPERATORS

As already noted, linear functional analysis aims to exploit the formal similarity that exists between addition of functions and addition of vectors in two, three, and more dimensions and, accordingly, tends to consider any collection of functions having the necessary property of linearity as an infinite-dimensional analogue of the familiar Euclidean spaces. In accordance with this fundamental point of view, a systematic attempt is made to carry as many useful geometric notions as possible over from two- and three-dimensional Euclidean space to arbitrary linear collections of functions. One notion of this kind, which has been quite successfully exploited in functional analysis, is that of convexity. A body B in two- or three-dimensional Euclidean space is convex if, and only if, the entire straight-line segment S between any two of its points belongs to the body; if B is closed, this will be the case if, and only if, the midpoint of each such segment S belongs to the body. Phrased in terms of vector algebra, this last condition amounts merely to the demand that the midpoint $(f+g)/2$ of S belong to B whenever the endpoints f and g of S are both points of B; in this formulation, the notion of convexity generalizes immediately to any linear space of functions. This last observation suggests that many of the properties of convex bodies in two and three dimensions can be carried over to convex collections of functions within a linear space of functions; and such is indeed the case. Thus, for example, if one is given a convex body B in three-dimensional space and a point p not belonging to the body, one can always draw a plane with the point on one side of the plane and the body on the other side. If the equation of the plane is written $F(x)=c$, it follows that $F(p)$ is smaller than c, but that $F(x)$ is larger than c for each point x of the body B. This simple but basic geometric fact generalizes at once to convex collections B in linear spaces of functions, giving the so-called Hahn-Banach theorem, a theorem possessing many applications. This case will serve to illustrate the more general remark that functional analysis contains generalizations of much of the rich geometrical theory of convex bodies, in which corners, shortest paths, inscribed and circumscribed bodies, cones, and so forth, are all studied to advantage. Many of the most useful inequalities belonging to the older analysis of single functions turn out to have simple geometric interpretations when they are reformulated in functional-analytic terms.

Linear operators

As has already been noted, an operation F on functions with values $F(f)$ that are themselves functions satisfying the algebraic condition (see 2) is called a linear operator. Such a linear operator may be regarded as the natural generalization, to linear spaces of functions, of the rectangular matrices of conventional finite-dimensional vector algebra. If F is a linear mapping in a space X—*i.e.*, if the values $F(f)$ of F belong to the same space X as do the functions f for which F is defined—then F is analogous not merely to a rectangular but to a square matrix; *i.e.*, to a matrix having equally many rows as columns. Matrices of this type may freely be added and multiplied; the study of their particularly interesting algebraic relationships, which in the finite dimensional case was initiated in the 19th century by Arthur Cayley, a British mathematician, generalizes at once to linear operators. Given that F is a linear operator in a linear space X, the "square" of F, for example, may be defined as the transformation G specified by the formula $G(f)=F(F(f))$, and higher powers of F may be defined in analogous fashion. The elementary infinite series that expresses non-negative the reciprocal of $1-x$ as a sum of all integer powers of x (see 3) then generalizes in a straightforward way to linear operators, its generalization proving to be the key to many of those results of the early-20th-century theory of linear integral equations from which the historic development of modern functional analysis began. Other power series can also be used to define functions of a linear operator F, provided always that F acts *in* a single linear space X of functions; *i.e.*, that F is defined for each f in X and that each value $F(f)$ of F belongs to X. For example, the exponential function e^x has the power series expansion, with coefficients formed of reciprocals of factorials (see 4), which suggests the possibility of defining an operator exponential by writing a formal power series for the linear operator f with coefficients as before (see 5); here I denotes the unit, or identity, linear operator specified by the equation $I(f)=f$. This definition of an exponential function for linear operators is in fact quite useful, and provides the key to several important calculations and theorems. Further extension of this same technique yields a systematic calculus of functions of linear operators.

Other analogies between collections of operators and ordinary algebra have been quite fruitful for functional analysis. To give a further example, the exponential function a^t has the property that the product of two function values evaluated at distinct points is equal to the function value evaluated at the sum of the two corresponding points (see 6). Moreover, any continuous numerical function $e(t)$ having this multiplicative property —*i.e.*, any function satisfying the condition that the product of two function values for distinct points is equivalent to the function value for the sum of the distinct points (see 7)—is necessarily expressible as an exponential function; *i.e.*, $e(t)=a^t$. This fact suggests the study of fam-

(3) $$\frac{1}{1-x}=1+x+x^2+x^3+\cdots$$

(4) $$e^x=1+x+\tfrac{1}{2}x^2+\tfrac{1}{6}x^3+\cdots$$

(5) $$e^F=I+F+\tfrac{1}{2}F^2+\tfrac{1}{6}F^3+\cdots$$

(6) $$a^{t_1}a^{t_2}=a^{t_1+t_2}$$

(7) $$e(t_1)e(t_2)=e(t_1+t_2)$$

ilies $E(t)$ of linear operators depending on a real parameter t and satisfying the algebraic condition that is defined by the formula given as (8), namely: $E(t_1)E(t_2) = E(t_1 + t_2)$. Operator families satisfying this condition are called semi-groups of linear operators; they arise in a variety of physical, mathematical, and probabilistic settings. Naturally enough, their study centres around techniques for defining fractional powers A^t of a linear operator A, and around attempts to find conditions guaranteeing that a particular semi-group $E(t)$ of linear operators must have the exponential form $E(t) = A^t$.

HILBERT SPACE; SPECTRAL ANALYSIS

In the preceding paragraphs it was stressed that many of the most fruitful ideas of functional analysis arise in analogy with algebraic and geometric notions belonging to the vector analysis and the geometry of two- and three-dimensional Euclidean space. These analogies become most compelling if there is imposed upon the spaces of functions a condition that plays a fundamental role in Euclidean geometry: namely, the relation contained in Pythagoras' theorem. For easy generalization it is useful to state Pythagoras' relationship in the following slightly unconventional way: the sum of the squares of the lengths of the diagonals of a parallelogram is equal to twice the sum of the squares of the length of its sides. A linear space of functions in which every parallelogram has this property is called a Hilbert space, in honour of Hilbert, whose penetrating investigations of integral equations in certain particular spaces of this kind influenced the subsequent development of formal functional analysis in a particularly profound way. For such spaces it may be shown that the Pythagoras principle allows many other familiar geometric facts to be derived. For example, one can define a notion of the angle between two vectors in such a way as to preserve all the ordinary geometric properties of angles. This is done by analogy with the law of cosines of Euclidean trigonometry, according to which the square of the length of the vector sum of two functions is equal to the sum of the squares of the lengths of the two vectors taken separately minus a term equal to twice the product of the lengths of the vectors multiplied by the cosine of the angle θ between them (see 9). The geometrically significant expression $|f||g|\cos\theta$, which has particularly simple algebraic properties, and which plays a fundamental role in geometric investigations of Hilbert space, is then called the inner product of the vectors f and g and is often designated by the symbol (f, g).

Relation of Pythagoras' theorem to Hilbert space

(8) $$E(t_1)E(t_2) = E(t_1 + t_2)$$

(9) $$|f+g|^2 = |f|^2 + |g|^2 - 2|f||g|\cos\theta$$

(10) $$[F(f), g] = [f, F(g)]$$

(11) $$g(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} f(y)e^{ixy}\,dy$$

In analogy with ordinary geometry, a linear transformation R in a Hilbert space is called a rotation if it preserves lengths; *i.e.*, if the length of the vector $R(f)$ is in every case the same as the length of the vector f. In Hilbert space, rotations and other related basic geometric transformations continue to have many of the familiar properties possessed by similar transformations in ordinary Euclidean two- and three-dimensional space. Orthogonal (mutually perpendicular) sequences $f_1, f_2, f_3, \cdots$ of functions in Hilbert space—that is, sequences of functions for which the inner product (f_i, f_j) is equal to zero whenever the indices i and j are different—appear as the geometric analogue of sets of mutually perpendicular axes in ordinary Euclidean space. The process of expanding a function as a linear combination of such vectors is then seen to be closely analogous to the geometric decomposition of a vector into components directed along several mutually orthogonal axes; this gives a transparent geometric meaning to the expansion into sines and cosines introduced by Fourier.

In Hilbert space it is possible to distinguish a class of linear operators, the study of which constitutes one of the most successful branches of the whole of functional analysis. A linear operator F in Hilbert space is said to be self-adjoint if it satisfies the simple algebraic condition that identifies two inner products, one with the left-hand element being the effect of the linear operator x operating upon a vector and one being an inner product in which a right-hand number is the effect of the linear operator f operating upon a second vector (see 10).

Self-adjoint linear operators

Given a bounded function $m(x)$, the mapping, which sends each function $f(x)$ into the product function $m(x)f(x)$, is linear; a mapping M of this particularly simple kind is called a multiplication operator. Multiplication operators, which are the simplest general type of linear operator, self-adjoint (see 10). The fundamental spectral theorem of Hilbert and von Neumann asserts that, given any self-adjoint linear operator in Hilbert space, one can always find a rotation R of the Hilbert space and a multiplication operator M such that $R(F(f)) = M(R(f))$ for each vector f of the Hilbert space; this means that in the rotated axes defined by the rotation, R, F is expressed, in the simplest possible way, as an operation of multiplication. Many other useful properties of self-adjoint linear operators in Hilbert space follow from this basic theorem, and, generally speaking, the spectral theorem tends to reduce the study of a particular self-adjoint operator F to the problem of finding the rotation R that allows F to be expressed as a multiplication. Many of the most interesting linear transformations of classical analysis turn out to be rotations of this type. This observation covers not only Fourier's initial sine–cosine series expansion and all the Sturm-Liouville generalizations of this expansion, but also many related integral expansions, of which the most famous is the Fourier integral, which expresses a rotated version $g = R(f)$ of an initially given function f by the formula that expresses g as a definite integral extended over the entire real line, the integrand being the product of f and an exponential function, multiplied by appropriate normalizing factor (see 11).

Linear operators satisfying the self-adjointness condition (see 10) are of common occurrence in classical mathematical physics, wherein the self-adjointness condition is often a consequence of some underlying law of energy conservation. They also are of fundamental importance in quantum theory, in which the self-adjointness condition appears as an assumption basic to the physical interpretation of the theory.

The spectacular conceptual and computational simplifications afforded in the study of self-adjoint operators by the spectral theorem has led to a long series of efforts to generalize this theorem to linear operators F not satisfying the basic condition (see 10) of self-adjointness. Unfortunately, these efforts, in which the works of the U.S. mathematician Nelson Dunford and the Romanian C. Foiaş have featured prominently, have been successful only in part. The fact is that general linear operators are considerably more complex than self-adjoint operators, even if attention is restricted to operators acting in Hilbert space. Moreover, phenomena appear in the study of general operators in Hilbert space that have no analogue in the more elementary algebraic theory of finite square matrices acting in spaces of vectors. In spite of this, it has been possible to find particular classes of non-self-adjoint operators having many properties in common with multiplication operators and with other special classes of linear operators of simple structure for which calculation is easy.

LINEARIZATION AS A METHOD; ERGODIC THEORY, CLASSICAL AND GENERALIZED HARMONIC ANALYSIS

As has been noted, linear functional analysis makes systematic use of the possibility of expressing one function

as a linear combination of others, thereby attaining great simplifications in the problems that it studies. Problems for which this is not possible—*i.e.*, cases in which it is necessary to deal with sets of points or of functions within which an operation of addition is not available in any useful way—are normally much more complicated. When problems of this kind are attacked it is generally necessary to have recourse to the elaborate tools of global (or algebraic) topology, and even so the attack may be unsuccessful. Global topological methods began to be used extensively in functional analysis around 1960, though work in the 1930s by the mathematicians Julius Schauder (Polish) and Jean Leray (French) and by the U.S. topologist Marston Morse introduced some basic techniques.

The use of global topological methods in combination with more conventional techniques of functional analysis defines what has come to be called non-linear functional analysis (see below). The relative simplicity of linear spaces suggests, however, that whenever it is possible to relate a linear space to an object to be studied, even if this object is initially nonlinear, analysis of the linear space may allow interesting properties of the original object to be established. This technique, which may be called linearization, has been applied successfully in a great variety of rather important investigations. Ergodic theory, which grew originally out of statistical mechanics (that is, out of the statistical analysis of physical systems containing a large number of particles of interest), constitutes an important instance of this general remark.

Ergodic theory

If at a given instant one records the momentary position and velocity of each particle in a system of particles, one obtains a multidimensional vector v; v is called the momentary state of the system, and the collection of all relevant vectors v is called the system's phase space. It follows from the laws of Newtonian mechanics that after t seconds have elapsed a system initially in a state v will find itself in a state $\phi_t(v)$, in which the mapping ϕ_t is determined by the details of the forces that each particle exerts on all the others. The Newtonian law of energy conservation implies that even if the time parameter t is allowed to vary, all the states $\phi_t(v)$ will lie on some single energy surface. Given this fact, detailed investigation of the overall motion of a system will naturally focus on one single such surface and on the transformations $\phi_t(v)$, all of which map each point of such a surface to some other point of the same surface. These energy surfaces are, however, described by complicated equations and are in consequence highly curved, nonlinear objects. This characteristic makes direct study of the transformations ϕ_t difficult. Using the technique of linearization described above, some of these difficulties may be alleviated, as follows.

An appropriate linear space of functions $f(v)$, defined for all the points v of an energy space, is set up, and the nonlinear maps ϕ_t are used to define linear maps F_t in this space of functions; F_t is defined simply by specifying that $F_t(f)$ is to be the function g given by the formula in which the value of g at v is equal to the function f evaluated at a nonlinear map of the argument v (see 12). A suitable choice of functions f will yield a Hilbert space within which the maps F_t turn out to have properties that allow them to be studied using the Hilbert space spectral theorem mentioned in the preceding section. This makes it possible to establish a number of interesting facts concerning the nonlinear maps ϕ_t. The method sketched above has been used in a series of investigations stretching from the 1930s to the present day, in which the names von Neumann, Birkhoff, Hopf, Kolmogorov, Sinai, Arnold, and Ohrenstein feature prominently. These investigations have elucidated many properties of nonlinear mappings significant for mechanics, probability theory, and geometry.

A closely related method has played a central role in the elucidation of important algebraic problems. In abstract algebra, a group is a collection of elements x, y, etc., any two of which may be multiplied to give a product xy, any one of which has an inverse x^{-1} satisfying $x^{-1}(xy) = y$ and $(yx)x^{-1} = y$. If a group is given, the linear space of all functions f on the group may be considered, and for each group element x a linear mapping F_x can be defined by requiring that $F_x(f)$ be the function g specified by the simple formula identifying the function value of g with the function value of f, the argument of the former being y, the argument of the latter being the group product yx (see 13). If appropriate conditions are imposed upon the functions f, a Hilbert space of functions will result; each of the mappings F_x will then be a linear transformation to which the spectral theorem can be seen to apply. This theorem then implies that one can find functions called group characters and irreducible representation coefficients, related to the mappings F_x in particularly simple ways. The line of reasoning just described was initiated by the German mathematicians Ferdinand Georg Frobenius and Issai Schur, who analyzed the properties of finite groups using this method. The gradual development of functional analytic technique subsequently made it possible to extend Schur's techniques to important groups containing infinitely many elements, including the groups of three-dimensional and of higher-dimensional rotations, and of the Lorentz group, which is important in all relativistic physical theories. In this connection the work of the German-born U.S. mathematician Hermann Weyl, of the contemporary Russian mathematicians I.M. Gelfand and M.A. Naymark, and of Harish-Chandra in the U.S. has been particularly important.

$$g(v)=f[\phi_t(v)] \tag{12}$$

$$g(y)=f(yx) \tag{13}$$

$$S(f)=F(f)+G(f) \tag{14}$$

$$P(f)=F[G(f)] \tag{15}$$

$$\begin{cases}(F+G)+H=F+(G+H)\\ F(GH)=(FG)H\\ F(G+H)=FG+FH\end{cases} \tag{16}$$

Because of the close correspondence (see 13) between the mappings F_x and the fundamental group multiplication operation yx, group characters and irreducible representation coefficients are most useful for analyzing the algebraic structure of a group. They inevitably play a central role in the understanding of any system of functions having a given group of symmetries. Such function systems occur in crystallography, in the study of sound and electromagnetic waves, and in the quantum theory of nuclear, atomic, and molecular energy levels and scattering processes.

FUNCTIONAL ALGEBRA

If a linear space of functions is given, and if F and G are both linear operators acting in this space, then F and G can be added and can be multiplied together. The sum of F and G is the linear operator S defined by the formula $S(f) = F(f) + G(f)$ (14). The product of F and G is the linear operator P defined by the formula $P(f) = F[G(f)]$ (15). The sum of F and G is customarily designated by $F + G$, and the product of F and G by FG.

The algebraic combinations of linear operators (see 14, 15) have many of the formal properties that are familiar from ordinary algebra; for example, the identities known as the laws of associativity for addition and multiplication as well as the distributive law (see 16) all hold. This observation suggests that the methods of algebra can fruitfully be applied to the study of collections of linear operators. This is indeed the case; the subject area that arises when algebraic techniques are combined with the methods of functional analysis is called functional algebra.

Gelfand's contribution

Functional algebra began to develop rapidly after 1941, in which year a famous paper of Gelfand demonstrated that important earlier work of Wiener on Fourier series could be conceptually simplified by reformulating it in algebraic terms, and that some of the methods of functional analysis could be fruitfully combined with abstract algebraic arguments and with reasonings adapted from the classical theory of functions of a complex variable. The theory initiated by Gelfand subsequently achieved many points of contact with profound questions in classical analysis.

A brilliant series of papers by von Neumann and F.J. Murray of the United States led at about the same time to the development of yet another major branch of functional algebra. Von Neumann and Murray considered algebras of operators in Hilbert space; that is, sets V of linear operators having the property that $F + G$, FG, and cF all belong to V whenever F and G belong to G (c is an arbitrary numerical constant). A linear operator F in Hilbert space can always be written as $F = H_1 + iH_2$, in which H_1 and H_2 are self-adjoint and i is the square root of -1; the operator F^*, which is called the adjoint of F, is then related to F in much the same way that the complex conjugate $\bar{z}$ of a complex number z is related to z.

It had been known for some time that *-closed algebras V of finite square matrices—that is, algebras having the property that F^* is a member of V whenever F is a member of V—had a particularly simple structure, and von Neumann and Murray set out to generalize this structural fact to *-closed algebras of operators in Hilbert space. The class of algebras that these two studied are currently known either as W^*-algebras or as von Neumann algebras.

In the von Neumann–Murray theory, a far-reaching generalization of the Hilbert space spectral theorem may be established, according to which a general W^*-algebra is decomposable into a family of simpler W^*-algebras, which were called factors by von Neumann and Murray. The simplest factors are finite or infinite matrix algebras. More complex factors also exist, and these are related in various interesting ways to algebraic structures such as infinite discrete groups, the decomposition of a general W^*-algebra into factors providing a quite useful tool for the treatment of a number of the main properties of group characters and irreducible representation coefficients.

NONLINEAR FUNCTIONAL ANALYSIS

As has been noted above, nonlinear functional analysis attempts to combine the global methods of topology with the approximation techniques used in linear functional analysis, thereby making it possible to study systems of functions in which the simplifications afforded by linearity are not available. This part of functional analysis was initiated in the late 1920s by Schauder, who showed that the Brouwer fixed-point theorem of topology could be generalized to linear functional spaces. Subsequent work by Schauder, and by Schauder and the French mathematician Jean Leray, extended other basic theorems of topology to linear functional spaces and applied these extended principles to the study of various interesting classes of partial differential equations, including the equations of fluid flow. Their work revealed the topological basis underlying several subtle existence theorems of classical analysis.

At about the same time Morse established some important results concerning the behaviour of smooth functions on curved surfaces, showing that the number of critical points of such a function (*i.e.*, the number of points at which the rate of change of the function in every direction is zero) is closely related to certain topological properties of the surface. Morse himself extended his results from finite-dimensional surfaces to spaces of curves; the suggestion, implicit in this pioneering work, that the topology of curved functional spaces ought to be studied, lay dormant for 20 years, probably because tools adequate for the study of the global topology of infinite-dimensional objects were initially not available. In the late 1950s, however, the leads left by Morse began to be pursued actively. Subsequently, the area of contact between global topology and functional analysis has broadened significantly, topology having contributed to the solution of nonlinear analytic problems, and the study of certain linear and nonlinear functional spaces having provided keys to important questions in pure topology.

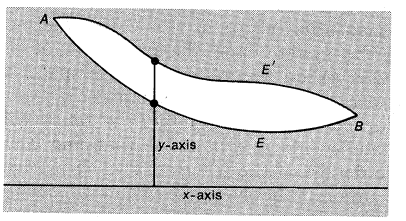
(J.T.Sc.)

II. Calculus of variations

The calculus of variations is an old subject; some of its problems were considered and partially solved by the ancient Greeks. Leonhard Euler in the mid-18th century deduced the first general rules for dealing with the subject, while much of the terminology was introduced soon after by the 18th-century French mathematician Joseph-Louis Lagrange.

When two points, A and B, are given in a plane, as shown in the accompanying figure, there is an infinity of arcs joining them. A simple problem of the calculus of variations is that of finding in this group of arcs (such as E and E') one that has the shortest length, the solution of the problem being a straight line segment. It may also be desired, however, to find in the group of arcs joining A with B one down which a particle, started with a given initial velocity, will fall from A to B in the shortest time; or which one of these arcs, when rotated about the x-axis, will generate a surface of revolution of minimal area.

These examples are typical problems of the calculus of variations of the so-called simplest type.



Simple problem of arc length (see text).

These problems illustrate the usual situation in the calculus of variations, in which mathematicians seek to find that arc from some given class for which some quantity, the value of which depends on the entire arc (its length, etc.), is a minimum or a maximum. The calculus of variations also deals with problems involving surfaces or functions of several variables. For example, if a circular wire is bent in any way, dipped in a soap solution, and then withdrawn, the soap film spanning the wire will assume the shape of the surface of least area bounded by the wire.

The calculus of variations has been useful as a unifying principle in mechanics and as a guide for the determination of new laws of physics.

Albert Einstein's theory of general relativity, for example, utilizes the calculus of variations extensively. One of the most widely applicable variational statements of classical mechanics is known as Hamilton's principle (after the 19th-century Irish mathematician William Rowan Hamilton); it states that the trajectories of many dynamical systems are the solutions of some variational problem involving an energy integral.

The introduction of the calculus gave great impetus to the study and to the solution of variational problems. After a number of special problems had been solved, Euler in 1744 deduced the first general rule, now known as Euler's differential equation, for the characterization of the maximizing or minimizing arcs.

MATHEMATICAL FORMULATION OF VARIATIONAL PROBLEMS

All of the problems mentioned in the first paragraph of this section reduce to the determination of that differentiable arc $y = y(x)$ joining the two points (x_1, y_1) and (x_2, y_2) that minimizes an integral extended over a segment of the x-axis and having its argument a function of

x, of y, and of the derivative of y (see 17). The length of the arc and the area of the surface of revolution obtained by revolving it around the x-axis are found from elementary calculus to be given respectively by the integrals of the general type described above in which the function in the integrand involves the square root of one plus the square of the derivative of y (see 18). In the case of the curve of steepest descent, the time required for a bead to descend along a wire in the shape of an arc from the point (x_1, y_1) to the point (x_2, y_2) under the action of gravity is given by the integral of the above type with a special form in the integrand that accounts for the special type of problem involved (see 19) if friction is neglected; here the term α in the integrand is of the form v_1^2 divided by twice g (see 20), in which v_1 is the initial velocity of the bead and g is the acceleration of gravity.

It is the case in the preceding examples that the function $f(x, y, p)$ occurring in the minimizing integral (see 17) has one of several forms, each involving the square root of one plus p^2 (see 21). The integrals arising from Hamilton's principle are of the type in formula (17), in which x is replaced by the time t (this usually does not appear in the function f in such applications) and the single variables y and p are replaced by several variables $y_1, \cdots, y_n$, and $p_1, \cdots, p_n$, the p_i standing for the derivatives dy_i/dt.

EULER'S EQUATION

Euler's differential equation for the determination of a minimizing arc is derived as follows: Suppose that $y = y(x)$ is a minimizing arc that passes through the two points, and assume that the second derivative y'' and the second partial derivatives of f are all continuous. If $\eta(x)$ is any differentiable function which vanishes at the end points x_1 and x_2 (see 22), then for any λ the arc obtained by adding to the function y the function λ times η (see 23) is another arc joining the two given points. Forming an integral for this new arc gives a function $I(\lambda)$ in the form of a definite integral in which λ appears as a parameter (see 24) which is differentiable and must have a minimum for $\lambda = 0$.

From elementary calculus, it follows that the integral involving λ can be differentiated with respect to the parameter λ. A necessary condition for a minimum is that this derivative should be equal to 0 (see 25). The integral that results is called the first variation of the original integral (see 17); this corresponds roughly to the total differential of a function of several variables, the function $\eta(x)$ corresponding to the set of differentials of the variables. Now, using the differentiability assumptions, the term involving $\eta'(x)$ in the first variation (see 25) may be eliminated by integrating by parts to obtain an integral of which the integrand includes η as a factor, the total integral being equal to 0 (see 26).

Because the last integral expression (see 26) holds for every function $\eta(x)$ allowed above, it follows from a well-known theorem in analysis that the first derivative of f_p is restricted to equal f_y (see 27), which is Euler's differential equation; it is understood in the derivation that the values x, $y(x)$ and $y'(x)$ are to be inserted in the expressions for f_p and f_y.

If Euler's equation (see 27) is written out in more detail, the rules of differentiation yield an equation in which the second derivative of y appears (see 28), which shows that Euler's equation is of the second order. Euler's equation becomes singular, however, if $f_{pp} = 0$. Therefore, regular variational problems are those for which f_{pp} never vanishes; in that case, it is assumed that $f_{pp} > 0$, which makes minimum problems more natural than maximum problems.

SOME QUESTIONS INVOLVING ARCS

The analysis leading to Euler's equation applies just as well to a maximizing arc as to a minimizing one. The first attempt to find a condition which would single out a minimizing arc was to consider the second variation, which is obtained, for any given $\eta(x)$, by finding $I''(0)$, $I(\lambda)$ being defined by (24).

Clearly $I''(0) \geqslant 0$ for every $\eta(x)$ if $y(x)$ is a minimizing arc, but the condition that $I''(0) \geqslant 0$ for every $\eta(x)$ is not sufficient to ensure that $y(x)$ is minimizing. The attempt to find conditions that would ensure a given arc being a relative minimum led to a variety of rather interesting discoveries and also raised a number of new questions.

After Weierstrass' famous example of a continuous function that does not possess a derivative at any point (thereby showing that a function with a continuous first derivative does not necessarily have to have a continuous second derivative), the question then was raised as to why a minimizing arc, or, more generally, why any arc for which the first variation is zero for every η, should possess a continuous second derivative, as is required in the derivation of Euler's equation. By the end of the 19th century, the problem of the existence of a minimizing arc had been raised, and the existence had been proved by Hilbert under certain hypotheses. Leonida Tonelli (an Italian mathematician) developed the so-called direct methods of the calculus of variations in order to simplify and amplify the results of Hilbert concerning the existence of minimizing arcs.

These results were applied to very general situations by the metric geometry methods of K. Menger (German) and Herbert Busemann (U.S.). Most of the remaining important problems involving arcs were solved between 1910 and 1940 by such mathematicians as O. Bolza (German), G.A. Bliss (U.S.), E.J. McShane (U.S.), and others.

The fact that not every minimum problem has a solution is seen by studying a definite integral extending from 0 to 1 symbolized by I in which the integrand in-

(17) $$\int_{x_1}^{x_2} f[x, y(x), y'(x)]\,dx, \qquad y' = \frac{dy}{dx}$$

(18) $$\begin{cases} \displaystyle\int_{x_1}^{x_2} \sqrt{1+\left(\frac{dy}{dx}\right)^2}\,dx, & \text{length of arc} \\ \displaystyle\int_{x_1}^{x_2} 2\pi y(x)\sqrt{1+\left(\frac{dy}{dx}\right)^2}\,dx, & \text{area of surface} \end{cases}$$

(19) $$\int_{x_1}^{x_2} (\alpha - y)^{-\frac{1}{2}} \sqrt{1+\left(\frac{dy}{dx}\right)^2}\,dx$$

(20) $$\alpha = y_1 + v_1^2/2g$$

(21) $$\sqrt{1+p^2},\ 2\pi y\sqrt{1+p^2},\ \sqrt{1+p^2}/\sqrt{\alpha - y}$$

(22) $$\eta(x_1) = \eta(x_2) = 0$$

(23) $$y = y(x) + \lambda\eta(x)$$

(24) $$I(\lambda) = \int_{x_1}^{x_2} f[x, y(x) + \lambda\eta(x), y'(x) + \lambda\eta'(x)]\,dx$$

(25) $$I'(0) = \int_{x_1}^{x_2} \{\eta(x) f_y[x, y(x), y'(x)] + \eta'(x) f_p[x, y(x), y'(x)]\}\,dx = 0$$

(26) $$\int_{x_1}^{x_2} \eta(x)\left\{f_y - \frac{d}{dx} f_p\right\} dx = 0$$

(27) $$\frac{d}{dx} f_p = f_y$$

(28) $$f_{pp}\frac{d^2y}{dx^2} + f_{py}\frac{dy}{dx} + f_{px} = f_y$$

cludes the derivative of y (see 29). In this case $0 < I < 1$ for every $y(x)$. Setting $y(x) = x^n$, $n = 1, 2, \cdots$, shows that I may be made arbitrarily close to 1. By taking $y(x)$ to be equal to x plus a sinusoidal function (see 30), it is seen that I may be made arbitrarily close to zero. There is, however, no $y(x)$ for which I is equal to zero or one.

(29) $$\begin{cases} I = \int_0^1 \{1 + [y'(x)]^2\}^{-1}\, dx \\ y(0) = 0, \qquad y(1) = 1 \end{cases}$$

(30) $$y(x) = x + n \sin n\pi x, \qquad n = 1, 2, \cdots$$

(31) $$\begin{cases} I = \iint_G f[x, y, z(x, y), z_x(x, y), z_y(x, y)]\, dx\, dy \\ z_x = \partial z/\partial x, \qquad z_y = \partial z/\partial y \end{cases}$$

(32) $$\frac{\partial}{\partial x}(f_p) + \frac{\partial}{\partial y}(f_q) = f_z, \qquad p = z_x, \qquad q = z_y$$

(33) $$\tfrac{1}{2} \iint_G (z_x^2 + z_y^2)\, dx\, dy$$

(34) $$f(x_1 + \eta_1, \cdots, x_n + \eta_n) - f(x_1, \cdots, x_n) = \tfrac{1}{2} \sum_{i,j=1}^{n} a_{ij}\eta_i\eta_j + R$$

(35) $$\alpha_{ij} = \partial^2 f / \partial x_i \partial x_j$$

(36) $$-\xi_1^2 - \cdots - \xi_r^2 + \xi_{r+1}^2 + \cdots + \xi_s^2$$

MULTIPLE-INTEGRAL PROBLEMS

Meanwhile interest had developed in problems involving functions of two or more variables, and the derivation of Euler's equation had been extended to integrals of a more complicated form (see 31). Replacing z by $z + \lambda\zeta$, in which ζ has a role similar to that of η in the preceding problem, $\zeta = 0$ on the boundary of the region G, and proceeding as in the derivation of Euler's equation (see 24–26), results in the equation known as Euler's equation (see 32) for the multiple integral (see 31).

Early in the 19th century Carl Friedrich Gauss (German) and W. Thompson had observed that Laplace's equation (which had already been recognized as having important applications) was the Euler equation for a particular integral (see 33), as is seen by setting $f(x, y, z, p, q) = (p^2 + q^2)/2$ in the derivation of Euler's equation for the multiple integral (see 31, 32).

Dirichlet's principle

This fact is the formal basis for what is now known as Dirichlet's principle (which was named after the German mathematician Peter Gustav Lejeune Dirichlet), which states that (subject to certain hypotheses) there is a unique function z that minimizes the integral of the sum of squares of z_x and z_y (see 33) among all differentiable functions having given boundary values and that that unique function is the solution of Laplace's equation with those boundary values.

This principle has proved to be of great importance not only in potential theory and in the theory of partial differential equations, but also in the theory of functions of a complex variable and in conformal mapping. The applications to these fields was begun by Bernhard Riemann (German) around 1850 and important advances have been made in them since 1935 by such mathematicians as Richard Courant, Lars V. Ahlfors, D.C. Spencer, M.M. Schiffer, P.R. Garabedian and others. Riemann simply assumed the principle but, of course, the objections that were raised concerning problems involving arcs applied to the Dirichlet principle with even greater force; the first proof of the principle was given by Hilbert in 1900.

The Plateau problem

The almost simultaneous solutions by Jesse Douglas (U.S.) and Tibor Radó (Hungarian-born) in 1930 of the problem of the Belgian physicist Joseph Antoine Ferdinand Plateau, *i.e.*, the proof of the existence of a minimal surface (surface of least area) bounded by a simple closed curve in space, aroused a renewed interest in variational problems for surfaces. The Plateau problem was solved for surfaces bounded by several curves and for more complicated surfaces by Douglas, Courant, and others. More general variational problems for surfaces were solved by Tonelli (Italian), McShane (U.S.), Charles B. Morrey, Jr. (U.S.), Lamberto Cesari (Italian-born), A.G. Sigalov, and others between 1930 and 1957 by extending the direct methods of Tonelli; these involve considering the surfaces as points in a space in which limits are defined and using the general theorem in functional analysis that a function that is lower-semicontinuous on a compact set attains its minimum on that set. In order to apply this theorem it was found expedient to enlarge the classes of arcs or surfaces, etc., within which the minimum is sought. Unfortunately, the enlarged classes usually contain surfaces that do not possess the properties of differentiability required in the derivation of Euler's equation.

In the case of the Plateau problem the differentiability of the solutions follows from the Dirichlet principle. The differentiability of the generalized solutions of certain other problems for surfaces was proved by Morrey, and some of his results were extended in 1957 to problems in higher-dimensional spaces by E. De Giorgi and J. Nash, Jr. (U.S.).

An interesting attack on problems for surfaces has been made by L.C. Young (U.S.), W.H. Fleming (Canadian), and others through their introduction of generalized surfaces.

MORSE'S THEORY OF CRITICAL POINTS

If a function $f(P)$ is defined on an n-dimensional manifold, its behaviour near a particular point may be studied by introducing coordinates, in which case f becomes just a function of n coordinates $x_1, \cdots, x_n$. A point (*i.e.*, set of numbers $x_1, \cdots, x_n$) is a critical point of f if all its partial derivatives are zero there. If $x_1, \cdots, x_n$ is such a point, then a formal finite difference constructed with the function f equals the summation of terms plus a final term symbolized by R (see 34), in which R consists of the higher order terms; here the coefficients in the summation of terms are obtained from the second-order partial derivatives (see 35) evaluated at the point. It is known that linear combinations $\xi_1, \cdots, \xi_n$ of the η_i can be chosen so that the double sum becomes expressed in terms of the squares of the ξ_i, the indices ranging from one through r to s (in which r is zero if there are no negative terms, etc.; see 36). Moreover, the numbers r and s are independent of the coordinate system and the way the ξ_i are chosen as long as the given quadratic form (see 36) is obtained. If $s = n$, the critical point is said to be nondegenerate, in which case r is called the index.

Marston Morse, a 20th-century U.S. topologist, first proved certain inequalities between the number of critical points of a given index of a (nondegenerate) function on a manifold and the so-called Betti numbers (after the 19th-century Italian mathematician Enrico Betti) of the manifold. His greatest contribution was to generalize the notion of a nondegenerate critical point and his inequality results to situations in which the functions and spaces were too general for differentiability to be defined. Morse applied his theory to spaces the elements of which were arcs or closed curves on a manifold or surfaces in space. He was able to prove the existence of unstable minimal surfaces (*i.e.*, surfaces furnishing some critical value other than a minimum to the area) and to prove such theorems as that there exist at least three closed shortest paths on a sphere. His theory has been applied to variational problems for arcs with great success. Stephen Smale (U.S.) and others have found a way to apply the Morse theory to multiple integral problems.

(C.B.My.)

III. Generalized functions, or theory of distributions

HISTORICAL BACKGROUND

In the second half of the 19th century mathematical analysts had raised their sights beyond the relatively orderly world of analytic functions—that is, functions with convergent Taylor expansions (after the English mathematician Brook Taylor)—and had begun to study and handle nonanalytic and even nondifferentiable functions. This led to the discovery of various kinds of pathological behaviour, or behaviour that was unfamiliar to the classically trained intuition. The complexity of the picture, in the theory of functions of real variables, was further compounded with the appearance of Lebesgue's integration theory (formulated by the French mathematician Henri-Léon Lebesgue), one of the great events in the history of mathematics. Standard ingredients of Lebesgue's theory are measurable and integrable functions and sets, which can exhibit irregularities of a degree and of a nature unknown until then. The Lebesgue theory had a strong influence on the investigation of integral equations. The space $\mathfrak{L}^2$—the space of (Lebesgue) square integrable functions—was recognized by Hilbert to be the right framework for the study of many problems in the calculus of variations. What is now called its Hilbert space structure was described and exploited, notably by Hilbert and by Riesz, to unify and reinterpret earlier results on integral equations, among them the pioneering work of Fredholm, in which ideas from finite-dimensional linear algebra had been transposed with unexpected success.

Lebesgue's integration theory

In 1894, quite off the mainstream of research in pure mathematics, Heaviside introduced his symbolic calculus to help solve the ordinary differential equations that arise in the theory of electrical circuits. His suggestion was simple and daring: replace the differentiation operator d/dx by the variable p, which amounts to replacing the differential equation under study, namely, an nth order linear inhomogeneous equation (see 37), by a multiplicative equation (see 38). Solve the latter for $\hat{y}(p)$ by division and revert, possibly with the help of appropriate tables of correspondences, from the symbolic solution $\hat{y}(p)$ to the actual solution $y(x)$. Although highly successful from the pragmatic viewpoint, Heaviside's calculus remained unexplained for several decades. Among other features it entailed the differentiation of non-differentiable functions, such as the so-called Heaviside function $Y(x)$ with a unit step at the origin (see 39).

The need to extend the operation of differentiation beyond its classical limitations is also strongly felt in pure mathematics, particularly in the theory of partial differential equations. One of the fundamental concepts of functional analysis, that of duality and of linear functionals, led, around 1930, to the introduction of weak derivatives. Two functions f and ϕ are considered with real or complex values, defined and continuously differentiable on the real line $\mathfrak{R}$. Furthermore, it is assumed that ϕ vanishes identically outside some finite interval. The classical formula of integration by parts leads to an equality in integrals involving the functions f and ϕ and their derivatives (see 40). It is clear that the right-hand side of this equality makes sense even when f is not differentiable, for instance when f is the Heaviside function (see 39). When $f = Y$ the value of the right-hand-side is $\phi(0)$, whereas the left-hand-side makes no sense at all: if the derivative of Y is to have a reasonable meaning, it must vanish for all $x \neq 0$. But then integration over the set consisting of a single point, $x = 0$, if understood in Lebesgue's sense, can only yield zero, not $\phi(0)$. In order to resolve this apparent paradox and to extend the concept of derivative, the expressions that continue to make sense when applied to nondifferentiable functions (in the classical sense), such as the right-hand side in the integration by parts (see 40), are retained. The latter is viewed as a linear functional in ϕ, which varies in the class C_c^1 of once continuously differentiable functions on the real line, vanishing outside a finite interval. It is precisely this functional that is called the weak derivative of f. Now, if f happens to be continuously differentiable (in the classical sense), its weak derivative will coincide with the integration of $\phi f'$, in other words, with the integration of ϕ with respect to the density $f'(x)dx$, a functional of ϕ that, from now on, shall be identified to the concept of f' itself, the classical derivative of f. As for integration by parts (see 40), it becomes a special identity, valid when f' is continuous or, more generally, locally integrable.

Weak derivatives

(37) $$y^{(n)} + a_1 y^{(n-1)} + \cdots + a_n y = f(x)$$

(38) $$(p^n + a_1 p^{n-1} + \cdots + a_n)\hat{y}(p) = \hat{f}(p)$$

(39) $$Y(x) = 1 \quad \text{for} \quad x > 0, \qquad Y(x) = 0 \quad \text{for} \quad x \leq 0$$

(40) $$\int_{-\infty}^{+\infty} f'(x)\phi(x)\,dx = -\int_{-\infty}^{+\infty} f(x)\phi'(x)\,dx$$

(41) $$\int_{-\infty}^{+\infty} \phi(x)\,\delta(x)\,dx = \int_{-\infty}^{+\infty} \phi(0)\,\delta(x)\,dx = \phi(0)$$

(42) $$\phi(0) = \int_{-\infty}^{+\infty} \phi(x)\delta(x)\,dx = \int_{-\infty}^{+\infty} Y'(x)\phi(x)\,dx$$

(43) $$\int \delta'(x)\phi(x)\,dx = -\int \delta(x)\phi'(x)\,dx = -\phi'(0)$$

It is not possible here to give an account of all the currents and crosscurrents that eventually culminated in the present-day formulation of the theory of generalized functions, from the work of Jacques Hadamard (French) on the Cauchy problem (after the 19th-century French mathematician Augustin-Louis Cauchy), which goes back to World War I, to that of T. Carleman, Arne Beurling, and especially of S. Bochner, just before and after World War II. A contribution particularly relevant to the present discussion, however, is that of the physicists, foremost among them Dirac, in their efforts to create the mathematical tools required by the development of quantum field theory. Dirac introduced in 1925, and systematically used thereafter, the mathematical object that was soon to be called Dirac's function and to be defined as a function $\delta(x)$ equal to 0 for all points $x \neq 0$ and to $+\infty$ at $x = 0$, the integral of which, moreover, is equal to $+1$. Though such a definition is inacceptable, that of the associated functional, which to every continuous function ϕ assigns the number $\phi(0)$ (see 41), is unobjectionable. As a matter of fact, δ turns out to be an object that rigorous mathematicians had been using for quite a while, for instance in the theory of summation of Fourier series, as the limit of sequences of true functions, the limit of which was never explicitly written down. Dirac's δ also entered into Heaviside's symbolic calculus: it is the function having a symbolic transform equal to the constant one. And returning to the earlier consideration of the weak derivative of Heaviside's function, and momentarily adopting the physicist's notation (see 42), it is seen that $Y' = \delta$: the weak derivative of Heaviside's function is Dirac's function or, rather, as it is now called, Dirac's distribution.

Dirac's function

There is no reason not to repeat the procedure of weak differentiation and apply it to δ itself. Again in physicists' notation, an integral identity can be obtained (see 43). In mathematical language this simply means that δ' is the functional which, to every continuously differentiable function ϕ, assigns the number $-\phi'(0)$. If one wishes to iterate this procedure and define the mth derivative of δ for any $m = 1, 2, \cdots$, one must evaluate the relevant functionals on functions ϕ, the degree of differentiability of which accordingly increases. Finally, it is convenient to use, as test-functions, the infinitely differentiable (or C^∞) functions. Thus, $\delta^{(m)}$ will be the functional that to every $\phi \in C^\infty$ assigns the value $(-1)^m \phi^{(m)}(0)$.

Laurent Schwartz's idea (in 1947) was to give a unified interpretation of all the generalized functions that had

infiltrated analysis as (continuous) linear functionals on the space C_c^∞ of infinitely differentiable functions vanishing outside compact sets. He provided a systematic and rigorous description, entirely based on abstract functional analysis and on duality. It is noteworthy that such an approach had a precedent, in the presentation by André Weil (French) of the integration of locally compact groups: there, Radon measures on the group G are introduced and handled as continuous linear functionals on the space $C_c^0(G)$ of the continuous functions in G that vanish outside some compact subset. Because of the demands of differentiability in distribution theory, the spaces of test-functions and their duals are somewhat more complicated. This has led to extensive studies of topological vector spaces beyond the familiar categories of Hilbert and Banach spaces, studies that, in turn, have provided useful new insights in some areas of analysis proper, such as partial differential equations or functions of several complex variables. Schwartz's ideas can be applied to many other spaces of test-functions beside C_c^∞, as he himself and others have shown; numerous examples can be found.

The Schwartz procedure is rather unsatisfactory if the analytic functions (on real space) are to be taken as test-functions, an action desirable in order to increase the pool of generalized functions: because analytic functions are fewer than C^∞ functions, the linear functionals on the space of the former are more numerous than those on C_c^∞. Such an increase of the pool is permitted when the functions intervening in the problem under consideration, for example the coefficients of a partial differential equation, are analytic. It was noted some time ago (in particular, by theoretical physicists) that many important distributions could be regarded as jumps across the real line of holomorphic functions of nonreal z. The simplest example is Dirac's δ expressed in terms of a limit (see 44), which means that the integration of the function under the limit sign, with respect to $\phi(x)dx$ with arbitrary $\phi \in C_c^\infty$, converges to $\phi(0)$. Observations such as these were developed and systematized by M. Sato and others, resulting in the theory of hyperfunctions, which is based not on functional analysis but on cohomology of sheaves and which appears to be the right framework for the treatment of differential problems within the (real) analytic category. The term sheaf involves the notion of functions or distributions defined only locally, while cohomology involves patching local distributions together to create a genuine distribution of the type described above.

DEFINITIONS AND BASIC PROPERTIES OF DISTRIBUTIONS

If generalized functions are to answer the needs of analysis and of applied mathematics, they must comply with a number of preconditions. First of all, they must truly generalize functions, which means not only that the great majority of the functions should be included in the new, expanded pool, but also that there should be a significant enlargement of it so as to include objects already in use that are definitely not functions, such as, for example, the Dirac distribution and its derivatives, the Radon measures, Hadamard's finite parts, and others. Secondly, the basic operations of analysis, such as differentiation, its inverse: the determination of primitives, multiplication (at least by smooth functions), convolution, Fourier transformation, and others, should be extended to sizable classes of generalized functions. This requirement means in particular that when the extensions are restricted back to standard functions, the classical operations should be recovered. For instance, consider distributions on the real line, say, with values in the set T of complex numbers; they are defined as continuous linear functionals on the space C_c^∞ of test-functions (infinitely differentiable functions vanishing outside some closed bounded subset; *i.e.*, some compact subset). It means that a distribution T is a mapping that to every test-function ϕ assigns a complex number $T(\phi)$, which is linear (see 45) and continuous. Continuity has the following meaning: consider any sequence of test-functions, $\{\phi_n\}$, $n = 1, 2, \cdots$, having the two properties: (i) there is a closed bounded set K outside of which all ϕ_n vanish, whatever n; (ii) the functions ϕ_n and each one of their derivatives, $\phi_n^{(\alpha)}$, $\alpha = 1, 2, \cdots$, converge uniformly, to limits $\phi^{(\alpha)}$, $\alpha = 0, 1, 2, \cdots$, respectively. Under such circumstances, the numbers $T(\phi_n)$ must converge to $T(\phi)$. The first thing to be noted is that most functions can be interpreted as distributions. More precisely, every locally integrable function f in the real line may be so interpreted: it defines the distribution T_f by the formula that identifies the map of ϕ through the distribution as an integral of a product (see 46) (continuity in the above sense, and linearity of this functional are checked at once). If f is to be identified with T_f, another requirement must be fulfilled: the functions such as f should not lose their identity in the process. More precisely, if two functions f and g are different, the associated distributions T_f and T_g should also be different. This is not quite true: $T_f = T_g$ is equivalent to the statement that f is equal to g almost everywhere, but the qualification is tolerable in modern analysis.

$$\delta = \lim_{y \to +0} \frac{1}{2\pi i}\left\{\frac{1}{x-iy} - \frac{1}{x+iy}\right\} \tag{44}$$

$$T(a\phi + b\psi) = aT(\phi) + bT(\psi) \tag{45}$$

$$T_f(\phi) = \int \phi(x) f(x)\, dx \tag{46}$$

$$\sum_{n=1}^{+\infty} a_n \delta(x - x_n) \tag{47}$$

$$\left(pv\frac{1}{x}\right)(\phi) = \lim_{\epsilon > 0,\, \epsilon \to 0} \int_{|x| > \epsilon} \phi(x)\frac{dx}{x}, \qquad \phi \in C_c^\infty \tag{48}$$

$$\left(pv\frac{1}{x}\right)(\phi) = \int_{-\infty}^{+\infty} \frac{\phi(x) - \phi(-x)}{2}\frac{dx}{x} = -\int_{-\infty}^{+\infty} \phi'(x) \log|x|\, dx \tag{49}$$

From now on, whenever a function is spoken of as a distribution, it will always mean a locally integrable function, defined almost everywhere.

Next it is observed that there are distributions that are not functions: examples are the Dirac distribution and its derivative. Other examples can be found among the Radon measures: these are the distributions $T(\phi)$ that, in a sense, do not depend on the derivatives of the test-function ϕ: $T(\phi_n)$ converges to $T(\phi)$ if ϕ_n converges uniformly to ϕ (while remaining equal to zero outside one and the same compact set K), regardless of whether their derivatives converge. Examples are infinite series of Dirac distributions (at different points), the distributions having arguments $x - x_n$ and coefficients a_n (see 47) in which either the points x_n tend to $\pm\infty$, in which case the numbers a_n can be chosen arbitrarily, or else infinitely many x_n remain in a bounded interval, in which case the corresponding numbers a_n must form a summable sequence. There are distributions that are not Radon measures, for instance $\delta^{(n)}$ for $n > 0$. Another example is the (Cauchy) principal value $pv\frac{1}{x}$ that can be defined in terms of the limit of an integral (see 48) and depends on the first derivative of the test-function ϕ. It can be defined by a number of different formulas (see 49).

Distributions need not be defined on the entire real line; they can also be defined on open subintervals $]a, b[$ and more generally, any open subset Ω of the N-dimensional Euclidean space $\mathfrak{R}^N$; the space of distributions in Ω is denoted by $\mathfrak{D}'(\Omega)$. Its elements are the continuous linear functionals on the space $C_c^\infty(\Omega)$ of test-functions in Ω that vanish outside some compact subset of Ω, that is to say, some closed bounded subset of $\mathfrak{R}^N$ contained in Ω.

There is an important precondition that distributions

satisfy and that essentially originates with their applications to physics, which can be called localization. Roughly speaking, it means that values may be assigned to distributions if not at every point, at least in every open subset, however small, of their domain of definition; more precisely, distributions may be evaluated on test-functions the support of which is arbitrarily small (the support of a continuous function is the smallest closed set outside which it vanishes identically). Simple examples of such test-functions can be constructed by modifying a basic function expressed in terms of the exponential with argument involving x (see 50). Here $x = (x^1, \cdots, x^N)$ is the variable in $\mathfrak{R}^N$ and the absolute value of x is defined through the extended Pythagorean theorem (see 51). It may be checked that $\rho \in C^\infty$. Given any $\varepsilon > 0$, $\rho_\epsilon(x)$ is set equal to $\varepsilon^{-N}\rho(\varepsilon^{-1}x)$. The support of $\rho\epsilon$ is the closed ball with radius ε centred at the origin. The constant κ is chosen so that the integral of ρ, and therefore also those of the ρ_ϵ, over the whole space $\mathfrak{R}^N$, is equal to $+1$. Given any point x_0, $\rho\epsilon(x - x_0)$ is a C^∞ function with support in the closed ball of radius ε, centred at x_0.

Now letting T denote a distribution in the open subset Ω of $\mathfrak{R}^N$ and O be any open subset of Ω, it may be said that $T = 0$ in O if $T(\phi) = 0$, for whatever $\phi \in C_c^\infty(\Omega)$, the support of which is contained in O; that is, whatever $\phi \in C_c^\infty(O)$. The two distributions T_1, T_2 are said to be equal in O if $T_1 - T_2 = 0$ in O. For such a notion to be really useful, the following property must be valid: if the distribution T vanishes in each one of a collection of open subsets $\{O_\iota\}$ $(\iota \in I)$, it also vanishes, then, in their union, O. This is indeed what happens, because of the existence of partitions of unity in C^∞, subordinate to an arbitrary locally finite covering of O by open sets $U_\alpha(\alpha \in A)$. The phrase locally finite means that every point of O has a neighbourhood that intersects only finitely many sets U_α. The union of the U_α is equal to O. For each index α, a C^∞ function ϕ_α in O can be found the support of which is contained in U_α, $\phi_\alpha \geqslant 0$ and such that $\Sigma\phi_\alpha \equiv 1$,the summation being over α, in O. If it is the case now that ϕ is an arbitrary test-function in O and $\phi = \Sigma\phi_\alpha\phi$, then, because ϕ vanishes outside some compact subset of O, there is only a finite number of functions $\phi_\alpha\phi$ that are not identically zero; therefore, $T(\phi)$ is expressible as a sum (see 52) by linearity.

On the other hand, $T(\phi_\alpha\phi) = 0$ because $T = 0$ in each O_ι and the support of $\phi_\alpha\phi$ is contained in at least one of them. It follows that $T(\phi) = 0$, that is $T = 0$ in O. As a corollary, it is seen that the union of all the open sets in which a distribution T vanishes is itself such a set. Its complement is called the support of T. If T is defined by a function f, its support is equal to the smallest closed set outside of which f vanishes almost everywhere; this is a reasonable notion, which coincides with what is called the support when f is continuous.

This rule, that the extended notion should coincide with the classical one when applied to functions, suggests in many instances how the generalization should be made. Consider the derivative: in distribution notation, the formula integration by parts (see 40) reads $f'(\phi) = -f(\phi')$. This suggests that the derivative of an arbitrary distribution T on the real line be defined as the distribution, denoted T', having a value at ϕ equal to $-T(\phi')$. In the case of N variables (see 53) a multi-index notation is used: $\alpha = \alpha_1, \cdots \alpha_N$ is an N-tuple of non-negative integers $|\,a\,| = \alpha_1 + \alpha_2 + \cdots + \alpha_N$ its length; and an abbreviated notation is used for the partial derivative operator raised to the α-power (see 54).

Now consider the multiplication by a smooth (that is to say, C^∞) function ψ. If f is a locally integrable function, regarded as a distribution, then the product ψf as a mapping is defined in terms of f as a mapping (see 55), for any test-function ϕ. This suggests that the product ψT of ψ be defined with any distribution T by a similar formula (see 56). The only functions that transform any test-function ϕ into another test-function after multiplication are the C^∞ functions. It follows from this that they are the only functions with which arbitrary distributions may be multiplied.

$$(50)\quad \begin{cases} \rho(x) = \kappa \exp\left(-\dfrac{1}{1-|x|^2}\right) & \text{if} \quad |x| \leqslant 1 \\ \rho(x) = 0 \quad \text{if} \quad |x| > 1 \end{cases}$$

$$(51)\quad |x| = \left[\sum_{j=1}^{N} (x^j)^2\right]^{1/2}$$

$$(52)\quad T(\phi) = \sum_\alpha T(\phi_\alpha\phi)$$

$$(53)\quad [(\partial/\partial x)^\alpha T](\phi) = (-1)^{|\alpha|} T[(\partial/\partial x)^\alpha \phi], \qquad \phi \in C_c^\infty$$

$$(54)\quad (\partial/\partial x)^\alpha = (\partial/\partial x^1)^{\alpha_1} \cdots (\partial/\partial x^N)^{\alpha_N}$$

$$(55)\quad (\psi f)(\phi) = f(\psi\phi)$$

$$(56)\quad (\psi T)(\phi) = T(\psi\phi), \qquad \phi \in \mathfrak{C}_c^\infty$$

$$(57)\quad \{a_n\}, \qquad n = 1, 2, \cdots$$

This is not to say that special distributions, about which enough information is available, may not be multiplied by functions that are not so smooth, or even by other distributions. Thus it is proper to multiply Radon measures with continuous functions, also distributions such as δ' or pv/x with other distributions, provided that the latter are equal to a C^1 function in some neighbourhood of zero.

Convergence and limits

Other notions of vital importance to mathematical analysis are those of convergence and limits. In this context the rule cannot simply be followed that the corresponding notion for functions should be extended. Convergence of functions is based on their assigning values to individual points and thus is not easily extended to such objects as Radon measures or to the derivatives of δ. Also, it is customary to require that the limits of convergent sequences of functions be functions, whereas it is desirable to allow such a limit to be a distribution, possibly not a function. There are advantages to a notion of convergence that is loose enough to apply to the totality of distributions and which gives rise to as many convergent sequences as possible, provided each one's limit remains unique. For it might then be possible to go to the limit in the basic operations of analysis, for instance in differentiating, and find that the limit of the derivative is equal to the derivative of the limit, a rare occurrence in classical analysis. This, however, is exactly what happens if a definition is adopted that comes naturally to mind, namely that a sequence of distributions T_n $(n = 1, 2, \cdots)$ converges if, for every test-function ϕ, the numbers $T_n(\phi)$ converge. One should be careful, however, not to extend this definition to "generalized sequences"—*i.e.*, nets or filters—for it would then lead to serious difficulties.

Within the new framework as now defined, there is obtained the distribution derivatives, of any order, of all the locally integrable functions—which achieves one of the main objectives of the generalization. It may be asked whether more and stranger generalized objects have been introduced than were called for. The answer, which is negative (at least locally), is contained in the structure theorem for distributions: any distribution is equal to a finite sum of derivatives of continuous functions, in any open subset of its domain of definition the closure of which is compact. It may be noted that the Dirac distribution δ on the real line is the derivative of Y, Heaviside's function, and that Y is the derivative of $Y(x)x$. Another example is that of $pv(1/x)$, which (see 49) is the derivative of $\log|\,x\,|$, itself the derivative of the continuous function $x(\log|\,x\,| - 1)$. It should be noted, furthermore, that if there is given (see 57) any sequence of nonzero numbers, then the infinite series constructed with delta functions and derivatives of delta functions, with

the members of the sequence as coefficients (see 58), cannot represent a distribution. On the contrary, if the sequence of points x_n (59) converges to $+\infty$ or to $-\infty$, the modified series (see 60) always does represent a distribution.

REGULARIZATION AND CONVOLUTION OF DISTRIBUTIONS

Convolution applies to functions and, as will be seen, to distributions defined in the whole real line or, in the case of several variables, in the whole space $\mathfrak{R}^N$. Like Fourier transformation it is linked to the structure $\mathfrak{R}^N$ as an additive group. The convolution of two Lebesgue integrable functions f, g on the real line is defined by either of two integrals (see 61). Extension of this definition to several variables is straightforward. The function $f * g$ is a Lebesgue integrable function. The convolution can be defined for many classes of functions besides the integrable ones. For instance, f and g need only be locally integrable (*i.e.*, integrable on every finite interval), provided one of the two has a bounded support. In general, as the requirements of decrease at infinity are relaxed on one of the two factors, f or g, these requirements must be strengthened on the other. The same observation applies to the requirements about regularity. For instance, g, or rather $g(y)dy$, may be replaced by a Radon measure $d\mu(y)$, but then it may be necessary to take f to be a continuous function; incidentally, in this case the convolution $f * \mu$ is a continuous function. If f is m-times continuously differentiable, the same is true of $f * \mu$, as can be seen by differentiating under the sign of integration with respect to μ. Going one step further, g may be replaced by an arbitrary distribution T, provided that f is replaced by a test-function ϕ (see 62). The notation T_y means that T acts with respect to the variable y; x should be regarded as a parameter, of which $T * \phi$ is a function. The definition of a distribution as a continuous functional of ϕ implies at once that $T * \phi$ is a continuous function of x on the real line and, in fact, a C^∞ one (in distribution theory, one may always differentiate with respect to a parameter under the "integral" sign).

Thus convolution of distributions with test-functions transforms them into C^∞ functions; that is why it is widely referred to as regularization (or smoothing). The general definition of convolution (see 62) makes sense also in the case in which ϕ does not have compact support, but is any C^∞ function, provided that, then, the support of T be bounded. In all cases, the support of $T\phi$ is contained in the vector sum of the support of T and of that of ϕ, the set of points $x + y$ with x in the support of T and y in that of ϕ.

Using regularization the convolution of two arbitrary distributions S, T may be defined, of which at least one has compact support. The definition is based on the remark that if S and T were associated with integrable functions f, g, their convolution could be defined by a simpler expression obtained by (see 63) setting $\check{g}(t) = g(-t)$. Extending this operation $^\vee$ to distributions by the formula that uses the unmodified form of the distribution (see 64), convolutions of two distributions (see 65) may be defined. If the support of T is bounded, so is the support of $T * \phi$, which is therefore a test-function. If ϕ_n, $n = 1, 2 \cdots$, converges to zero in the sense of test-functions, so do the $\check{T} * \phi_n$. Thus (65) is a sound definition. Some basic properties of convolution are now listed: (i) the support of $S * T$ is contained in the vector sum of the support of S and of that of T; (ii) if S_n $(n = 1, 2, \cdots)$ is a sequence of distributions converging to S and T_n $(n = 1, 2, \cdots)$ a sequence of distributions, all having their support contained in one and the same bounded set K, converging to T, $S_n * T_n$ converges to $S * T$; (iii) in order to differentiate the convolution $S * T$, any one of the factors may be differentiated (see 66); (iv) if δ is the Dirac distribution, $S * \delta = S$.

The functions ρ_ϵ introduced above may now be considered. When $\varepsilon > 0$ goes to zero, ρ_ϵ, regarded as a distribution, converges to δ. Indeed, given any test-function ϕ, ρ may he defined in terms of an integral involving ϕ (see 67). By the properties (ii) and (iv) above it is concluded that, given any distribution T in the whole space,

(58) $$\sum_{n=1}^{+\infty} a_n \delta^{(n)}$$

(59) $$\{x_n\}, \qquad n = 1, 2, \cdots$$

(60) $$\sum_{n=1}^{+\infty} a_n \delta^{(n)}(x - x_n)$$

(61) $$(f * g)(x) = \int_{-\infty}^{+\infty} f(x-y)g(y)\,dy = \int_{-\infty}^{+\infty} f(y)g(x-y)\,dy$$

(62) $$(T * \phi)(x) = T_y\{\phi(x-y)\}$$

(63) $$(f * g)(x) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(y)g(x-y)\phi(x)\,dx\,dy = f(\check{g} * \phi)$$

(64) $$\check{T}(\phi) = T(\check{\phi})$$

(65) $$(S * T)(\phi) = S(\check{T} * \phi), \qquad \phi \in C_c^\infty$$

(66) $$(\partial/\partial x)^\alpha (S * T) = S * (\partial/\partial x)^\alpha T = [(\partial/\partial x)^\alpha S] * T$$

(67) $$\rho_\epsilon(\phi) = \int \phi(\varepsilon x)\,\rho(x)\,dx \to \phi(0)$$

(68) $$(-\infty, x_0), \qquad x_0 < +\infty$$

(69) $$S^{(m)} + a_1 S^{(m-1)} + \cdots + a_m S = A * S$$

$\rho\epsilon * T$ converges to T. Because each $\rho\epsilon * T$ is a C^∞ function, then every distribution is the limit of a sequence of C^∞ functions. From this may be derived that every distribution T in an open set Ω is the limit of a sequence of test-functions in Ω. Indeed, by using a C^∞ partition of unity in Ω a sequence may be formed of test-functions α_j in $\Omega (j = 1, 2, \cdots)$ equal to one on ever increasing open subsets of Ω, the union of which is equal to Ω. It is seen that the distributions $\alpha_j T$, which have compact support contained in Ω, converge to T. For each j, a number $\varepsilon_j > 0$ may be selected so small that the support of $(\alpha_j T) * \rho\epsilon^j$ will be a compact subset of Ω, and therefore $(\alpha_j T) * \rho\epsilon^j$will be a test-function in Ω; and if ϵ_j tends to zero, as $j \to +\infty$, this function will converge to T. This procedure of approximation, which in view of its importance has been described in some detail, is known as cutting and regularizing.

Convolution between two distributions S and T can be defined, in various instances, even when one of them does not have bounded support (see 61). Another important case is convolution between distributions on the real line that vanish in the half-line $x < 0$. (Such distributions are said to belong to $\mathfrak{D}_+'$; there is no simple extension of this concept to several variables.) The reason is that if $S, T \in \mathfrak{D}_+'$ and $\phi \in C_c^\infty$, the support of $\mathfrak{D} * \phi$ is contained in some half-line (see 68) and intersects the support of S along a bounded set, from which it follows that $S(T * \phi)$ appearing in the right-hand side of a formula (see 65) makes sense. It is checked at once that $\mathfrak{D}_+'$ is commutative convolution algebra, having a unit element, the Dirac distribution δ. As a matter of fact, $\mathfrak{D}_+'$ is an integral domain: if $S * T = 0$, then either $S = 0$ or $T = 0$. This follows from the important theorem of supports (valid in any number of variables): if S, T are two nonzero distributions with compact supports and if K, L respectively are the smallest convex sets containing the support of S and the one of T, the smallest convex set containing the support of $S * T$ is exactly equal to the vector sum $K + L$.

There is a close relationship between the convolution algebra $\mathfrak{D}_+'$ and Heaviside's symbolic calculus. Indeed, given any $S \in \mathfrak{D}_+'$, a formal convolution between A and S can be constructed (see 69) in which A is the distribution that is composed of a linear combination of δ-functions

(see 70). This follows from properties (iii) and (iv) of convolution, stated earlier. If it can be shown that the distribution A, obviously belonging to $\mathcal{D}_+'$, has an inverse E in $\mathcal{D}_+'$ (that is, $A * E = \delta$), it will be possible to solve all the differential equations $A * S = T \in \mathcal{D}_+'$: for it will suffice to set $S = E * T$ and, as a matter of fact, S will be the unique solution in $\mathcal{D}_+'$. The inverse $E = A^{-1}$ can now be constructed: letting $U(x)$ be the (unique) analytic function satisfying a linear equation (see 71) for all x in which certain conditions are met at 0 (see 72) and if Y denotes Heaviside's function, then $E = YU$.

LOCALLY CONVEX SPACES AND THEIR DUALS

A deeper understanding of the relations between the various spaces of test-functions and distributions is impossible without the help of abstract functional analysis and the general theory of locally convex spaces. The structure of these is better described by means of seminorms. Consider a complex (or a real) vector space E; a seminorm p on E is a mapping of E (see 73) into the set $\mathfrak{R}_+$ of numbers $\geqslant 0$ that is subadditive, that is, satisfying an inequality (see 73), and positive-homogeneous of degree one (see 74) for all vectors x, y and scalars α. The seminorm p is called a norm if $p(x) = 0$ implies $x = 0$. A topology on E is said to be locally convex if given any point x_0 of E and any neighbourhood U of x_0, there is a continuous seminorm p on E and a number $\varepsilon > 0$ such that U contains the set of points x satisfying $p(x - x_0) < \varepsilon$ (any set U having this property is automatically a neighbourhood of x_0). Convergence in such a topology is not hard to describe: a sequence (or a net) $\{x_\alpha\}$ converges to x if the nonnegative numbers $p(x - x_\alpha)$ converge to zero whatever the continuous seminorm p. The locally convex space E is Hausdorff if, given any two points $x \neq y$, there are neighbourhoods U of x and V of y that do not intersect each other. This is equivalent to saying that, given any vector $x \neq 0$, there is a continuous seminorm p such that $p(x) \neq 0$. A basis of continuous seminorms on E is a collection Q of such seminorms such that every continuous seminorm p satisfies $p \leqslant Cq$ for some $q \in Q$ and some constant $C > 0$. If there is a basis of continuous seminorms on E consisting of a single norm, N, the locally convex space E is said to be normable; the pair (E, N) is called a normed space. If E has a countable basis of continuous seminorms $\{P_n\} (n = 1, 2, \cdots)$ and if, moreover, E is Hausdorff, E is metrizable, which means that its topology can be defined by a metric such as will now be constructed. There is an easy way of constructing a metric $d(x, y)$ defining the topology of E: possibly after replacing each p_n by the upper envelope of the p_j's, $1 \leqslant j \leqslant n$, it may be assumed that $p_n \leqslant p_{n+1}$ for every n; set then the metric equal to an infinite sum involving the infimum of 1 and p_n with coefficients α_n (see 75), where $\alpha_n > 0$ for every n and the sum of the coefficients converges (see 76). A sequence $\{x_n\}$ in a locally convex space is called a Cauchy sequence if, given any continuous seminorm p and any number $\varepsilon > 0$, there is $N_\epsilon > 0$ such that $p(x_m - x_n) < \varepsilon$ whenever $m > n > N_\epsilon$. The space is said to be sequentially complete if every Cauchy sequence converges. It is complete when the space is metrizable; it is then called a Fréchet space. A complete normed space is called a Banach space. One of the fundamental theorems of functional analysis is the open mapping theorem, proved by Banach: if E and F are two Fréchet spaces and if u is a continuous linear map of E onto F, u transforms open sets into open sets (or, equivalently, any neighbourhood of any point into a neighbourhood of its image). An equivalent form of this theorem is the *closed* graph theorem: if E and F are two Fréchet spaces and if u is a linear map of E into F the graph of which (see 77) is closed (in $E \times F$), then u is continuous. These two theorems have been extended to wide classes of spaces besides the Fréchet spaces, but there are simple instances in which they fail to hold.

Examples of Banach spaces

Important examples of Banach spaces are: the space $\ell^\alpha (1 \leqslant \alpha < \infty)$ of sequences of numbers the absolute values of which, raised to the αth power, are summable; the Lebesgue space $\mathfrak{L}^\alpha$ of equivalence classes (for the relation of being equal almost everywhere) of measurable functions the absolute values of which, raised to the αth power, are integrable; the space ℓ^∞ of the bounded sequences; the space $\mathfrak{L}^\infty$ of the equivalence classes of measurable functions that are bounded in the complement of a set of measure zero. All these spaces carry natural norms. Another important example of Banach space is the space $C(K)$ of continuous functions on a compact Hausdorff topological space K, equipped with the norm equal to maximum of the absolute value. Every finite dimensional linear space can be turned into a Banach space: by means of a linear basis and the associated Euclidean norm. It should be noted that on a finite dimensional Hausdorff locally convex space every seminorm is continuous and any two norms, N_1, N_2, are equivalent: $N_1 \leqslant \text{constant } N_2 \leqslant \text{constant } N_1$.

(70) $\delta^{(m)} + a_1\delta^{(m-1)} + \cdots + a_m\delta$

(71) $U^{(m)} + a_1U^{(m-1)} + \cdots + a_mU = 0$

(72) $U^{(j)}(0) = 0 \quad \text{for} \quad j < m-1, \qquad U^{(m-1)}(0) = 1$

(73) $p(x+y) \leqslant p(x) + p(y)$

(74) $p(\alpha x) = |\alpha|\, p(x)$

(75) $d(x, y) = \sum_{n=1}^{+\infty} \alpha_n \inf[1, p_n(x-y)]$

(76) $\sum_{n=1}^{+\infty} \alpha_n < +\infty$

(77) $\mathrm{Gr}\, u = \{(x, y) \in E \times F;\ y = u(x)\}$

(78) $|f|_{K,m} = \max_K \sum_{|\alpha| \leqslant m} |(\partial/\partial x)^\alpha f(x)|$

The spaces of test-functions used in distribution theory are seldom Banach spaces. The simplest of them, $C^\infty(\Omega)$, the elements of which are the C^∞ functions in the open subset Ω of $\mathfrak{R}^N$, is a Fréchet space. When m ranges over the integers 0, 1, 2, $\cdots$, and K over the family of all compact subsets of Ω, the maxima over K of a sum of terms involving f (see 78) form a basis of continuous seminorms on $C^\infty(\Omega)$. As a matter of fact K only may be taken ranging over an exhausting sequence of compact subsets K_j of Ω ($j = 1, 2, \cdots$), which means that the union of the K_j is equal to Ω and that, for each j, K_j is contained in the interior of K_{j+1}. It is not difficult to prove that $C^\infty(\Omega)$ is complete, and thus it is a Fréchet space.

If K is any compact subset of Ω, $C_c^\infty(K)$ denotes the space of C^∞ functions in Ω that vanish outside K; the test-functions proper are the elements of the union of the spaces $C_c^\infty(K)$, $C_c^\infty(\Omega)$. Its locally convex structure is completely described by saying that a seminorm p on $C_c^\infty(\Omega)$ is continuous if, and only if, its restriction to every subspace $C_c^\infty(K)$ is continuous when the latter carries the structure induced by $C^\infty(\Omega)$, that is to say, defined by the seminorms (78). A sequence of functions in $C_c^\infty(\Omega)$, ϕ_n, $n = 1, 2, \cdots$, converges, if it is contained in some subspace $C_c^\infty(K)$ and converges there: this is the notion of convergence that has been used in defining the distributions. The space $C_c^\infty(\Omega)$ is not a Fréchet space; it is a space so-called LF, that is, a strict inductive limit of countably many Fréchet spaces. The countability here comes from the fact that $C_c^\infty(\Omega)$ and its topology may be defined by means of a sequence of subspaces $C_c^\infty(K_j)$, in which the K_j form an exhausting sequence of compact subsets of Ω. Each $C_c^\infty(K_j)$ is a closed subspace of C (Ω), hence a Fréchet space; it has no interior points in $C_c^\infty(\Omega)$; otherwise it could absorb the latter by dilation. Thus

(79) $y' \circ u(x) = y'[u(x)]$

(80) ${}^t u(y') = y' \circ u$

(81) $\begin{cases} {}^t u(y') \neq {}^t u(z'), \text{ for every pair} \\ (y', z') \text{ in } E' \text{ in which } y' \neq z' \end{cases}$

(82) $F = L^\alpha(\Omega), \qquad 1 \leqslant \alpha < +\infty$

(83) $p_S(x') = \sup_{x \in S} |x'(x)|$

(84) $\mathfrak{F}f(y) = \int e^{-2\pi i x \cdot y} f(x)\, dx, \qquad i = \sqrt{-1}$

(85) $dx = dx^1 \cdots dx^N, \qquad x \cdot y = x^1 y^1 + \cdots + x^N y^N$

(86) $(f, g) = \int \bar{f}(x) g(x)\, dx$

(87) $(f, g) = (\mathfrak{F}f, \mathfrak{F}g)$

(88) $f(x) = \int e^{2\pi i x \cdot y}\, \mathfrak{F}f(y)\, dy$

$C_c^\infty(\Omega)$ is the union of a countable family of closed subsets without interior points, in other words, it is not a Baire space, whereas every complete metric space, in particular every Fréchet space, is one.

A linear mapping of the locally convex space E into the scalar field ($\mathfrak{R}$ or $\mathfrak{C}$) is usually called a linear functional (or form). The set of all continuous linear functionals is called the dual of E and most often denoted by E' (or E^*). It is naturally equipped with a vector space structure. If the space E is Hausdorff, there are enough continuous linear functionals on E to separate its points, which means that if $x \neq y$, there is $f \in E'$ such that $f(x) \neq f(y)$. This is a consequence of the fundamental theorem of duality of locally convex spaces, the Hahn-Banach theorem, which essentially states that any linear functional on a vector subspace M of E, continuous for the topology induced by E, can be extended as a continuous linear functional to the whole of E. This result, in one form or another, is of frequent use in the applications of functional analysis. For instance, the proof of the structure theorem for distributions is based on it.

Let E, F be two locally convex spaces (with the same scalars, either real or complex), u a continuous linear mapping of E into F. If y' is any continuous linear functional on F, the composition $y' \circ u$, defined by y' followed by u (see 79), is a continuous linear functional on E, and there results (see 80) a linear map of the dual F' of F onto that of E, E', called the transpose of u and denoted by ${}^t u$. An often used consequence of the Hahn-Banach theorem is that, if E and F are Hausdorff, $u(E)$ is dense in F if, and only if, ${}^t u\colon F' \to E'$ is injective or one-to-one (see 81) applying this to $E = C_c^\infty(\Omega)$ and to F, any Hausdorff locally convex space of functions (or distributions) in Ω, in which $C_c^\infty(\Omega)$ is contained and dense. Take u to be the natural inclusion (or injection) of $C_c^\infty(\Omega)$ into F; its transpose is injective and the dual F' of F may therefore be identified to its image, which is a linear subspace of the dual $\mathfrak{D}'(\Omega)$, of $C_c^\infty(\Omega)$, the space of all distributions in Ω. Thus F' can be regarded as a space of distributions in Ω. If this is applied to a special case (see 82), F' is identified to $L^{\alpha'}(\Omega)$ with $\alpha' = \alpha/(\alpha - 1)$. It may also be applied to $F = C^\infty(\Omega)$ the dual of which can be identified to $\mathscr{E}'(\Omega)$, the space of distributions with compact support in Ω.

The dual E' of a locally convex space E can be equipped, in a natural fashion, with various locally convex topologies, of which the most used are the weak topology and the topology of bounded convergence; the latter is often called the strong dual topology. The weak topology is the topology of pointwise convergence in E or of uniform convergence on the finite subsets of E. The following form a basis of continuous seminorms in it: let $S = \{x_1, \cdots, x_n\}$ be any finite subset of E and set P equal to a supremum extended over S (see 83) in which sup θ is the lowest upper bound of the numbers θ; in the present case, it is the largest of the numbers $|x'(x)|$. The weak topology on E', sometimes denoted by $\sigma(E', E)$, is always Hausdorff. The strong dual topology is the topology of uniform convergence on every bounded subset of E. A subset of a locally convex space E, B, is said to be bounded if every continuous seminorm on E is bounded on B. In a Hausdorff locally convex space E, no neighbourhood of a point can be bounded unless E is normable. In general, bounded sets are much smaller than neighbourhoods. The Lebesgue-Borel-Heine theorem (after the French mathematicians Henri-Léon Lebesgue and Émile Borel, and the German Heinrich Eduard Heine) asserts that, in a finite-dimensional Hausdorff space, every bounded and closed set is compact. This is not true in an infinite-dimensional normed space. It happens, however, to be true in many spaces of test-functions in distribution theory, such as $C_c^\infty(\Omega)$ or $C^\infty(\Omega)$ (it follows from Ascoli's [after the Italian Giulio Ascoli] theorem about equicontinuous sets of functions): this is another important difference between these and the Banach spaces. It has the consequence that, for sequences of distributions, there is no difference between weak and strong convergence. Otherwise it is convenient to equip $\mathfrak{D}'(\Omega)$ with the strong topology, for it then becomes complete (which it is not with the weak topology). The strong topology on the dual E' of a locally convex space E is defined by the following basis of continuous seminorms: $p_B(x') = \sup |x'(x)|$, as B ranges over the collection of all bounded subsets of E; the strong topology is finer than the weak one. Neither $\mathfrak{D}'(\Omega)$, nor $\mathscr{E}'(\Omega)$, the strong dual of $C^\infty(\Omega)$, are metrizable. They have, however, the convenient property that every closed bounded set is compact; and their duals are identifiable to $C_c^\infty(\Omega)$ and $C^\infty(\Omega)$ respectively. In other words, all these spaces are Montel spaces (after the Frenchman Paul-Antoine Montel), and they are reflexive (the dual of the dual is the original space).

FOURIER TRANSFORMATION OF DISTRIBUTIONS

The Fourier transform of an integrable function, f, in the space $\mathfrak{R}^N$ can be conveniently defined as the function composed of the integral of an exponential times f (see 84) in which the differential is N-fold (see 85) and the integration symbol stands for N-tuple integration, to be performed over the entire space. It can be observed that the factor 2π enters in a different way from that of the previous one-dimensional definition (see 11). The difference is simply in scale and does not affect the essential properties. A theorem of Lebesgue states that $\mathfrak{F}f(y)$ is a continuous function in $\mathfrak{R}^N$, tending to zero as $|y| \to +\infty$. An excellent framework for the study of the Fourier transformation is the space L^2 of square-integrable functions (defined almost everywhere); L^2 is equipped with the scalar product defined in terms of integrals (see 86) in which $\bar{f}$ is the complex conjugate of f. The norm in L^2 is $\| f \| = (f, f)^{1/2}$; L^2 is complete, thus it is a Banach space, in fact a Hilbert space, which refers to the fact that its structure is determined by a scalar product (see 86). A formula for $\mathfrak{F}f(y)$ (see 84) defines the Fourier transformation $\mathfrak{F}$ on the elements of L^2 that are integrable, for instance on the continuous functions with bounded support. The latter are dense in L^2. One of the most important theorems of modern analysis, Plancherel's theorem (formulated by the Swiss Michel Plancherel), states that $\mathfrak{F}$ can be extended to the whole of L^2, and becomes thus a Hilbert space isomorphism of L^2 onto itself. There follows the Plancherel-Parseval formula that the inner product of f and g equals the inner product of the transforms of these functions (see 87). The inverse of this isomorphism $\mathfrak{F}$ is $\check{\mathfrak{F}}$, defined by a formula similar to that for $\mathfrak{F}$, except that $-i$ is replaced by i (see 88).

The test-functions in $\mathfrak{R}^N$ are not transformed by the Fourier transform (see 84) into test-functions: the image of C_c^∞ under $\mathfrak{F}$ is a space of analytic functions in $\mathfrak{R}^N$, the dual of which cannot be identified conveniently to a

space of distributions. If it is desired to perform the Fourier transformation on distributions in such a way as to get distributions, it is necessary to restrict the domain of validity of that transformation to distributions that do not grow too fast at infinity. These, often called tempered distributions, are introduced as continuous linear functionals on the space $\mathscr{S}$ of C^∞ functions decaying at infinity. These are the C^∞ functions ϕ such that, given any pair of nonnegative integers m, k, the numbers obtained by summing absolute values of partial derivatives of the function ϕ and taking a supremum operation (see 89) are finite. The functionals involved (see 89) form a basis of continuous seminorms in the natural locally convex topology on $\mathscr{S}$; thus, $\mathscr{S}$ becomes a Fréchet space, which contains C_c^∞ as a dense subspace. Its dual, $\mathscr{S}'$, can therefore be identified to a linear subspace of $\mathscr{D}'$. The structure theorem for tempered distributions states a distribution in $\mathfrak{R}^N$ belongs to $\mathscr{S}'$ if, and only if, it is a finite sum of derivatives of continuous functions that grow at infinity slower than some power of $|x|$. Every function belonging to a Lebesgue space L^α for some α, $1 \leqslant \alpha \leqslant +\infty$, defines a distribution that is tempered, and so do polynomials. The distribution on the real line defined by e^x is not tempered. Now, it is easy to check that the Fourier transformation $\mathfrak{F}$ is a continuous linear map of $\mathscr{S}$ into itself; of course, the same is true of $\bar{\mathfrak{F}}$, the inverse of $\mathfrak{F}$ in L^2. From which it follows that $\mathfrak{F}$ is an isomorphism of $\mathscr{S}$ onto $\mathscr{S}$. Its transpose is an isomorphism of $\mathscr{S}'$ onto $\mathscr{S}'$, which is defined to be the Fourier transformation of tempered distributions. If $T \in \mathscr{S}'$ and $\phi \in \mathscr{S}$, it follows that $\mathfrak{F}T(\phi) = T(\mathfrak{F}\phi)$. When T is a square-integrable function, this agrees with the Plancherel-Parseval formula.

$$|\phi|_{m,k} = \sup_{x \in \mathfrak{R}^N} (1 + |x|)^k \sum_{|\alpha| \leqslant m} |(\partial/\partial x)^\alpha \phi(x)| \tag{89}$$

$$\mathfrak{F}(S * T) = \mathfrak{F}S\mathfrak{F}T \tag{90}$$

$$\mathfrak{F}T(y) = T_x(e^{-2i\pi x \cdot y}) \tag{91}$$

The main reason, perhaps, for the spectacular success of Fourier transformation is that, roughly speaking, it transforms convolution into (ordinary) multiplication and vice-versa. For tempered distributions this can be made precise as follows. It is known, that, when T is an arbitrary distribution, it may be convolved with any distribution S the support of which is compact. When it is known that T grows slowly at infinity, that is, T is tempered, however, the constraints on S may be relaxed with only the requirement that, in a suitable sense, it decays rapidly at infinity. This, in turn, means that the Fourier transform of S, a priori a tempered distribution, is in fact a C^∞ function slowly growing at infinity; *i.e.*, each one of the derivatives of this function grow at infinity slower than some power of $|x|$. Such functions are precisely the multipliers on the tempered distributions: namely, if a C^∞ function α is such that αT is tempered for every tempered distribution T, α must be a C^∞ function slowing growing at infinity. Finally, if S is rapidly decaying at infinity and if T is tempered, then the transform of the convolution is the product of the transforms (see 90).

Every distribution with compact support is rapidly decaying at infinity. In this connection, there is the important Paley-Wiener-Schwartz theorem: a tempered distribution T has compact support if, and only if, its Fourier transform can be extended to the complex space C^N as an entire (that is, everywhere analytic) function of exponential type of z (this means that it grows, for $|z| \to +\infty$, slower than some exponential $\exp[C|z|]$). It is to be noted that, in the case of a distribution T with compact support, the transform of the distribution is expressible in terms of the distribution itself (see 91). The Fourier transform of the Dirac measure δ is the constant function one. That of $(\partial/\partial x^j)\delta$ is the linear function $2\pi i y^j$ $(1 \leqslant j \leqslant N)$; and more (generally, if $P(z) = P(z^1, \cdots, z^N)$ is a polynomial in N variables with complex coefficients, the Fourier transform of $P(\partial/\partial x)\delta$ is $P(2\pi i y)$. Consequently it is seen that the Fourier transform of $P(\partial/\partial x)T$ is equal to the product $P(2\pi i y)T$. This fact makes it possible to transform the linear partial differential equations with constant coefficients within $\mathscr{S}'$ into multiplicative equations, and reduces their solution to division problems, a standard aim of any symbolic calculus. (F.Tr.)

BIBLIOGRAPHY. A comprehensive account of a large part of linear functional analysis will be found in NELSON DUNFORD and JACOB T. SCHWARTZ, *Linear Operators*, 3 vol. (1958–71). The first volume covers the general theory of linear operators and the second the theory of self-adjoint operators in Hilbert space. This work includes a bibliography of several thousand items, and gives historical accounts of many of the topics that it discusses. A systematic study of semigroups of linear operators, together with more general background material on functional analysis, is contained in EINER HILLE and RALPH S. PHILLIPS, *Functional Analysis and Semi-groups*, rev. ed. (1957). For an account of the application of Hilbert space spectral theory to the analysis of differential equations, see M.A. NAIMARK (NAYMARK), *Linear Differential Equations* (1967; orig. pub. in Russian, 1954); and ch. 13 in vol. 2 of Dunford and Schwartz (*op. cit.*). For further developments in the application of functional analysis to differential equations see TOSIO KATO, *Perturbation Theory for Linear Operators* (1966); and PETER D. LAX and RALPH S. PHILLIPS, *Scattering Theory* (1967). The spectral analysis of non-self-adjoint operators is reviewed in vol. 3 of Dunford and Schwartz (*op. cit.*); and in NELSON DUNFORD, "A Survey of the Theory of Spectral Operators," *Bull. Am. Math. Soc.*, 64:217–274 (1958); ION COLOJOARA and CIPRIAN FOIAS, *Theory of Generalized Spectral Operators* (1968), contains additional material concerning this developing area. PATRICK BILLINGSLEY, *Ergodic Theory and Information* (1965), gives an elegant short account of many of the main theorems of ergodic theory; for work since 1965, some of it quite important, see WILLIAM PARRY, *Entropy and Generators in Ergodic Theory* (1969). Harmonic analysis is treated in many works; in particular, see EDWIN HEWITT and KENNETH A. ROSS, *Abstract Harmonic Analysis*, 2 vol. (1963–70), which gives a useful picture of this area. GERALD M. LEIBOWITZ, *Lectures on Complex Function Algebras* (1970), gives an account of some of the deeper issues in classical harmonic analysis. The generalized Fourier integral theory central to harmonic analysis is closely related to some of the principal theorems of group representation theory and of functional algebra. MARTIN BURROW, *Representation Theory of Finite Groups* (1965), contains an account of the representation theory of finite groups. For the generalization to compact continuous groups, see HERMANN WEYL, *The Classical Groups: Their Invariants and Representations*, 2nd ed. (1946, reprinted 1961). The representation theory of noncompact continuous groups, which draws much more heavily upon the most advanced techniques of functional analysis, is treated in I.M. GELFAND and M.A. NAIMARK, *Unitary Representations of Semisimple Lie Groups* (American Mathematical Society Translation, vol. 64, 1967). The final chapter of SIGUROUR HELGASON, *Differential Geometry and Symmetric Spaces* (1962), contains additional information concerning this field, which is an active frontier of current research. M.A. NAIMARK, *Normed Rings*, rev. ed. (1970; orig. pub. in Russian, 1956), covers many results of functional algebra. A more specialized discussion, devoted to the theory of von Neumann or W–* algebras, can be found in JACOB T. SCHWARTZ, *W–* Algebras* (1967). For a general account of the use of global topological methods in functional analysis, see the same author's *Nonlinear Functional Analysis* (1969). The reader interested in the present state of this rapidly developing field will have to consult the current journal literature; the survey article by JAMES EELLS, JR., "A Setting for Global Analysis," *Bull. Am. Math. Soc.*, 72:751–807 (1966), gives a good introduction.

There are many works available on the calculus of variations. For an introduction to the theory for problems involving arcs only, see GILBERT A. BLISS, *Calculus of Variations* (1925, reprinted 1962) and *Lectures on the Calculus of Variations* (1946, reprinted 1961). For problems involving multiple integrals, see PAUL FUNK, *Variationsrechnung und ihre Anwendung in Physik und Technik*, 2nd ed. (1970); and CHARLES B. MORREY, JR., *Multiple Integrals in the Calculus of Variations* (1966). For expositions of Morse's theory of critical points, see MARSTON MORSE, *The Calculus of Variations in the Large* (1934); and HERBERT SEIFERT and W. THRELFALL, *Variationsrechnung im Grossen (Theorie von Marston Morse)* (1938).

The standard text on the theory of distributions is LAURENT SCHWARTZ, *Théorie des distributions*, 3rd ed. (1966). STEFAN BANACH, *Théorie des opérations linéaires*, 2nd ed. (1963), is the classic treatise on Banach and Fréchet spaces and is still very inspired reading. ANDRE WEIL, *L'Intégration dans les groupes topologiques et ses applications*, 2nd ed. (1965), is a classic text on the subject of harmonic analysis on locally compact groups. The most used text on generalized functions is I.M. GELFAND *et al.*, *Generalized Functions*, 5 vol. (1964–68; orig. pub. in Russian, 1958–62), a very clearly written work, with numerous applications to partial differential equations, mathematical physics, and harmonic analysis.

The reader interested in the general theory of locally convex spaces may consult NICOLAS BOURBAKI, *Espaces vectoriels topologiques*, 2 vol. (1953–55), the first of all the treatises on the general theory of topological vector spaces, very systematic and exhaustive. An alternative is ALEX P. and WENDY J. ROBERTSON, *Topological Vector Spaces* (1964), which is very clear and very short. FRANCOIS TREVES, *Topological Vector Spaces, Distributions and Kernels* (1967), is another reference on distributions. HANS BREMERMANN, *Distributions, Complex Variables and Fourier Transforms* (1965), is a short and well-written book, with a slant toward applications to physics and to electrical circuits, emphasizing the boundary values of analytic functions. KOSAKA YOSHIDA, *Functional Analysis* (1965), is a wide-ranging presentation of functional analysis, from the theory of linear operators in Hilbert spaces to Banach algebras, ergodic theorems, and diffusion equations.

(J.T.Sc./C.B.My./F.Tr.)

Analysis, Real

Analysis is one of the main divisions of mathematics, the others being history and foundations, algebra, combinatorics and number theory, geometry, topology, and applied mathematics. In extent, analysis is the largest of these, comprising subdivisions that are nearly autonomous and easier to describe than the division as a whole. This article is concerned primarily with a subdivision of analysis that is dependent upon or suggested by the properties of the real-number system; *i.e.*, the system that includes the natural numbers (0, 1, 2, · · ·) and all numbers that may be derived from these by taking ratios, multiplying by plus or minus 1, and applying to such ratios a basic operation of analysis called the passage to the limit of a sequence.

The article structure is as follows:

I. Origins and concepts of analysis
  Meaning of real analysis
    Relation to science
    The concept of the derivative
    The concept of real numbers
    The concept of magnitude
  Stages of development
    Irrational numbers
    Areas and volumes
    Set theory
    Analytic geometry
    Developments in the 19th and 20th centuries
II. Number
  The real-number system
    Number as quantity
    Multiplication; number as ratio
    Rational numbers
    Irrational numbers
  Generalizations of real numbers
    Complex-number system
    Quaternions and hypercomplex numbers
    Algebraic and transcendental numbers
    Modular numbers: fields
    Infinite cardinal numbers
    Infinite ordinal numbers
    Ordered fields
III. Functions and differential calculus
  Functions
    Higher derivatives
    General functions
    Power series
    Piecewise analytic functions
    Fourier series
    Analytic theory of numbers
    Least-square approximation
  Differential calculus
    The derivative
    Maxima and minima
    Differentiation formulas; computation of derivatives
IV. Measure and integral calculus
  Measure
    The empirical period
    The Greek period
    The period of indivisibles
    The calculus period
    The analysis period
    The extension period
  Integral calculus
    The definite integral
    Integrable and nonintegrable functions
    Fundamental theorem of calculus
    Conditions for existence of the Riemann integral
    Improper integrals
    Applications

I. Origins and concepts of analysis

MEANING OF REAL ANALYSIS

Relation to science. Greek mathematics was largely geometrical. It initiated some durable topics of analysis, but the organized creation of analysis began only around 1600. Analysis then developed in stages in intimacy with mechanics and theoretical physics. Whenever work in basic physics instigated a topic in mathematics spontaneously, however, it was largely in analysis. Differential and integral calculus, ordinary differential equations, and the calculus of variations have thus arisen from mechanics; so-called Fourier series from acoustics and thermodynamics; complex analysis from optics, hydrodynamics, and electricity; and partial differential equations from elasticity, hydrodynamics, and electrodynamics. Even mathematical probability, which is classified under analysis, although born from problems of gambling and human chance, drew much of its syllogistic strength in the 19th century from statistical theories of mechanics and thermodynamics.

The concept of the derivative. Geometry deals with spaces and configurations; topology with spatial deformations; algebra with the general nature of the basic operations of addition, subtraction, multiplication, and division; and arithmetic with additive and multiplicative properties of general integers—that is, natural numbers and negative forms of these. Analysis also deals with specific operations; namely, with differentiation and integration. It is more appropriate to say, however, that it deals with the mathematical infinite in many of its aspects, such as those of an infinite multitude, the infinitely large, infinitely small, infinitely near, and infinitely subdivisible. Its first objects and concepts, introductory and basic, are infinite sequence; infinite series; a function such as $y = f(x)$; continuity of a function, which is a property of a function in an infinitely small neighbourhood of a point; derivative of a function such as $y' = df/dx$; and integral of a function, both of which can be defined in terms of infinite limits. The derivative is a master concept, one of the most creative in analysis as well as in all of human cognition. Without it there would be no proper expression of such fundamental properties in physics as velocity, acceleration, density, gradient, and potential. The wave equation would not exist, and, indeed, physics as a discipline would be vastly different if the concept were unavailable. The formal textbook definition of the derivative took more than 150 years to evolve, and it is rewarding to begin with Newton's own description of it under the name of "ultimate ratio":

Concern of analysis for the infinite

> For those ultimate ratios with which quantities vanish are not truly the ratios of ultimate quantities, but [they are] limits towards which the ratios of quantities, decreasing without a limit, do always converge: and to which they approach nearer than by any given difference, but never go beyond, nor in effect attain to, until the quantities have diminished *in infinitum*.

Newton's generation was able to listen to a recondite mathematical definition, all in words, without symbols; and the Marquise du Châtelet, woman of the world, could undertake to translate Newton's formidable *Principia* from Latin into French, ultimate ratios and all.

The concept of real numbers. Mathematical formulation of scientific statements imparts a peculiar kind of lucidity and precision to such statements, and it suggests and establishes logical and cognitive relationships. It also introduces challenging analogies and unifications.

Most phenomena of wave propagation, for instance, whether in acoustics, hydrodynamics, electricity, or optics, are assumed to be governed at the outset by virtually the same set of differential equations. For a while, therefore, the theories run parallel, until the actual physical differences begin to be accounted for.

Measurement and real numbers

The most basic trait of unification is the fact that ultimately all specific and tangible information in science is quantitative—it may be expressed in ordinary real numbers. Every physical magnitude, whatever its own unit of dimension (simple or composite) may be, is measured by one or perhaps several real numbers, and all these numbers come from the same joint supply. Without this fact, there would be no conservation laws in physics; the law of the conservation of energy, for example, could not be stated or even conceived. Space extensions, time intervals, noises, colours, tactile impressions, and emotions—all these things are in the end valued quantitatively. The "meaning in depth" of formulas, equalities, and inequalities in science comes to life by accompanying interpretations. But the transmittable specific information is all in numbers, whether on graphs or on punched cards or on tape.

$$(1)\qquad 1+\frac{4}{10}+\frac{1}{100}+\frac{4}{1{,}000}+\frac{2}{10{,}000}+\cdots$$

$$(2)\qquad \frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+\cdots+\frac{1}{2^n}+\cdots$$

This all-pervading arithmetization in physics and mathematics was begun systematically in the 17th century, and in mathematics its manifest outcome was first of all the analytic, or coordinate, geometry. Mathematics is less insistent than physics on its exclusiveness. To the contrary, 20th-century mathematics abounds and delights in ever introducing and rarely discarding concepts and objects that need not be numerical. It somehow frequently comes back to numbers, however, and the effect of the 17th-century arithmetization of space on the course of mathematics cannot be overstated.

The concept of magnitude. The Greeks never introduced the system of real numbers as the ultimate quantitative measure for geometry, and in science (excepting astronomy) only a few of their results became unmistakably quantitative at all. Instead, in mathematics, they tried to operate directly with the concept of a general "magnitude" that somehow was always geometrical. In the fifth book of Euclid's *Elements* the technical term for "magnitude" is the noun *megethos*. In Homer, this noun still means personal greatness or stature (of a hero, say); in the French noun *grandeur* and the German noun *Grösse*, the two meanings, personal greatness and mathematical magnitude, similarly reside simultaneously. Twentieth-century mathematics has "magnitudes" on various levels of generality, but it has them in addition to and not instead of real numbers.

STAGES OF DEVELOPMENT

Irrational numbers. In the 5th and 4th centuries BC the Greeks gave the first known satisfactory solution to a problem in analysis. In a square the side of which has length equal to 1, the diagonal has a length equal to $\sqrt{2}$. The decimal expansion of $\sqrt{2}$ is a nonterminating quantity, $1.4142\cdots$, so that the value $\sqrt{2}$ is determined as the sum of an infinite series (see Box, sum 1), each term of which is an integer times a (nonpositive) power of 10. The Greeks proved that $\sqrt{2}$ is not a rational number; that is, not a quotient p/q of any integers p and q. But they apparently found for $\sqrt{2}$ an approximate representation by certain rational numbers, or rather they found for the diagonal of the square approximations by line segments that are commensurate with the side of the square. The approximate numbers were later called the partial sums of the continued fraction for $\sqrt{2}$, and apparently from the study of this problem there evolved the so-called Euclidean algorithm, a chain of operations (indicated by equations) for finding the largest integer that divides two given integers.

The most profound achievement in this entire context was the development of a general theory of irrational numbers, or of incommensurate ratios of certain types of "magnitudes" and of proportions between ratios. This achievement was the work of Eudoxus of Cnidus of the 4th century BC but was recorded by Euclid.

Areas and volumes. The Greeks also investigated the number $\pi = 3.14159\cdots$, which is the area of a circle the radius of which is 1, and the problem of squaring the circle that was widely known to the Athenian public. Computations of other areas and volumes and of centres of masses are known from the works of Archimedes, and these results taken cumulatively can be viewed as a precursor to the 17th-century theory of integration. But these "direct" Greek calculations of many particular areas and volumes by circumscribed and inscribed approximating figures, when compared with the procedures of the integral calculus proper, are rather awkward and tedious.

Set theory. The Greeks were also puzzled by the fact that the ordinary number 1 is the sum of the geometric series (see 2) composed of the numbers ½, the square of ½, the cube of ½, and so forth, and this was for them part of a range of problems relating to the concepts "infinitely small" and "infinitely large." The preoccupation with such problems is reflected in the paradox of Achilles and the Turtle put forward by Zeno of Elea and in the distinction between potential and actual infinite as proposed by Aristotle. Partly to refute this distinction, the 19th-century mathematician Georg Cantor (*q.v.*) developed his set theory (*q.v.*), or theory of aggregates, the beginnings of which go back even to Galileo in 16th- and 17th-century Italy. This theory not only strengthened the logical framework of analysis but soon became a basic innovation in all of mathematics and in all the speculations about its essence and its general background.

Analytic geometry. The Greek achievements in analysis were but a prelude. A notable step was taken in the 14th century by the French bishop Nicole d'Oresme, who discovered the logical equivalence between tabulation and graphing, more or less, and, in a way, proposed the use of a graph for plotting a variable magnitude the value of which depends on that of another.

The work of Descartes

The systematic theoretical basis for this possibility, however, evolved only later from the work of a leading French philosopher and mathematician, René Descartes, who in 1637 laid the foundation for analytic geometry—that is, a geometry in which everything is reduced to numbers. In this geometry a point is defined as a set of numbers called its coordinates. A figure is viewed as an aggregate of points, but it is usually described by formulas, equations, and inequalities. This approach to geometry is the basis for the maps, graphs, and charts that have penetrated into all walks of life, and for the mathematical concept of a function or rule of correspondence identifying with each value of x at most one value of y, which, first intended only for analysis, eventually penetrated into all corners of mathematics and into all areas of rational thinking as well.

Later in the 17th century the culminating event took place simultaneously in England and Germany when Newton and Leibniz introduced derivatives and laid the foundation for calculus and mechanics. Archimedes had been groping for the concept of a derivative but hit upon it only dimly in his book on spirals. Archimedes' methods, comparable to those of the 16th and 17th centuries, were in fact analytical in the modern sense of the term. The 17th and 18th centuries produced much mathematics in all directions, but analysis was the underlying theme.

Developments in the 19th and 20th centuries. In the 19th century, topology, algebra, and arithmetic were emerging as independent divisions of mathematics, rivalling analysis in attention; but analysis continued to proliferate and, by sheer bulk, to predominate. Many French

treatises on general mathematics simply bore the title *Cours d'analyse* ("Course in Analysis"). But by a peculiarity of usage, in memory of the fact that Euclid's *Elements* was always being viewed as a book on geometry, in these same treatises a mathematician in general was frequently referred to as *un géomètre*, especially when he was well-known.

Only in the 20th century did topology and algebra begin to achieve a status approximating that of analysis. At the same time the lines of demarcation between the main divisions of mathematics were becoming more flexible and uncertain. After the 1930s, on the one hand analysis penetrated into geometry and arithmetic further than ever, and on the other hand topology and algebra began to fit tighter and more streamlined structural schemes on the large body of analysis. Duplications and redundancies could be reduced by syllogistic identifications and assimilations, and in some areas an increase of the substance offered itself naturally. Furthermore, digital computing machines injected into analysis, if only marginally, a certain algebraization. The structure of the machines and the problem of coding both involve algebra of logic, and in computational assignments from analysis, limiting processes must be replaced by approximations within certain maximal numbers of steps.

II. Number

THE REAL-NUMBER SYSTEM

Number as quantity. Mankind was early led to the concept of a real number by the need for symbols to represent such geometric quantities as lengths, areas, and volumes and such physical quantities as weight and (more recently) electric charge.

Characteristics of geometric quantities

The characteristic features of such quantities are that (1) two quantities of the same kind can be added by some obvious geometrical or physical operation (such as laying two segments next to each other on a straight line), and (2) any quantity can be divided into parts. The second feature, of infinite divisibility, is the crudest way of expressing the principle that the real-number system is continuous and not discrete like the system of positive integers.

The laws of addition are frequently intuitively obvious from the physical definition of this operation. In the case of lengths, it is obvious that $a + (b + c) = (a + b) + c$, because both represent the sum of three segments of lengths a, b, and c in order on a line (see Figure 1). It is also obvious that $a + b = b + a$, because the segment $a + b$ can be transformed into the segment $b + a$ by rotating the segment through 180°.

In representing geometrical or physical quantities by numbers, it is usually necessary to choose first an arbitrary unit, such as a foot, a square inch, or a pound. This unit of quantity is assigned arbitrarily the numerical value 1, and exact multiples of this quantity (*i.e.*, sums of an integral number of unit quantities) are assigned corresponding integral values. Because the entire addition table for positive integers can be constructed from the law for positive integers, which states that the sum of an integer a and an integer $(b + 1)$ is the same as the sum of the integers $(a + b)$ and 1, the sum of two such quantities must be represented by the sum of the corresponding positive integers.

Multiplication; number as ratio. A simple definition of the product of two quantities is not always possible. If the product of two lengths a and b is defined as the area of the rectangle with sides a and b, it may logically be objected that this is a quantity of a different kind.

To get around this logical difficulty, the Greeks suggested that a number should be considered as representing the ratio between a certain quantity and the unit quantity and never as representing the quantity itself. Thus, 160 pounds is a weight; the number 160 is the ratio of this weight to the weight of a pound.

Numbers as dimensionless ratios

This idea that numbers are dimensionless ratios has wide applicability: it leads to plausible "proofs" of the laws of multiplication. For example, the hypothesis that $a \times (b \times c) = (a \times b) \times c$ for all real numbers is given substantial support by the fact that both quantities may

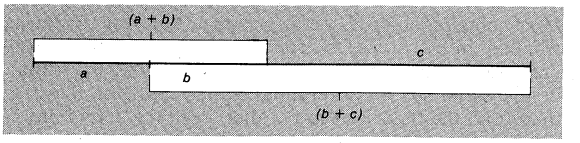


Figure 1: *The associative law of addition.* The length $a + (b + c)$ is the same as $(a + b) + c$ because both represent the sum of three segments of lengths a, b, c in order on a line.

be regarded as representing the volume of a box of sides a, b, and c. These laws $a \times b = b \times a$ and $a \times (b + c) = (a \times b) + (a \times c)$ can be given similar plausible geometric interpretations in terms of areas. But these plausible arguments are not proofs in any rigorous sense. They depend not only on postulates for geometry but also on definitions of area and volume.

The only known way of getting rid of these difficulties is to define real numbers abstractly in terms of the system of positive integers, deducing all properties of real numbers by pure logic from these definitions and properties of the positive integers and using geometrical and physical concepts only to suggest possible postulates and definitions. In this way the properties of real numbers may be made to depend on pure logic and the concept of a class alone (see MATHEMATICS, FOUNDATIONS OF). This abstract procedure will be followed henceforth.

Laws of arithmetic

Rational numbers. If it be assumed that division (except by zero) and subtraction are always possible, the system of positive integers then leads to a system of rational numbers (positive and negative fractions and integers, and zero). For the fundamental laws of arithmetic, see ARITHMETIC. The assumption that every equation $by = a$ in which a and b are positive integral coefficients has a solution corresponds to the idea that a quantity, a, can be divided into any positive whole number, b, of equal parts. The fundamental laws of arithmetic and the cancellation law for multiplication that states that if $a \times m = a \times n$ then $m = n$ can all be preserved.

By the cancellation law for multiplication, $by = a$ can have only one solution, which is written a/b. It may be shown that the usual rules for adding and multiplying fractions must hold. For example, if $by = a$ and $dz = c$, then $bdy = ad$, $bdz = bc$; and so $bd(y + z) = bdy + bdz = ad + bc$, which proves the rule for adding fractions having different denominators. It can also be proved that division by fractions as well as by integers is possible in the new system.

Conversely, it can be proved that the rules for adding and multiplying fractions do give a system in which the laws of arithmetic are valid, as well as the cancellation law for multiplication and certain other basic laws.

A more thorough study would reveal that a considerable reduction in the number of postulates (fundamental laws needed to imply the others) is possible. All the laws for positive fractions thus can be deduced from the associative law for addition; the distributive laws, $a \times (b + c) = (a \times b) + (a \times c)$ and $(a + b) \times c = (a \times c) + (b \times c)$; and the unit laws, $a \times 1 = 1 \times a = a$ and $1 + 1 \neq 1$.

Irrational numbers. Fractions were used as early as 1700 BC by the ancient Egyptians, but, as mentioned above, it was not until Pythagoras (530 BC) that the need for other numbers, such as $\sqrt{2}$, was discovered. The need for such irrational numbers is corroborated in modern mathematical analysis, in which they play a fundamental role in the integral calculus, trigonometry, and other fields. Pythagoras showed that the ratio x of the diagonal of an isosceles right triangle to the length of a side must satisfy the equation $x^2 = 2$ (Pythagorean theorem). No fraction m/n, however, can satisfy $(m/n)^2 = 2$; that is, $m^2 = 2n^2$ has no solution among the integers, because 2 divides m^2 an even number of times, and it divides $2n^2$ an odd number of times.

Eudoxus of Cnidus pointed out in about 367 BC that although $\sqrt{2}$ could not be represented exactly by any one fraction, it could be represented as a limit of a sequence of fractions. Thus, $\sqrt{2}$ can be represented in the form of

an infinite decimal: $\sqrt{2} = 1.4142\cdots$; this amounts to specifying $\sqrt{2}$ as the limit of the sequence of decimal fractions.

Cantor's definition of numbers

The clear exposition of these ideas from the modern point of view is due to Georg Cantor (1871). Real numbers, including both rational and irrational numbers, are defined by Cantor as infinite sequences, $x = (x_1, x_2, x_3, \cdots)$, $y = (y_1, y_2, y_3, \cdots)$, $\cdots$, of fractions $x_n, y_n, \cdots$, that converge in the sense that $(x_m - x_n)$, $(y_m - y_n)$, $\cdots$, approach zero as m and n increase indefinitely. Here x is regarded as the limit of the sequence $(x_1, x_2, x_3, \cdots)$.

Equality is defined by making $x = y$ mean that $(x_n - y_n)$ approaches zero as n increases indefinitely. Addition and multiplication are defined by adding and multiplying corresponding terms (see 3). Laws valid for rational numbers can be extended to all real numbers by the principle of continuity. Each approximating term $(x + y)_n = x_n + y_n$ of $x + y$, for example, is equal to the corresponding approximation $(y + x)_n = y_n + x_n$ of $y + x$; hence, $(x + y)_n - (y + x)_n = 0$ for all n, and $x + y = y + x$ by definition of equality. In general, the principle of continuity holds that laws involving continuously varying functions like $x + y$ and $x \times y$ that are valid for arbitrarily good approximations $x_n, y_n, \cdots$ of $x, y, \cdots$ must be valid also for the limit values $x, y, \cdots$.

The Dedekind cut

Another interesting definition of real numbers was by the 19th-century German mathematician Richard Dedekind: by a section in the class R of fractions is meant a division of all fractions into two classes L and U, such that $x \leqslant y$ for every x in L and every y in U. Each fraction (rational number) determines a section: L consists of the $x \leqslant a$ and U of the $y \geqslant a$. The other sections define the irrational numbers; thus, the section dividing the fractions into the $x < \sqrt{2}$ and the $y > \sqrt{2}$ (more precisely, the positive y with $y^2 > 2$) may be regarded as defining $\sqrt{2}$. Dedekind's definition is in fact equivalent to Cantor's.

The real numbers defined by this process have so far been found adequate for the mathematical treatment of most geometrical and physical quantities such as length, area, weight, electric charge, and the like. Not only are the fundamental laws of arithmetic, the cancellation laws, and the order properties true, but division (except by zero) and subtraction are always possible. Finally, there is the property of any increasing sequence $x_1 < x_2 < x_3 \cdots$ the terms of which are all bounded above by a fixed constant c tending to a limit.

Other types of numbers remain to de discussed; in every case, some of the properties of real numbers will be lost.

GENERALIZATIONS OF REAL NUMBERS

Complex-number system. The equation $x^2 = -1$ can have no real solution because the square of any real quantity is either positive or zero. But $i = \sqrt{-1}$ can be introduced as an imaginary number, and those laws of addition and multiplication that do not involve the relation $a \geqslant b$ may thus be preserved. To do this, all combinations $a + bi = a + b \times \sqrt{-1}$ must be introduced; these are the so-called complex numbers. If addition and multiplication of complex numbers are defined (see 4) appropriately, the basic laws familiar from the properties of real numbers will be satisfied. It can be verified by substitution, for example, that the equation $(c + di)z = (a + bi)$ has a solution (see 5). Furthermore, any quadratic equation $Ax^2 + Bx + C = 0$ with real coefficients has complex roots (see 6), because even if $B^2 - 4AC = -D$ is negative, the roots (see 7) can be found as complex numbers.

This is a special case of a fundamental theorem of algebra that asserts that any polynomial equation $x^n + a_1x^{n-1} + \cdots + a_n = 0$ with real or complex coefficients has a complex root. From a strictly algebraic standpoint, no further generalization is called for.

Although the theory of algebraic equations is greatly simplified by the use of complex numbers, it is necessary to sacrifice the properties of order. That is, if $z_1, z_2, \cdots, z_n$ are a collection of n distinct complex numbers, it is generally not possible to order these such that, say, $z_1 < z_2 < \cdots < z_n$. This is in contrast to the corresponding case of a collection of n distinct real numbers for which such an ordering is always possible.

$$(3)\quad \begin{cases} x + y = (x_1 + y_1, x_2 + y_2, x_3 + y_3, \cdots) \\ x \times y = (x_1 \times y_1, x_2 \times y_2, x_3 \times y_3, \cdots) \end{cases}$$

$$(4)\quad \begin{cases} (a + bi) + (c + di) = a + c + bi + di \\ \qquad\qquad = (a + c) + (b + d)i \\ (a + bi) \times (c + di) = ac + (ad + bc)i + bdi^2 \\ \qquad\qquad = (ac - bd) + (ad + bc)i \end{cases}$$

$$(5)\quad z = \frac{a + bi}{c + di} = \frac{ac + bd}{c^2 + d^2} + \frac{bc - ad}{c^2 + d^2}\,i$$

$$(6)\quad \frac{(-B \pm \sqrt{B^2 - 4AC})}{2A}$$

$$(7)\quad \frac{(-B \pm i\sqrt{D})}{2A}$$

Complex numbers are useful not only in pure mathematics (theory of equations and function theory); they are also used in describing alternating electric currents, and they greatly simplify the solution of many problems throughout the physical sciences.

Addition and multiplication

Quaternions and hypercomplex numbers. If the properties of order can be sacrificed as well as the commutative law of multiplication, then an interesting further extension of the complex-number system can be made to the so-called quaternion; *i.e.*, to numbers of the form $a + bi + cj + dk$ in which a, b, c, d are real.

The sum $(a + bi + cj + dk) + (a' + b'i + c'j + d'k)$ is defined as $(a + a') + (b + b')i + (c + c')j + (d + d')k$. Multiplication is defined from the equations (generalizing $i^2 = -1$); $i^2 = j^2 = k^2 = -1$, $ij = jk = ki = -ji = -kj = -ik = -1$. Certain laws of arithmetic are satisfied, and subtraction and division (except by zero) are possible. Further, any polynomial equation with quaternion coefficients has a quaternion root.

For many years quaternions were widely used in solving physical problems.

The construction of quaternions can be further generalized. For any set of n^3 real coefficients $C_k{}^{ij}$, a linear algebra, or system H of hypercomplex numbers, can be defined as follows. The elements of H are the expressions $a_1\epsilon_1 + \cdots + a_n\epsilon_n$, in which the units $\epsilon_1, \cdots, \epsilon_n$ take the place of the special quaternions i, j, and k, and are the same for all elements; $a_1, \cdots, a_n$ are arbitrary real numbers. The sum of $a_1\epsilon_1 + \cdots + a_n\epsilon_n$ and $b_1\epsilon_1 + \cdots + b_n\epsilon_n$ is defined as $(a_1 + b_1)\epsilon_1 + \cdots + (a_n + b_n)\epsilon_n$. Their product is defined as the sum $a_1b_1\epsilon_1{}^2 + a_1b_2\epsilon_1\epsilon_2 + \cdots + a_nb_n\epsilon_n{}^2$, in which the $\epsilon_i\epsilon_j$ are given by $\epsilon_i\epsilon_j = C_1{}^{ij}\epsilon_1 + C_2{}^{ij}\epsilon_2 + \cdots + C_n{}^{ij}\epsilon_n$.

Although the study of hypercomplex numbers forms an interesting branch of algebra, the complex numbers and quaternions are the only systems of hypercomplex numbers with real coefficients in which multiplication is associative and division is possible.

Algebraic and transcendental numbers. Numbers like $\sqrt{5}$ and $\sqrt{-1}$, which represent roots of polynomial equations $a_0x^n + a_1x^{n-1} + \cdots + a_n = 0$ with integral coefficients, are called algebraic numbers. It can be shown that any sum, difference, product, or quotient of algebraic numbers is again algebraic, and that so is any root of a polynomial equation the coefficients a_i of which are all algebraic numbers. For this reason, algebraic numbers form a closed subsystem of the system of complex numbers that can be studied without using limits or other infinite processes. Algebraic numbers have many fascinating properties (see NUMBER THEORY).

Real or complex numbers that are not algebraic are called transcendental. In 1851 the French mathematician Joseph Liouville first proved the existence of transcenden-

Existence of transcendental numbers

tal numbers; in 1873 Charles Hermite, also of France, proved that the base e of natural logarithms is transcendental, and in 1882 Ferdinand von Lindemann of Germany proved that π is transcendental. Indeed, the vast majority of real numbers are now known to be transcendental (see below *Infinite cardinal numbers*).

Modular numbers: fields. Another interesting class of number systems, each containing a finite number of elements, can easily be constructed from the positive integers, with the system of prime numbers serving as a point of departure. (Prime numbers are those integers [greater than one] that cannot be expressed as the product of two other integers each greater than one.) By subtracting unity from each prime number, $p = 2, 3, 5, 7, \cdots$, the integers $0, 1, 2, \cdots, p-1$ are obtained. The sum $a+b$ and the product $a \times b$ of any two of these integers are defined as the remainders of the ordinary sum $a+b$ and the ordinary product $a \times b$ when divided, respectively, by p. Thus, if $p = 3$, then $2+1=0$ and $2 \times 2 = 1 = 4-3$; the complete addition and multiplication tables are found in Figure 2. These number systems satisfy the basic laws of arithmetic, and subtraction and division (except by zero) in them is always possible. They were first studied by the German Carl Friedrich Gauss in his *Disquisitiones Arithmeticae* of 1801.

Number systems satisfying the basic laws of arithmetic, in which subtraction and division (except by zero) are always possible, are called fields. All the fields having a finite number of elements are known—they are called Galois fields; for each prime power p^n there is exactly one field having p^n elements. A theorem, proved by the Scottish-born mathematician J.H.M. Wedderburn in 1910, asserts that for systems containing only a finite number of elements, the commutative law of multiplication is a consequence of the other laws.

Hypercomplex-number systems with coefficients in a modular field or any other field can be constructed just as easily as if the coefficients were in the field of real numbers.

Infinite cardinal numbers. Cardinal numbers are defined as a mark associated with a class and with all other classes in one-to-one correspondence with this particular class. There is no requirement, however, for this class to be finite. It is perfectly legitimate, therefore, to ascribe the mark d (denumerable infinity) to the class J of all positive integers and to every other class that can be put in one-to-one correspondence with J.

Similarly, it is perfectly legitimate to ascribe the mark c (power of the continuum) to the class R of all real numbers and to every other class that can be put in one-to-one correspondence with R.

The definitions of addition and multiplication are still valid, as are the proofs of the basic laws of arithmetic. The definition of the relation $a \leqslant b$ is also valid, and laws of order can be proved for infinite numbers as well as for finite numbers.

In fact, exponentiation can be defined for infinite as well as finite cardinal numbers. The usual elementary definition $a^n = a \times \cdots \times a$ (n factors) must be avoided, and a^s must be defined as the cardinal number of the class of all single-valued functions or rules f, assigning to each element x in a class of b elements a single value $y = f(x)$ in a class of a elements. The usual laws of exponents (see 8) can then be easily proved. The infinite cardinal numbers c and d can be connected by the interesting formula $c = 2^d$. The main loss in dealing with infinite cardinal numbers is that the laws of cancellation for multiplication and the similar law for addition are no longer valid. In fact, for every infinite cardinal number, symbolized by the Greek letter gamma, γ, it can be proved that $\gamma = \gamma + 1$, $\gamma + \gamma = \gamma$, and $\gamma \times \gamma = \gamma$. These equations imply that every infinite class can be placed in one-to-one correspondence with a proper subclass of itself. There is an obvious one-to-one correspondence between the class J of positive integers $1, 2, 3, \cdots$ and the class E of even integers $2, 4, 6, \cdots$; another between J and the class O of odd integers $1, 3, 5, \cdots$; it follows from the definition of addition that $d + d = d$.

From $\gamma = \gamma + 1$, the relation $\gamma + 1 = \gamma + 2$ arises, although $\gamma + \gamma = \gamma$ implies $2 \times \gamma = 1 \times \gamma$; hence neither law of cancellation holds. Thus the system of finite and infinite cardinal numbers cannot be extended to larger systems in which subtraction or division is possible.

From $\gamma + \gamma = \gamma$ it can easily be shown that the class of all integers has the same cardinal number d as the class of positive integers; from $\gamma \times \gamma = \gamma$, it follows that the class of all rational numbers $\pm m/n$ also has the cardinal number d; this is even true of the class of all algebraic numbers. This suggests the conjecture that perhaps the class R of all real numbers also has the cardinal number d; this will now be disproved. If this conjecture were correct, it would mean that all real numbers could be written in a sequence, S, just as the integers $1, 2, 3, \cdots$ can be so written.

Cardinal number of the reals

First, one may imagine the numbers of this sequence to be written in decimal form. Now, certain rational numbers can be represented in two ways, which is evident from the fact that $0.879999\cdots = 0.880000\cdots$. For this discussion, the mode of representations by 9s will be used, so that every number shall have a unique decimal representation. Another number, $0.x_1x_2x_3\cdots$, can now be constructed as follows: The first digit, x_1, after the decimal point is chosen as other than 0 or the first digit (after the decimal point) of the first number of S; similarly x_2 is chosen as other than 0 and the second digit of the second number of S, and so on. There will then be defined a number in decimal form that is obviously distinct from all the numbers of S, which disproves the conjecture.

This disproof by the so-called diagonal process of Cantor can be generalized so as to apply to any cardinal number. Thus, there is an infinite sequence of distinct infinite cardinal numbers: $d < 2^d = c < 2^c < \quad < \cdots$. It is not known whether or not there are others in between those listed.

As the cardinal number of the class of algebraic numbers is d, the preceding argument proves that there exist real transcendental numbers and, in fact, that there are infinitely more transcendental numbers than algebraic numbers.

Infinite ordinal numbers. The infinite sequence of non-negative integers $0, 1, 2, 3, \cdots$ has the property that every non-empty subset has a first member (known as the well-ordering property). It was observed by Cantor that this sequence can be extended in exactly one way without losing the well-ordering property. Indeed, there must be a first infinite ordinal number, symbolized by the Greek letter omega, ω, after all the integers, a first ordinal $\omega + 1$ after ω, a first ordinal $\omega + 2$ after $\omega + 1$, and so on. Immediately after the sequence $\omega, \omega+1, \omega+2, \cdots$, there must be a first ordinal 2ω, followed by $2\omega+1$, $2\omega+2, \cdots$; then $3\omega, 3\omega+1, \cdots$. Evidently this process of ordinal numeration can be continued (see 9). Each ordinal number, symbolized by the Greek letter alpha, α, in this sequence can be identified with the set of (kappa) κ preceding α, in order. Thus, 3 corresponds to the ordered set (0, 1, 2), ω with the ordered sequence of all non-negative integers, and so on.

(8) $(a \times b)^s = a^s \times b^s, \qquad a^{s+t} = a^s \times a^t, \qquad (a^s)^t = a^{s \times t}$

(9)
$$\begin{cases} 0, 1, 2, \cdots \\ \omega, \omega+1, \omega+2, \cdots \\ 2\omega, 2\omega+1, 2\omega+2 \\ \quad\vdots \\ \omega^2, \omega^2+1, \omega^2+2, \cdots \\ \omega^2+\omega, \omega^2+\omega+1, \omega^2+\omega+2, \cdots \\ \quad\vdots \end{cases}$$

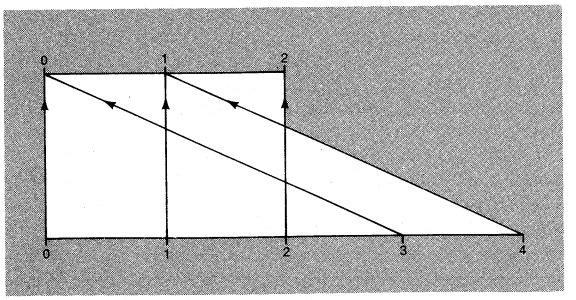


Figure 2: Addition and multiplication for the modular number system containing three elements: add or multiply by the rules of ordinary arithmetic and map the result (follow the arrows) to the sum or product in the modular system.

Proceeding in this manner, one can define $\alpha + \beta$ as the result of laying the sequence β after the sequence α; $\alpha\beta$ as the result of substituting the sequence β for each term of the sequence α and α^2 as $\alpha\alpha$. These definitions are consistent with the notations introduced above; however, $1 + \omega = \omega \neq \omega + 1$ and $\omega 2 = \omega \neq 2\omega$, so that neither addition nor multiplication is commutative. On the other hand, both operations are associative—*i.e.*, $(\alpha + \beta)\gamma = \alpha\gamma + \beta\gamma$—and the one-sided cancellation laws are preserved: $\alpha + \beta = \alpha + \gamma$ implies $\beta = \gamma$, and $\beta\alpha = \gamma\alpha$ implies $\beta = \gamma$.

There appears to be no reason why the process of ordinal enumeration of the elements of a class C cannot be continued indefinitely until all the elements of the class have been counted. That is to say, it seems to be plausible to assume that every class can be well ordered. Using this assumption, it can be proved that $\alpha + \alpha = \alpha$, and $\alpha^2 = \alpha$ for all infinite cardinal numbers. Inasmuch as no infinite class not equivalent to (that is, in one-to-one correspondence with) the class of integers has ever been constructively well ordered, however, this assumption must still be accepted with reservations (see MATHEMATICS, FOUNDATIONS OF).

Ordered fields. In each of the preceding generalizations of the real-number system, either the order properties have been lost, as in the cases of complex, quaternion, and modular numbers; or the possibility of subtraction and division has been lost, as in the cases of infinite cardinal and ordinal numbers. It will now be shown that there exist ordered fields, not contained in the real-number system, that have both the order properties and the arithmetic properties of the rational and real numbers.

Operations on power series

Thus, the class of infinite formal power series $a(x)$ (see 10) begins with an arbitrary positive or negative integral power $a_{-m}x^{-m}$ of the symbol x with real nonzero coefficient a_{-m} (and is composed of a sum of such terms, each with successively increasing powers of x). To add two such series, corresponding coefficients are added; thus, $c(x) = a(x) + b(x)$ means $c_n = a_n + b_n$ for all n. To multiply, the coefficient c_n of x^n in the product $c(x) = a(x)b(x)$ is formed from the sum of all products $a_i b_j$ such that $i + j = n$ (see 11). Defining $a(x) \geqslant 0$ to mean that $a_{-m} \geqslant 0$, adding 0 to the system, and using the relations $0 + a = a$ and $a \times 0 = 0$ to define addition and multiplication by zero, the fundamental laws of arithmetic can readily be proved. It is also easy to show that subtraction is always possible. Further, division (except by zero) is always possible (see 12). The formal power series and 0, therefore, form a field (see above *Modular numbers: fields*). Finally, if $a(x) \geqslant b(x)$ is defined to mean $a(x) - b(x) \geqslant 0$, then the order properties (such as "$a \leqslant b$ implies $a + c \leqslant b + c$, if $c \geqslant 0$") hold.

In fact, the only property of real numbers that is not shared by this new system is the completeness property; *i.e.*, that any increasing sequence of real numbers, whose terms are all bounded above by a fixed quantity, must tend to a limit. In fact, the increasing sequence $-1 < -\frac{1}{2} < -\frac{1}{4} < \cdots$, whose terms are all bounded above by 0, does not tend to a limit in the new system. It does not approach 0, for example, because $-x$ separates 0 from every term of the sequence. Because it can be proved that any ordered field with the completeness property is equivalent to the real-number system, no closer analogue of the real-number system can be proposed. (Ed.)

III. Functions and differential calculus

FUNCTIONS

Higher derivatives. A function $y = f(x)$, for real numbers x and y, is equivalent to a graph in a Cartesian coordinate system. A graph looks like a curve, but a curve is a figure on ordinary paper whereas a graph is a figure on graph paper; and graph paper represents a Euclidean plane onto which some fixed coordinate system has been introduced beforehand.

The derivative df/dx is the limit of the difference quotient $\Delta f/\Delta x$ as Δx approaches 0 (see *The derivative*). It is defined for the numerical function but has an interpretation for the graph—as the slope of the straight line that is tangent to the figure at the given point with abscissa x. The numerical value of the slope, however, does depend on the coordinate system given. This dependence becomes more emphatic when the second derivative of $f(x)$—that is, the derivative of the derivative (see 13)—is introduced. This is formed by interpreting the first derivative as a distinct function to which the limiting process of differentiation is again applied. Second derivatives were in the forefront of Newton's mechanics, mainly because they defined the kinematic concept of acceleration as the instantaneous rate of change of the velocity function, and, by an interpretative correlation with it, the all-important dynamic concept of force. (The Greeks knew no functions or accelerations and only varying notions of force.)

Graphical interpretation of the derivative

General functions. A comprehensive concept of function began to emerge only in the first half of the 19th century, at first only for numbers x, y and gradually for more general objects x, y. In the case of general objects x, y, the word function is sometimes replaced by the words functional or operator, especially when functions for various types of objects are used in the same context simultaneously. For "working" mathematics—that is, when problems relating to foundations of mathematics are not at issue—the unsophisticated description of a function, $y = f(x)$, generally seems sufficient.

Given two aggregates of mathematical objects, an aggregate X and an aggregate Y, it is customary to let x denote an element of X and y an element of Y. With each element x of X there is associated a certain element y of Y and a single one only, but the same element y may be associated with several different elements x. On the other hand, it is not necessary that each element y shall appear as an associate to some element x. Indeed, all elements x may have as associates the same element y, in which case the function is called a constant. In general, the element y associated with x is denoted by a symbol like $y = f(x)$. It is also called the image of x, or the value of f at the "point" x. The entire object $y = f(x)$ is called a function from X into Y or, more explicitly, from the aggregate X into the aggregate Y.

Syllogistically, a mathematical statement is a certain conclusion drawn from particular assumptions and previous statements. A function in general, as just described,

(10) $$a(x) = a_{-m}x^{-m} + a_{-m+1}x^{-m+1} + \cdots + a_0 + a_1x + a_2x^2 + \cdots$$

(11) $$(x + x^2 - x^3 + \cdots)(2x^{-2} + 3x^{-1} - 4 + \cdots) = (2x^{-1} + 5 - 3x + \cdots)$$

(12) $$\frac{1}{(1-x)} = 1 + x + x^2 + \cdots + x^n$$

(13) $$\frac{d^2f}{dx^2} = \frac{d}{dx}\left(\frac{df}{dx}\right)$$

$$\frac{1}{1-x}=1+x^1+x^2+x^3+\cdots+x^n \tag{14}$$

$$a_0+a_1x^1+a_2x^2+\cdots+a_nx^n \tag{15}$$

$$\begin{cases} a_0=f(0), \quad a_1=\dfrac{1}{1}f'(0), \quad a_2=\dfrac{1}{1\times 2}f''(0), \quad \cdots \\ a_n=\dfrac{1}{1\times 2\times\cdots\times n}f^{(n)}(0) \end{cases} \tag{16}$$

$$a_0+a_1(x-h)+a_2(x-h)^2+\cdots+a_n(x-h)^n \tag{17}$$

$$a_0=f(h), \quad a_1=\frac{1}{1}f'(h), \quad \cdots \tag{18}$$

is not sufficiently tangible for statements to be made concerning it, even if X and Y are real numbers (except if one of them is a finite set). Rather, in order to start, it is necessary to impose on the functions some restrictions or qualifications; that is, to single out classes of functions with some particular features that may be analyzed. Together with the general notion of function, certain overall descriptive properties also began to emerge by means of which classes of functions could be characterized: continuity, differentiability, integrability, bounded variation, and others. Earlier, however, especially in the 17th century, it had been taken for granted, expressly or not, that a function is a "formula" or "expression" that is a tangible prescription for finding y if x is given. Certain functions, then familiar, were taken as being basic, as it were; for instance, x, x^2, $\cdots$, sin x, tan x, arctan x, e^x, log x; and a formula arose by taking one or several such functions and performing a finite number of "natural" operations—primarily the four arithmetical basic operations, but including also extraction of roots and, what is important, the substitution of one function into another.

There soon appeared in these manipulations, however, certain expressions involving an infinite, or unending, succession of operations: certain infinite series, infinite products, and infinite continued fractions. Of these, the infinite series proved the most important by far, and two types of such series have advanced into positions of leadership and command that no phase of development has seriously impaired since, although there have been cases of evasion from authority, half in mathematical sport and a little in earnest. These two are the power series and the Fourier series.

Infinite series

Power series. One of the simplest power series is the infinite geometric series (see 14) that expresses the reciprocal of $(1-x)$ as an infinite sum of non-negative integer powers of x (provided $|x|<1$). In a broad sense, this is ancient, since, for instance, Zeno's series, (2), is merely a special case of (4) for $x=\frac{1}{2}$ and the number 1 is subtracted from both sides. The power series (see 15) has a slightly more general form in which the typical power x^n is multiplied by the coefficient a_n. Here, the quantities a_0, a_1, a_2, $\cdots$ are fixed numbers, and x is a variable that, for the present, is assumed to be real-valued. Now if such a series is convergent for all x in an interval $-R<x<R$, in which R is some positive number, and if the sum is denoted by $f(x)$, then this function has "very good" properties in the sense that it is continuous, differentiable, and, in fact, has a second derivative, third derivative, and so on. Furthermore, if the values of all those derivatives are known at one point (namely, at the origin, $x=0$), then the coefficients $\{a_n\}$ of the series (see 15) can be computed. They are expressed (see 16) as multiples of nth derivatives of the function f evaluated at zero, as deduced by Brook Taylor, an English mathematician, in 1715. It is possible to shift the origin into any other fixed point, $x-h$, by starting out with a series (see 17) formally identified with the generalized power series by substituting $(x-h)$ for x. If this series converges in some interval $x_0<x<x_1$ containing the special point h, say in an interval $(-R+h)<x<(R+h)$, and if the sum function is denoted again by $f(x)$, then it has derivatives of all orders in the interval; and, moreover, the coefficients (see 18) are expressible as before, as multiples of nth derivatives of the function f evaluated, however, at the point h rather than at zero.

In order to compute all these derivatives only for the point h, it suffices to know the function $f(x)$ in some partial interval containing the point h, no matter how small this partial interval may be, and the coefficients a_0, a_1, a_2, $\cdots$ are already determined thereby. If, therefore, the class of functions $f(x)$ that can be represented as sums of power series are considered, the following uniqueness property arises: if two such functions are equal in some small interval, then they are equal in any larger interval containing it. Investigators in the 18th century were very impressed with this property because it agreed with a scientific presumption that if the history of a closed mechanical system has been known for a certain time interval, then this knowledge ought to determine the mechanical system for later time as well. The French mathematician Joseph-Louis, comte de Lagrange, especially claimed, with more insistence perhaps than self-conviction, that mechanics, physics, and even "significant" analysis in general ought to be able to do with such functions exclusively. He firmly attached to these functions the name analytic (probably already current before) that they have borne since, and in his works he tried to use them in an exclusive sense.

Piecewise analytic functions. This exclusiveness, however, still allowed for a certain generalization that Lagrange (like others) was taking for granted due to the following situation: if a ball is thrown against a wall and it hits it at the time t_0, then in the time interval $0<t<t_0$ the components of position and velocity and other quantities for the ball are analytic functions of the time variable and of each other. After colliding with the wall at time t_0, the ball flies on its rebound during a certain time interval $t_0<t<t_1$. During this rebound the functions are again analytic, as they were before, but at the instant of collision this is not so. At $t=t_0$, for instance, the velocity components are not even continuous, let alone analytic. Such "instantaneous" interruptions of analyticity, when the "instants" were only finitely many, were acceptable to Lagrange and others. Their occurrence was taken for granted and viewed as a necessary generalization of the demand on analyticity, if not of its basic definition.

In the 20th century, such functions have been called "piecewise" analytic, and this manner of generalization became important in many contexts. There are functions that are piecewise continuous, piecewise differentiable, etc. Thus, an indispensable method in algebraic topology, which was initiated in 1911 by the Dutch mathematician L.E.J. Brouwer, is built on approximating to continuous functions in several variables by piecewise linear ones.

The uniqueness property for analytic functions is indicative of great regularity of structure. Given a small piece of the function—that is, given the function in only a small interval—it acquires a natural growth into a larger interval, and uniquely so. Furthermore, if the x-axis on which the function is defined is placed into the plane of the complex variable $z=x+iy$, then the function spreads in a unique manner from the one-dimensional interval on the x-axis into a two-dimensional domain of the plane. Syllogistically, the decisive phenomenon of this spread is the fact that if a power series (see 15) is convergent on a real interval $-R<x<R$, and if the real variable x is replaced by the complex variable $z=x+iy$, then the power series with the same coefficient $a_0+a_1z^1+a_2z^2+\cdots$ will also converge in the entire disk $|x+iy|<R$.

The complex analytic functions thus arising lead to a theory of great structural variety and beauty. The mathematicians of the 19th century loved and admired it all. They gave it their best and molded into it all the glamour of the "Victorian era" then regnant everywhere and in everything. Around 1900 and later, a number of books on such functions bore the generic title "Theory of Func-

tions," in the expectation that the "other" functions would be identified in some other manner; and so they were and came to be called functions of a real variable or of real variables.

The theory of the latter functions, too, was growing into a compound of many parts, and in the 20th century it was even beginning to gain an ascendancy of interest over the complex analytic area. Subsequently, the theory of complex analytic functions became revitalized, however, and branched out from one into several variables. A balance of evenly divided interest was established between the two theories.

Fourier series. Fourier series are intimately related to the theory of functions of real variables, although not to all of it, and yet they were a great ferment for the development of the theory in all its parts. Only the high points of Fourier series will be described here. (A more detailed discussion is presented in ANALYSIS, FOURIER.) Functions of bounded variation arose from a paper by the German mathematician P.G.L. Dirichlet, giving the first non-obvious convergence criterion for Fourier series. Also in Germany, Bernhard Riemann developed the theory of the integral, bearing his name, in a paper on trigonometric series; and Cantor, by pursuing a problem raised in that paper, was led to construct the first part of his theory of sets; namely, the theory of point sets in Euclidean space. The theory of summability of divergent series, originally designed as an aid for the analytic continuation of power series, became lastingly indispensable when Leopold Fejér applied it to Fourier series. The importance of the Lebesgue integral was recognized when it led to the Riesz–Fischer theorem for Fourier series and to the Plancherel theorem for Fourier integrals. The Stieltjes integral, originally designed for the so-called moment problem, acquired its stature when it entered Fourier analysis in the theory of probability.

In physics and technology, many phenomena are viewed as waves, including phenomena known as vibrations and oscillations, particularly in studies such as acoustics, hydrodynamics, electricity, optics, and so on. In each context there is a certain first type of wave, not yet too complicated and yet already important, and the interpretations and names associated with it are closely analogous from context to context. There is always a certain decomposition of the general wave within that type into a series of component waves that themselves are no longer decomposable. The first in the series is the "basic," "lowest," "primary," etc.; the later ones are "refinements," "higher," "secondary," etc., and they are supposed to decrease in effect and in influence the later they appear in the series. Each component has certain fixed attributes, such as (wave) length, frequency, phase, and amplitude.

This analogy in the phenomena is simply due to the identity of the mathematical instrument used for their "decomposition." Invariably, a certain function of one variable x in the interval $-\infty < x < \infty$ is represented as a Fourier series, and the attributes of the phenomena are simply attributes of its series.

Representation of periodic functions

Such a function is a periodic function; that is, there is a fixed number p such that $f(x+p) = f(x)$. It is normalized by putting $p = 2\pi$ so that $f(x+2\pi) = f(x)$, and its Fourier series is then an expansion (see 19) formed by summing a constant term $\frac{1}{2}a_0$ and successive cosine and sine terms, each with a coefficient a_n or b_n and each with a typical argument, nx, composed of an integral multiple of x. The first coefficient is written with the number $\frac{1}{2}$ in order to make the formulas (see 20) for the coefficients $\{a_n\}$ and $\{b_n\}$ simple. The latter formulas, written as integrals, express the average behaviour of the products of f times cosine and sine factors over an interval. All terms in the Fourier series have the period 2π; however, for the computation of the coefficients a_n, b_n, it suffices to know the function $f(x)$ only in the interval $-\pi < x < \pi$. This agrees with the fact that if a function $f(x)$ is initially given in the latter interval then it can be extended to a periodic function by repeating it successively in the intervals $(\pi, 3\pi)$, $(3\pi, 5\pi)$, $\cdots$ to the right and similarly in the intervals $(-3\pi, -\pi)$, $(-5\pi, -3\pi)$, $\cdots$ to the left.

For purposes of theoretical reasoning and of applications in physics it is appropriate to rewrite the Fourier series (19) into two more forms. The first is the complex version (see 21) that arises by the substitution into the earlier formula of the cosine and sine (see 22), expressed in terms of linear combinations of exponentials with complex arguments. In this version, the coefficients (see 23) in the series are integral averages, as before, with the complex exponential replacing the trigonometric functions.

The second form arises if for $n \geqslant 1$ are introduced the number $\rho_n = (a_n^2 + b_n)^{1/2}$ and an angle θ_n for which $a_n = \rho_n \cos\theta_n$, $b_n = \rho_n \sin\theta_n$. The series (see 24) then becomes a sum of cosine terms with ρ_n as the typical coefficient and $(nx - \theta_n)$ as the typical argument.

In physical applications it is frequently pertinent to replace x by $x - at$, where t is time and a is the velocity of the wave; and the result is again (see 25) a cosine series. The constant term does not represent anything "wavy" and it is frequently zero. The next term, however, (see 26), represents the lowest indecomposable component of the wave, and the successive terms (see 27) the higher ones. The number $2\pi/n$ is the length of the nth component of the wave, $na/2\pi$ its frequency, θ_n its phase, and ρ_n its amplitude. The interpretation of the amplitude attaches itself to an important relation, called the Parseval equality. It states that for any Fourier series there appears, (see 28) the integral average of the square of the function, f, equal to the sum of the squares of the Fourier coefficients. For some wave phenomena, the quantity (see 29) that appears in the Parseval equality (namely,

The Parseval equality

(19) $$f(x) = \frac{1}{2}a_0 + (a_1\cos x + b_1\sin x) + \\ + (a_2\cos 2x + b_2\sin 2x) + \cdots + \\ + (a_n\cos nx + b_n\sin nx)$$

(20) $$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos nx\,dx, \quad b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin nx\,dx$$

(21) $$f(x) = c_0 + (c_1 e^{ix} + c_{-1}e^{-ix}) + \cdots + \\ + (c_n e^{nix} + c_{-n}e^{-nix})$$

(22) $$\cos nx = \frac{e^{inx} + e^{-inx}}{2}, \qquad \sin nx = \frac{e^{inx} - e^{-inx}}{2i}$$

(23) $$\begin{cases} c_m = \dfrac{1}{2\pi}\displaystyle\int_{-\pi}^{\pi} f(x)e^{-imx}\,dx \\ \text{for all integers } m, \text{ positive and negative.} \end{cases}$$

(24) $$f(x) = \frac{1}{2}a_0 + \rho_1\cos(x-\theta_1) + \rho_2\cos(2x-\theta_2) + \\ + \cdots + \rho_n\cos(nx-\theta_n)$$

(25) $$f(x-at) = \frac{1}{2}a_0 + \rho_1\cos(x-at-\theta_1) + \cdots + \\ + \rho_n\cos(nx-nat-\theta_n)$$

(26) $$f_1(x-at) = \rho_1\cos(x-at-\theta_1)$$

(27) $$f_n(x-at) = \rho_n\cos(nx-nat-\theta_n)$$

(28) $$\begin{cases} \dfrac{1}{\pi}\displaystyle\int_{-\pi}^{\pi}[f(x)]^2 dx \\ \qquad = a_0^2 + (a_1^2 + b_1^2) + \cdots + (a_n^2 + b_n^2) \\ \qquad = a_0^2 + \rho_1^2 + \rho_2^2 + \cdots + \rho_n^2 \end{cases}$$

(29) $$\frac{1}{\pi}\int_{-\pi}^{\pi}[f(x)]^2\,dx$$

the integral average of the square of f) represents a certain volume of energy associated with the propagation of the wave. The term ρ_n^2 is the volume of energy for the component wave $f_n(x)$; therefore, if the neutral term a_0 is ignored then the Parseval equality states that this energy for the whole wave is the sum of the energies for all component waves.

Analytic theory of numbers. The trigonometric expansion (19) is named after the early 19th-century French mathematician and physicist Baron Fourier largely because he found a counterpart to it in which the sum is replaced by an integral. It is the pair of formulas (see 30) expressing g, a function of the real variable y, as an integral transformation of a function f, the argument of which is x, and vice versa, in which either integral extends over the whole real line and applies to a product of f or g multiplied by one of two complex exponential functions of both variables x and y. The formulas belong together in the following sense: For certain classes of functions $f(x)$ in $-\infty < x < \infty$ the first integral exists for all y in $-\infty < y < \infty$ so that a function $g(y)$ arises. With this function the second integral can be formed, and the resulting function $f(x)$ is the one from which the process started. This pair of formulas has given rise to a theory of duality in abstract analysis.

Fourier's contemporary and countryman Siméon-Denis Poisson and others before him found a formula that combines the Fourier series with the Fourier integrals, and the simplest case of it (see 31) identifies the sum of integral transforms of f (evaluated on the integers) with the sum of real numbers mapped by f from the integers. It became one of the most important tools in the analytic theory of numbers when Dirichlet in Germany used it in 1837 for giving a new proof for the Gauss reciprocity law. In number theory, the greatest master of this Poisson summation formula was the German-born mathematician Erich Hecke.

Least-square approximation. There is an amplification to the Parseval theorem that has shaped many conceptions in physics: Let $f(x)$ be a function on $-\pi < x \leqslant \pi$. For a fixed integer $n \geqslant 1$ take any real numbers $A_0, A_1, \cdots, A_n; B_1, \cdots, B_n$ and set up a trigonometric sum, $\sigma_n(x)$ that is (see 32) a constant plus a linear combination of cosine and sine terms in which the $\{A_k\}$ are coefficients of cosine and the $\{B_k\}$ are coefficients of sine terms.

Partial sum of terms

A particular case of this sum is the nth partial sum (see 33) of the Fourier series, written $s_n(x)$; that is, the sum that includes the first n trigonometric terms of each kind in the Fourier series is, among the sums $\sigma_n(x)$, the one that arises if for the coefficients A_m, B_m are chosen the numbers a_m, b_m that are given by the Fourier integral formulas (20). Now form the difference $f(x) - \sigma_n(x)$ and ask for which choice of the coefficients A_m, B_m this difference will be a least possible one. A good answer to this question arises if the size of the difference is measured by a certain average of it; namely, by the value of the integral (see 34) of the square of the difference (multiplied by a constant). This quantity is always greater than or at best equal to the equivalent expression (see 35) that involves $s_n(x)$ rather than $\sigma_n(x)$, so that it is smallest if the trigonometric sum is the corresponding partial sum of the Fourier series for $f(x)$. Furthermore, the function $f(x) - s_n(x)$ has itself a Fourier series that arises by cutting off from the entire Fourier series (see 36) the partial sum $s_n(x)$, thus yielding an infinite sum of sine and cosine terms. Therefore, by the Parseval equality, the integral (see 35) that measures, in the least-squares sense, the difference between $f(x)$ and $s_n(x)$ has as a value the sum of squares of all Fourier coefficients with subscript larger than n, and altogether the following result arises: If the degree of approximation is measured by the integral of the square of the difference, and if $\sigma_n(x)$ is a trigonometric polynomial of prescribed degree n, then the best approximation is brought about by the partial sum of the Fourier series, and its size is that given by (see 37) the sum of squares of all Fourier coefficients with subscripts larger than n.

This result is not restricted to expansions in which the simple functions are cos nx, sin nx, but it holds if they are a so-called complete orthonormal system, and in this way it broadens into a comprehensive proposition in many settings.

$$(30)\quad \begin{cases} g(y) = \int_{-\infty}^{\infty} f(x)e^{2\pi iyx}\,dx \\ f(x) = \int_{-\infty}^{\infty} g(y)e^{-2\pi ixy}\,dy \end{cases}$$

$$(31)\quad \sum_{m=-\infty}^{\infty} \int_{-\infty}^{\infty} f(x)e^{2\pi imx}\,dx = \sum_{n=-\infty}^{\infty} f(n)$$

$$(32)\quad \sigma_n(x) = \frac{1}{2}A_0 + (A_1 \cos x + B_1 \sin x) + \cdots + (A_n \cos nx + B_n \sin nx)$$

$$(33)\quad s_n(x) = \frac{1}{2}a_0 + (a_1 \cos x + b_1 \sin x) + \cdots + (a_n \cos nx + b_n \sin nx)$$

$$(34)\quad \frac{1}{\pi}\int_{-\pi}^{\pi} |f(x) - \sigma_n(x)|^2 dx$$

$$(35)\quad \frac{1}{\pi}\int_{-\pi}^{\pi} |f(x) - s_n(x)|^2 dx$$

$$(36)\quad f(x) - s_n(x) = a_{n+1}\cos(n+1)x + b_{n+1}\sin(n+1)x + \cdots$$

$$(37)\quad a_{n+1}^2 + b_{n+1}^2 + a_{n+2}^2 + b_{n+2}^2 + \cdots$$

$$(38)\quad \frac{f(1+h) - f(1)}{h} = \frac{(1+h)^2 - 1^2}{h} = \frac{2h + h^2}{h} = 2 + h$$

$$(39)\quad \frac{f(0+h) - f(0)}{h} = \frac{h - 0}{h} = 1$$

$$(40)\quad \frac{f(0+h) - f(0)}{h} = \frac{-h - 0}{h} = -1$$

DIFFERENTIAL CALCULUS

The derivative. If $f(x)$ is a real-valued function of a real variable and x_0 a real number, consider the difference quotient constructed by taking the ratio of the difference $f(x_0 + h) - f(x_0)$ to h, in which h is any real number except zero. If this difference quotient approaches a number as h approaches zero, this number is called the derivative of the function $f(x)$ at the point x_0. If $f(x) = x^2$ and $x_0 = 1$, for example, the difference quotient is, by a direct calculation (see 38), seen to be $2 + h$. As h approaches zero the difference quotient approaches the number 2. The function thus has for its derivative the value 2 at the point 1.

The difference quotient

On the other hand, there exist other functions, called "well-behaved," that fail to have a derivative at some given point. A simple example is the following: Let $f(x) = x$ if x is positive or zero and $f(x) = -x$ if x is negative (that is, $f(x) = |x|$, the absolute value of x) and choose x_0 to be 0. It follows that if h is positive the difference quotient is the number 1, as shown by calculation (see 39); whereas, if h is negative, the quotient is -1 (see 40). Thus $f(x)$ does not have a derivative at $x_0 = 0$ because, arbitrarily close to 0, the difference quotient assumes the values 1 and -1; *i.e.*, this difference quotient does not approach one unique number as h approaches zero.

The derivative of $f(x)$ at x_0 (if it exists) is usually denoted by the symbol $f'(x_0)$ or $\frac{df}{dx}(x_0)$.

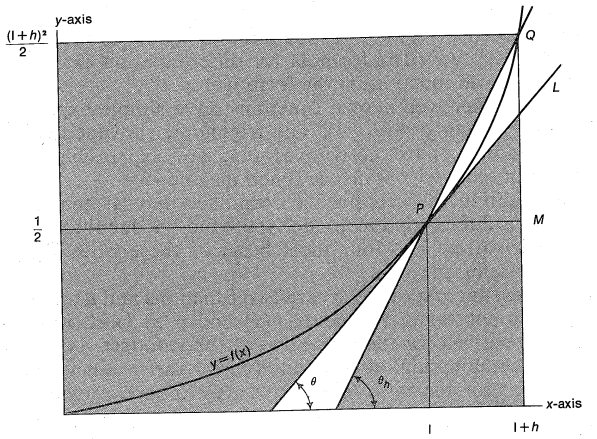


Figure 3: Geometric representation of the concept of the derivative (see text).

Geometrical interpretation. Perhaps the quickest way to obtain some insight into the concept of the derivative is to consider its geometrical interpretation. For example, if $f(x) = x^2/2$ and $x_0 = 1$ then inspection of the graph of this function near the point $P = [x_0, f(x_0)] = (1, \frac{1}{2})$ shows that the difference quotient, the ratio of $f(x_0 + h) - f(x_0)$ to h, is equal to the ratio QM/PM (see Figure 3). But, as is well-known from analytic geometry, this ratio is the slope of the line passing through the two points P and Q; that is, it is the trigonometric tangent of the angle, symbolized by the Greek letter theta followed by an inferior h, θ_h, made by this line and the x-axis. Now if L be the line tangent to the graph of $f(x)$ at the point P, then as h approaches zero, the point Q moves along the graph of $f(x)$ toward the point P and the angle θ_h approaches the angle θ made by the line L and the x-axis. Consequently, the trigonometric tangent of θ_h approaches the trigonometric tangent of θ. The latter, by definition, is the slope of the line L. On the other hand, by the definition of the derivative, this slope must also be the derivative of $f(x)$ at x_0. The derivative of $f(x)$ at x_0, therefore, is the slope of the line tangent to the graph of $f(x)$ at the point $[x_0, f(x_0)]$.

This geometrical interpretation and the argument justifying it extend equally well to other functions, provided that their graphs are sufficiently smooth; *i.e.*, smooth enough to guarantee the existence of a unique tangent line passing through the point in question. Conversely, if the graph of $f(x)$ fails to satisfy this condition at a point, then the derivative fails to exist; it has already been shown that the function $f(x) = |x|$ fails to have a derivative at $x_0 = 0$. This is reflected geometrically by the fact that the graph of $f(x)$ has a "corner" at the point $[x_0, f(x_0)] = (0, 0)$ (see Figure 4).

Non-existence of derivative

Instantaneous velocity. An important physical application of the derivative is embodied in the concept of instantaneous velocity. In fact, a precise definition of the velocity of a moving particle at a given time, t, involves exactly the same limiting process that occurred in the definition of the derivative, as will now be shown.

If, for simplicity, a particle is considered to be moving along a straight line and in a fixed direction, then its position as a function of time can be represented in the following way. If some point along the particle's path of motion is selected, then the particle's position at time t is completely determined by the distance, $f(t)$, between the particle and the chosen point, and no other dimension is needed to locate the particle.

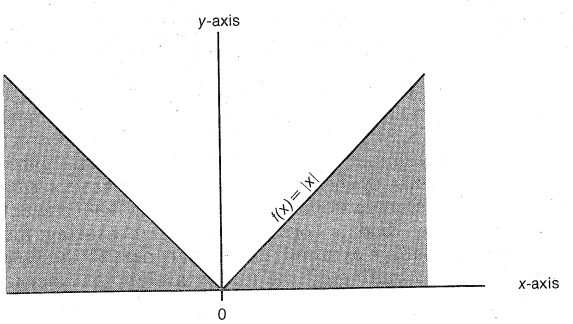


Figure 4: Graph of $f(x) = |x|$ showing the absence of a derivative for the function at $x_0 = 0$.

Using this notation, the average velocity of the particle during an interval of time h, beginning at time t_0, is defined to be the ratio of $f(t_0 + h) - f(t_0)$ to h, as in the non-temporal case; that is, the distance travelled during this interval of time divided by the length of the interval.

Particle velocity and freely falling bodies

On the other hand, a little reflection will show that, whatever meaning is attached to the concept of the instantaneous velocity of the particle at time t_0, it should, first of all, be a number, and, second, the average velocities of the particle during small intervals of time starting at time t_0 should approximate this number. Comparison of the ratio just considered with the difference quotient shows that these considerations make it obligatory to define this instantaneous velocity to be the derivative $f'(t_0)$.

An important example of this situation is given by a freely falling body. It is an experimentally established law that if a body is allowed to fall from rest it will traverse, in time t, a distance that is proportional to t^2 (provided the medium through which it falls offers no resistance). This law can also be expressed by asserting that the distance travelled by the body is a function of time, the expression having the form $f(t) = ct^2$ (in which the constant c depends on the units used in measuring distance and time). Thus the average velocity of the body during an interval of time h, beginning at time t_0, is by calculation (see 41), found to be $c(2t_0 + h)$. The limiting value of this ratio, as h approaches zero, is $2ct_0$. Thus, the instantaneous velocity of the freely falling body at time t_0 is $2ct_0$ and also $f'(t_0) = 2ct_0$.

$$(41)\qquad \frac{f(t_0+h)-f(t_0)}{h} = \frac{c(t_0+h)^2 - ct_0^2}{h} = \frac{2ct_0h + ch^2}{h} = c(2t_0+h)$$

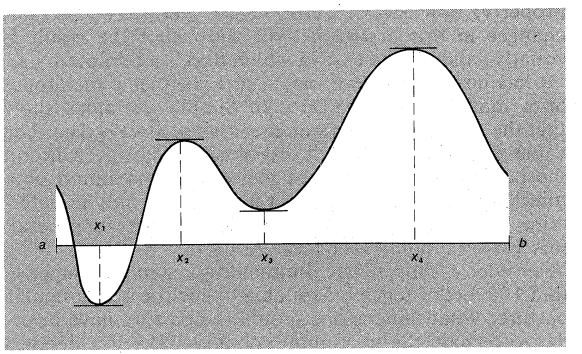


Figure 5: Maxima and minima of a function.

Maxima and minima. It is now possible to discuss one of the more important applications of elementary differential calculus: the location of the maxima and minima of a function. If a function $f(x)$ of a real variable, x, is at least as large at $\bar{x}$ as it is at all other values of x throughout an interval about $\bar{x}$, then $f(x)$ is said to have a relative maximum, $f(\bar{x})$, at $x = \bar{x}$. If, on the other hand, $f(\bar{x})$ does not exceed the values $f(x)$ when x ranges throughout an interval about $\bar{x}$, then $f(x)$ is said to have a relative minimum at $x = \bar{x}$. In Figure 5, x_1 and x_3 are the locations of minima, while x_2 and x_4 specify the locations of maxima.

Now, if $f(x)$ has a derivative $f'(x)$ for all values of x satisfying $a < x < b$, then the tangent to the graph of $f(x)$ passing through a point $[x, f(\bar{x})]$, corresponding to either a maximum or a minimum, must be horizontal, the slope of the tangent must be zero (see Figure 5). (It

$$(42)\quad \frac{f(x+h)-f(x)}{h}=\frac{50h-2xh-h^2}{h}=50-2x-h$$

$$(43)\quad \begin{cases}(\sin x)'=\cos x\\ (\cos x)'=-\sin x\\ (\tan x)'=\sec^2 x\\ (\cot x)'=-\csc^2 x\end{cases}$$

$$(44)\quad [af(x)+bg(x)]'=af'(x)+bg'(x)$$

$$(45)\quad [f(x)g(x)]'=f(x)g'(x)+f'(x)g(x)\quad \text{(Leibniz's rule)}$$

$$(46)\quad \left[\frac{f(x)}{g(x)}\right]'=\frac{f'(x)g(x)-f(x)g'(x)}{[g(x)]^2}$$

$$(47)\quad h'(x)=f'[g(x)]\cdot g'(x)$$

$$(48)\quad p(x)=a_0+a_1x+\cdots+a_kx^k$$

$$(49)\quad p'(x)=a_1+2a_2x+\cdots+ka_kx^{k-1}$$

$$(50)\quad \begin{cases}f(x)=\dfrac{\sin u}{1+x}, & u=x^2+3x+2\\ f'(x)=\dfrac{(1+x)(2x+3)\cos u-\sin u}{(1+x)^2}\end{cases}$$

Meaning of zero derivative

has been assumed here that $a<\bar{x}<b$.) The rigorous proof of this fact is not difficult: For, assuming $f(x_0)$ to be a maximum, it is seen that if h is positive and sufficiently small, the difference quotient cannot be positive, because $f(x_0)\geqslant f(x_0+h)$ if x_0+h is near enough to x_0; similarly, for h negative, but of small absolute value, the difference quotient cannot be negative (because in this case both the numerator and the denominator are negative). The number $f'(x_0)$, therefore, must be arbitrarily close to numbers that cannot be negative and to numbers that cannot be positive. The only number having this property, however, is zero; hence, $f'(x_0)=0$. Obvious changes in this argument will show that the result is equally valid for the case in which $f(x_0)$ is a minimum.

It has just been shown that if $f(x)$ is either a maximum or a minimum, then $f'(x)=0$. Simple examples show that the converse of this statement is not always true. The value of the positive result just obtained, however, lies in the fact that if it is desired to locate a maximum or a minimum of a function $f(x)$, it is necessary only to determine those numbers x for which $f'(x)=0$. In general, this greatly simplifies the task.

Consider a simple, but illuminating, example. Suppose that 100 feet of fence is available to enclose a rectangular pasture. What dimensions should the pasture have in order to contain the maximum possible area? If x denotes one length of the rectangle, then $50-x$ must be the other length. Thus, the area of the pasture is $(50-x)x=50x-x^2$. The problem, then, reduces to finding the largest value (hence, a maximum) of the function $f(x)=50x-x^2$, $0\leqslant x\leqslant 50$. It is found by calculation (see 42) that the difference quotient is $50-2x-h$; therefore $f'(x)=50-2x$. A maximum must satisfy, thus, the equation $f'(x)=50-2x=0$; therefore, $x=25$. The area is largest when $x=25$; that is, when the rectangle is a square of side 25 feet.

Differentiation formulas; computation of derivatives. It has already been shown that if $f(x)=x^2$ then $f'(x)=2x$. This is just a special case of one of several formulas for differentiation that are obtained in elementary calculus. It follows from the definition of the derivative that if $n=1, 2, 3, \cdots$ is any positive integer and $f(x)=x^n$, then $f'(x)=nx^{n-1}$.

In fact, by means of an appropriate limiting process, it can be shown that, for each real number n, the expression x^n for positive x has a natural definition as a real number (for example, $n=\frac{1}{2}$ gives the familiar square root function: $x^{1/2}=\sqrt{x}$). The formula for differentiation extends to include all functions of the form $f(x)=x^n$.

Many other well-known functions have simple expressions for their derivatives. The trigonometric functions, for example, have derivatives that are expressible in terms of other trigonometric functions (see 43).

In addition to such specific formulas, several general rules for differentiation are derived from the definition of the derivative. The four most basic of these rules will now be referred to.

Suppose that $f(x)$ and $g(x)$ are two functions and a and b are two constants. Whenever $f'(x)$ and $g'(x)$ exist, rules can be derived for the derivative of the sum (see 44), or linear combination, of $f(x)$ and $g(x)$, for the product (see 45), and for the ratio (see 46) of $f(x)$ to $g(x)$, provided $g(x)\neq 0$.

The chain rule

The fourth basic rule, the chain rule, may be stated in the following way. Suppose that, for each real number x for which $g(x)$ is defined, the number $g(x)$ is such that $f[g(x)]$ is defined. The composite function $h(x)=f[g(x)]$ can then be formed. The chain rule asserts that, if both $g'(x)$ and $f'[g(x)]$ exist, then $h'(x)$ exists and equals (see 47) the product of the former two derivatives.

These four rules, together with formulas for the derivatives of the powers of x and the basic trigonometric functions of x, permit the computation of the derivatives of a large class of functions. For example, the reader can check that from the results already discussed, the derivative of the most general polynomial (see 48) can be obtained (see 49).

In addition, a formula for the derivative of the most general rational function (*i.e.*, the quotient of two polynomials) can be obtained.

Use of all four of the above rules reveals that functions having fairly complicated expressions have explicit formulas for their derivatives. Many special cases can be solved, such as the derivative (see 50) of the ratio of the sine of a polynomial to a polynomial. It was the development of formulas and rules such as those that have been discussed, enabling mathematicians to calculate derivatives in this manner, that motivated the use of the name calculus for this mathematical discipline.

IV. Measure and integral calculus

MEASURE

The abstract concepts of measure and of integration are best understood in the light of the historical evolution and growth of the problem of measure. This problem always has been present in everyday life as well as in science. It pertains to measures of such properties as the length of a curve (rectification), area of a surface (quadrature), volume of a solid (cubature), mass or weight or electric charge of a body, probability of a random event, and so on. Yet the abstract concept of measure, of which the preceding measures are but particular concrete cases, was not isolated until the 20th century. The corresponding evolution may be loosely divided into several periods, each characterized by its particular attitude toward the problem of measure.

The empirical period. The purely empirical period began with prehistory and lasted until the 6th century BC. The most advanced organized knowledge of this era was that of the ancient Babylonians and Egyptians, providing precise and correct rules for finding areas and volumes of such geometric figures as the triangle, trapezoid, and circle. For the Babylonians, π equalled 3 (this value of π is also found in the Old Testament), and for the Egyptians π equalled $256/81$ $(=3.160\cdots)$. They also had correct rules for the volumes of the parallelepiped, prism, pyramid, and circular cylinder, thought of in concrete terms —mainly as storage containers for grain. The ancient Egyptians even had a correct rule for measuring the frustum of a square pyramid; *i.e.* the part next to the base and formed by cutting off the top by a plane parallel to the base.

Babylonian and Egyptian knowledge

Apparently these rules were discovered empirically. To become proved mathematical propositions, deduced from

a set of definitions and postulates, they had to await the "Greek miracle": the emergence of organized abstract thinking in Greece, which includes the earliest known treatment of mathematics as a deductive discipline. Available information prompts some historians to attribute the birth of abstract mathematics to Thales of Miletus (6th century BC), one of the Seven Wise Men of ancient Greece.

The Greek period. This period began with Thales of Miletus and lasted until about 200 BC. By the 5th century BC, problems of quadrature and cubature were so popular that Aristophanes in his satirical comedy *The Birds* alludes to the quadrature of circles, and the problem of duplication of cubes is attributed to the Delphic oracle. The first abstract proofs of rules for finding some areas and volumes are said to have been developed by Eudoxus of Cnidus in about 367 BC. About a century later, his method of approximation was developed and exploited by Archimedes, called the greatest mathematician of antiquity. This procedure, called the method of exhaustion in the 17th century AD, was at the root of all further developments of the mathematical problem of measure. The method culminated in the 19th-century isolation of the concept of Riemann integration, defined by means of approximating Riemann sums (which, in all justice, could be called Eudoxus or Archimedes sums or, as will be seen below, Cauchy sums).

Method of exhaustion

In modern terms, the method of exhaustion applied to surfaces, for example, may be stated as follows: If S is a surface of unknown area s, then one may choose another surface of known area s' contained in S, and yet another surface S'' of known area s'' containing S; thus $s' < s < s''$. The approximating surfaces S' and S'' are polygons or sums of slices, mainly trapezoidal or rectangular, selected by Eudoxus and by Archimedes according to the particular figure S. In fact, sums of rectangular slices (the areas of which are particularly simple to compute) were to become predominant from the 16th century AD on. To complete the proof, Archimedes used the following proposition: If a known area a is such that for every positive ϵ there exist surfaces S' and S'' with $s' < a < s''$ and $s'' - s' \leqslant \epsilon$, then $s = a$. Archimedes did not isolate this proposition. He proved it, in concrete geometric terms for every particular figure that he considered, by a double *reductio ad absurdum* (double, because negative numbers were not available to him): If $s < a$, then $\epsilon < a - s$ is taken; but if $s > a$, then the choice $\epsilon < s - a$ is made. In either case, a contradiction is reached, so that $s = a$. The method is not complete, for it does not specify how the value of a is to be found. In fact, Archimedes did fill this gap, but his letter to Eratosthenes of Cyrene that described what is sometimes called the Archimedes method remained unknown until 1906. In this letter, Archimedes gave his heuristic approach: a surface is thought of as a sum of lines to each of which the lever principle is applied.

A surface as a sum of lines

The idea of a sum of lines was rediscovered in Italy in the 16th century. Hindu–Arabic numerals were adopted. Algebra was stimulated, letters began to stand for quantities, and symbols were introduced to denote operations. Inexorably, through the period of indivisibles to the emergence of calculus, these representations led to a new attitude, an analytic one in lieu of the geometric attitude of the Greeks.

The period of indivisibles. This period extended from 1615 to about 1670, heavily influenced by Archimedes, whose works had become available in 1544 in Greek and in Latin. The concept of integration (see below *Integral calculus*) gradually emerged, but neither in Archimedean nor in modern dress—especially in England; there, integration became for a time not a primary operation but the inverse of differentiation.

After the efforts of the Dutch mathematician and physicist Simon Stevin and other followers of Archimedes came the pioneering work of the German astronomer Johannes Kepler, *Nova Stereometria Doliorum Vinariorum* (1615; "Solid Geometry of a Wine Barrel"). From then on and until the 19th century, Archimedean rigour was abandoned in favour of an intuitive "infinitesimal"

(51) $\sum_{k=0}^{n-1} f(ka/n), \qquad \sum_{k=0}^{n-1} g(ka/n)$

reasoning. For Kepler, the circle consisted of an "infinity" of infinitesimal triangles with a common vertex at the centre; the sphere consisted of an "infinity" of pyramids, and so on. Nevertheless, he obtained over 80 new results for certain areas and volumes. The language of infinitesimals, or method of indivisibles, was then codified by the Italian mathematician Bonaventura Cavalieri in 1635, in his *Geometria Indivisibilibus Continuorum Nova Quadam Ratione Promota* ("A Certain Method for the Development of a New Geometry of Continuous Indivisibles"). For the first time in the history of the problem of measure, instead of measures of specific geometric figures, general principles for quadratures and cubatures were announced. While not proved rigorously, the principles were at least justified in the language of infinitesimals.

The Cavalieri principle

For quadratures, for example, the Cavalieri principle may be stated as follows: if the ratio of lengths of sections of two surfaces generated by a straight line that moves parallel to a fixed direction is a constant, then the ratio of their areas is equal to this constant. The justification of this principle, in the language of infinitesimals, can be described in modern terms as follows: Every real-valued non-negative function f, defined on a set A of reals, is to be associated with its ordinate set in the (x, y)-plane; that is, the set of all points (x, y) such that x belongs to A, and y varies from 0 to $f(x)$. If A is the segment $[0, a]$, then the ordinate set of f is the surface in the (x, y)-plane, the boundaries of which are the vertical lines $x = 0$ and $x = a$; the horizontal line $y = 0$; and the graph of the function f. For two such functions f and g on $[0, a]$ with coordinate sets $S(f)$ and $S(g)$ of areas $s(f)$ and $s(g)$, the ratio of the sums of equidistant ordinates (see 51) for f and for g can be made as close as desired to the ratio $s(f)/s(g)$ of the areas, provided n is sufficiently large. (The question of existence of these areas was not raised until the middle of the 19th century.) Letting n approach infinity, the unknown ratio therefore becomes the "limit" of the ratio of the sums of ordinates. The concept of limit was not isolated at that time, however, and Cavalieri actually took n to be directly equal to infinity. Thus appears the *summa omnium linae* ("sum of all lines"). Some mathematicians of this period, however, such as Blaise Pascal and Gilles Personne de Roberval, both of France, spoke of "rectangles with infinitely small bases" instead of lines.

During this period, the method of indivisibles yielded a very large number of special integrals as quadratures, cubatures, and centres of gravity. Most were formulated geometrically. Yet more abstract algebraic formulations were already apparent: Pierre de Fermat of France and Cavalieri integrated x^n for rational values of $n \neq -1$; the French mathematician Grégoire de Saint-Vincent integrated it for $n = -1$ in a geometric form that is readily recognized as the logarithmic function; and Pascal integrated $\cos x$. Fermat, Pascal, and the Scottish mathematician James Gregory geometrically manipulated changes of variables and integrations by parts. In fact, geometric methods, formulations, and results were gradually made analytic. Pascal first discussed the types of sums that he would need and then only computed them in the concrete geometric case under consideration. The English mathematician John Wallis in *Arithmetica Infinitorum* (1655), which has to be translated as "infinitesimal algebra," started straight from algebraic grounds, introducing infinite series and infinite products, and Nicolaus Mercator, a German mathematician, discovered the series expansion of $\log(1 + x)$.

The birth of analytic geometry

The impetus and the tools for abstraction were provided by the growth of algebra, the birth of analytic geometry with René Descartes's *Géométrie* (1637), Fermat's *Isagoge* (written before 1636), and the birth and growth of kinematics, the consideration of motion leaving aside

considerations of mass and force. Consideration of time as a concept in kinematics, and the notion of abscissa (the horizontal coordinate of a point in a plane Cartesian coordinate system), led to a gradual emergence of the idea of independent variable. Velocities and slopes of tangents led to the concept of differentiation. Performance of rectifications, such as that of parabolic arcs, forged a first link between differentiation and integration, because the rectification of curves $y = f(x)$ was reduced to the quadrature of $\sqrt{1 + [f'(x)]^2}$, where $f'(x)$ is the derivative of $f(x)$. In fact, integration lost its predominant role in favour of differentiation. The basic link was provided by the recognition that differentiation and integration are inverse operations, to be found in the English mathematician Isaac Barrow's *Lectiones Geometricae* of 1670. Although a crowning of the period of indivisibles, Barrow's work also represented the dying spirit of that period. He began with kinematical considerations that seemed to lead at once to the basic link. The relation between differentiation and integration, however, was given only much later in two widely spaced propositions and in geometric terms: the derivative in terms of slopes of tangents, and the integral in terms of areas.

Development by Newton and Leibniz

The calculus period. This period extended from about 1670 to the beginning of the 19th century, beginning with the creation of the calculus by Leibniz and Newton.

Newton, influenced by his teacher Barrow and his preoccupation with dynamics, dealing with forces and their relation, primarily to motion, used a universal independent variable conceived primarily as time; thus, he had no concept of functions of several variables, and hence none of partial derivatives. The primary concept was that of fluxion (derivative) and arose from kinematic considerations. The concept of integral was not isolated, and no symbol for integration was introduced. His first basic problem was to find fluxions. Integration was treated in a geometric form as a search for fluents (antiderivatives, primitives, indefinite integrals), functions the fluxions of which are given.

Newton depended mainly on the fact that the derivative of a variable area $F(x)$ under a curve is the ordinate $f(x)$ of this curve. For Newton, moreover, integration was only a particular case of his second basic problem: Given an equality containing fluxions, find the corresponding relation for fluents (solve ordinary differential equations). He solved this problem formally by means of series.

Leibniz, influenced by Pascal, conceived the derivative as the slope of a tangent and the integral as *summa omnium linae* ("the sum of all lines"). The main purpose of his activity was to devise a universal language; that is, a general formalism for the acquisition and organization of knowledge. He largely succeeded in the creation of such a formalism for the calculus, and the present formalism is still essentially and properly his, including the integral symbol (a stylized form of the letter S standing for *summa omnium*). His also are the terms constant, variable, parameter, and (with the help of the Swiss mathematician Johann Bernoulli) the terms function and integral.

It is an unfortunate fact that approaching the fourth quarter of the 20th century, freshman college calculus was still being taught largely in the spirit of 1700 with a thin modern varnish, ignoring the 250 years of evolution of mathematical thought that had followed.

The heroic years of the calculus at the end of the 17th century saw the constant flow of new results from Leibniz and his school, as embodied in the work of Jakob Bernoulli, Johann Bernoulli, and G.F.A. de l'Hospital. During the 18th century, the Scottish mathematician Colin MacLaurin pursued and developed Newton's ideas and attempted to clarify them in his *Treatise of Fluxions* (1742), while the Swiss mathematician Leonhard Euler pursued further the ideas of Leibniz and his school in his *Introductio in analysin infinitorum* of 1748 and attempted to found the calculus upon an obscure passage to the limit, starting with the calculus of differences.

Mathematicians of the 18th century devoted little effort to the shaky foundations of the calculus, apparently happy with the powerful tools at their disposal, the fecundity of the methods, the beauty and consistency of the results. Discussions about the metaphysics of calculus were left to philosophers and theologians, among whom the Irish-born Anglican Bishop George Berkeley figured prominently. The whole subject of metaphysics of calculus vanished into thin air, however, with the French mathematician Jean Le Rond d'Alembert's two articles on "Limite" and "Différentielle" in the *Encyclopédie* (1751–65). At last the concept of limit was isolated, a concept that had run as an almost invisible thread in the fabric of infinitesimals as woven since Stevin and others at the turn of the 17th century. The concept of differentiation, too, was finally defined rigorously in terms of limits. (Yet, after d'Alembert, Lagrange still tried to define integration in terms of Newton's conception of a function as a power series.)

The concept of limit as a basis

The analysis period. A new era opened in France with Augustin-Louis Cauchy's *Resumé des leçons sur le calcul infinitésimal* (1823), the basis of which was the concept of limit. Differentiation was defined by a passage to the limit. At last, the integral was defined directly and in purely analytic terms by returning to the very essence of the method of exhaustion as follows: Given a bounded function $f(x)$ defined for $a \leqslant x \leqslant b$, the interval $[a, b]$ may be divided into n subintervals by means of points $a = x_0 < x_1 < \cdots < x_n = b$, and the Eudoxus–Archimedes–Cauchy sums (see 52; nowadays called Riemann sums) may be formed by adding n terms, each composed of a product of a value of f and the length of a subinterval, with f evaluated at an intermediate point of the subinterval. The limit of these sums, as the largest length of the subintervals converges to zero, is, by definition, the integral of the function f on $[a, b]$, provided that this limit exists and is independent of the choice of the t_k, the intermediate point in the subinterval at which f is evaluated.

A new way of thinking was thus born; existence of integrals as well as of derivatives was no longer a question of faith but a problem to be studied. Cauchy gave an incomplete proof of the existence of integrals of continuous functions, a proof that was completed only in 1875 by the French mathematician Gaston Darboux.

Introduction of bounds

Meanwhile, Riemann in 1854 (published in 1867) gave necessary and sufficient conditions for the existence of the integral called the Riemann integral and showed that continuous and piecewise monotone functions satisfy these conditions. Instead of Riemann sums, Darboux introduced the upper and lower Darboux sums (see 53), replacing the value of f at an intermediate point in an interval by M_k and m_k, which, respectively, denote the least upper and the greatest lower bounds of $f(x)$, as x varies between x_k and x_{k+1}. The passage to the limit yields the upper and lower Darboux integrals. These two integrals coincide if and only if the bounded function f is Riemann-integrable on $[a, b]$.

Meanwhile, the general concept of a real-valued function of a real variable that had appeared in a more or less vague or restrictive form throughout the calculus period was finally isolated by Dirichlet in 1829. Dirichlet examined the problem of convergence of Fourier series. Since the coefficients of these series are given by integrals, the knowledge of Fourier series expanded with the evolution of the concepts of integration and of function. From the time of Dirichlet and Riemann on, the knowledge of trigonometric series and, more generally, of harmonic analysis thus gained in depth and in extension.

In fact, modern set theory, which changed the whole outlook on measure and integration, began in 1872 with

$$(52) \qquad \sum_{k=0}^{n-1} f(t_k)(x_{k+1} - x_k), \qquad x_k \leqslant t_k \leqslant x_{k+1}$$

$$(53) \qquad \sum_{k=0}^{n-1} M_k(x_{k+1} - x_k), \qquad \sum_{k=0}^{n-1} m_k(x_{k+1} - x_k)$$

Cantor's investigation of the structure of some exceptional sets of trigonometric series: sets S such that if a trigonometric series (see 54) converges to zero except at the points of S, then each coefficient $c_n = 0$. The evolution of set theory between 1872 and 1890, due primarily to Cantor's genius, together with the parallel investigation of the system of reals and of functions of reals by such mathematicians as Richard Dedekind, Karl Weierstrass, Paul de Bois-Reymond, all of Germany, and Ulisse Dini of Italy brought on the "great mathematical revolution" at the turn of the 20th century.

The great mathematical revolution

The problem of foundations was revived but was applied to the tremendous wealth of accumulated knowledge at a level of abstraction far beyond that known to the ancient Greeks. In particular, the problem of measure was no longer confined to a few special sets—the geometric figures available to Eudoxus or Pascal. Moreover, the very nature of the problem was different.

$$(54) \qquad \sum_{n=-\infty}^{+\infty} c_n e^{inx}$$

Eudoxus' or Pascal's problem was to find the measure (area, volume) of some specific geometric figure. The existence of this measure was a question of faith, and it was approximated by sums of measures of simple figures. The new problem was to define a measure for as large a class of sets as possible by means of such approximations. The first attempts were made in 1884–85 by the German mathematicians Harnack, Stolz, and Cantor. They did not succeed, because they used only approximations from above; they assigned to any set S the greatest lower bound of sums of ordinary measures of simple sets—intervals, rectangles, parallelepipeds, the finite unions of which cover S.

The outer content so defined does not correspond to the intuitive idea of measure, for it is not additive: in everyday experience as well as in experimental science, the "measure" of a union of two disjoint sets is the sum of their "measures." For example, the volume (or mass) of two disjoint bodies is understood as the sum of their volumes (or masses). This requirement was recognized in Italy and France by Giuseppe Peano and Camille Jordan, who introduced approximations both from above and from below. To a set S, contained in a bounded interval I, they assigned also its inner content (the difference between the ordinary measure of I and the outer content of S). If the inner and outer contents of S are the same, then their common value defines the content of S. This content measure is additive. The most important types of sets encountered in analysis, however, such as the set of all rationals or general open sets, have no content. The content measure is thus of limited use.

The extension period. The extension (or theory of measure) period began in 1898 with Émile Borel, the modern French Eudoxus. It is characterized by the extension of the idea of measure to wider and more abstract classes of sets and by the parallel extension of the idea of integration to wider and more abstract families of functions, until both ideas coalesce. It was stimulated by and intricately interwoven with the explosive growth of general topology, following Maurice-René Fréchet's thesis in France in 1902.

The Borel sets and measure

At the root of the theory of measure lie a few crucial remarks by Borel in his *Leçons sur la théorie des fonctions* (1898). While studying series of functions, Borel found the need for a measure such that sets whose measure is not zero, are not countable. He therefore introduced the Borel measure on the Borel field, prototypes of all measures and all σ-fields considered below.

In the present discussion, the following terminology will be used: Unions of disjoint sets will be called sums, and the symbol $\cup$ for union will be replaced by the Greek letter sigma, Σ. "Countable" will mean finite or denumerable. Countable (or finite) set operations are complementations taking countable (or finite) unions and intersections. A σ-field (or field) of sets is a class of sets closed under countable (or finite) set operations.

Borel began by extending the ordinary measure on intervals to a measure on open sets, as follows: As Cantor had shown that every open set G of reals is a countable sum of open intervals I_j, Borel realized that the "natural" extension consists in assigning to G the sum of ordinary measures of its constituent intervals I_j. Hence, if G is not bounded, this sum may be infinite; but it always exists. By indefinite iteration of countable set operations and using the same extension procedure, he then extended the class of open sets and their measure to the class of Borel sets and their Borel measure. This class is a σ-field (the smallest σ-field containing the class of open sets), and this measure is σ-additive; the measure of a countable sum of sets is the sum of their measures. From then on, the Borel field of Borel sets in any topological space will be the smallest σ-field containing the topology; that is, the class of open sets in the space. (In fact, some authors use the term Borel field as a synonym for σ-field).

The Archimedes of the extension period in France was Henri-Léon Lebesgue. He took the decisive step in his thesis *Intégrale, aire, volume* (1902), which he immediately applied to trigonometric series (1903), and pursued further in *Sur l'intégration des fonctions discontinuous* (1910). Lebesgue combined the ideas of his immediate predecessors, Jordan and Borel, and created and exploited the powerful tools of modern measure and integration. The evolution of his ideas, methods, and results consisted primarily in the gradual shedding of their extraneous coverings.

The "Archimedes of France"

The Lebesgue measure is defined by extending the ordinary measure on intervals by means of outer and inner Lebesgue measures, exactly as the Jordan content was defined, except that coverings by countable unions of intervals replace coverings by finite unions. The intervals are one-dimensional or N-dimensional, and the same method applies. Thus the Lebesgue integral of a function $f \geqslant 0$, defined on a set, becomes simply the Lebesgue measure of the ordinate set of the function, provided this set is Lebesgue-measurable. As may be expected, every Riemann-integrable function is Lebesgue-integrable, but the converse is not true. Riemann's integrability condition for bounded functions f becomes: the discontinuity set of f is of Lebesgue measure 0. Thus, the famous Dirichlet function $f(x) = 0$ or 1 (according as x is irrational or not), which is Lebesgue-integrable on every bounded interval, is not Riemann-integrable.

Lebesgue's main contribution, however, does not lie in the above concepts. The Lebesgue measure, for example, is simply the completed Borel measure; that is, a Lebesgue-measurable set is a Borel-measurable set up to a subset of a set of Borel-measure 0. What really mattered were Lebesgue's discovery and use of the properties of his concepts of measure and of integration.

Using the Lebesgue approach, the problem of finding a process of integration that includes both that of Leibniz (*summa omnium*) and that of Newton (primitive) was solved by Arnaud Denjoy of France (1912) and also by Oskar Perron of Germany (1914).

The cleaning up of Lebesgue's ideas was essentially achieved in the 15 years or so that followed the publication of his thesis. It yielded a general answer to the problem of measure for areas, masses, probabilities, and the like and finally isolated its conceptual substance. The answer and the substance, however, were to be formulated in terms far removed from the everyday experience. The steps in the cleaning-up process included the generalization that stemmed from Thomas Jan Stieltjes of France and was due to J.K.A. Radon of Germany, and the abstraction that began with Fréchet in 1915 and with Constantin Carathéodory of Germany in 1918.

Stieltjes, in *Recherches sur les fractions continues* (1894), in order to express the limit of a particular sequence of analytic functions, replaced (in the approximating sums for Riemann integration) lengths of inter-

(55) A set Ω with a σ-field $\mathfrak{A}$ of its subsets (to be called measurable sets) is a measurable space $(\Omega, \mathfrak{A})$. A function f on Ω, say, to the real line R is said to be measurable if the inverse images by f of Borel sets in R are measurable. (The inverse image by f of a set A is the set of all points ω belonging to Ω, such that the values $f(\omega)$ of f at these points belong to A.) The tremendous advantage of a family F of all measurable functions on a measurable space, as compared, say, to the family of all continuous functions on a topological space, is that not only is F closed under arithmetic operations but it is also closed under sequential limits. In particular, if there is a topology on Ω, and $\mathfrak{A}$ is the σ-field of Borel sets in this topological space, then F contains the family of continuous functions on it. This is the abstraction of Borel's scheme.
If measurable sets are denoted by A (with or without subscripts), then to every such A one can assign its indicator $I_A = 1$ on A, and 0 elsewhere. A measurable function with a finite number of distinct values $x_1, \cdots, x_n$ is called a simple function and is denoted by $\sum_{k=1}^{n} x_k I_{A_k}$.

(56) A measure μ on a class $\mathfrak{E}$ of sets in a space Ω is a non-negative countably additive function of sets (not necessarily finite). If, to simplify the statements, Ω is one of the sets of the class, then this measure can be extended to an outer measure μ^0 on all sets A in Ω by assigning to every A the greatest lower bound of sums of measures of sets of $\mathfrak{E}$ whose countable union covers A. This is the abstraction of the Lebesgue outer measure. The restriction of the outer measure μ^0 to μ^0-measurable sets yields the extension of μ on $\mathfrak{E}$ to a measure on the σ-field (containing $\mathfrak{E}$) of μ^0-measurable sets; a μ^0-measurable set B is such that $\mu^0 C = \mu^0(B \cap C) + \mu^0(B^c \cap C)$ for all sets C in Ω (the symbol $\cap$ means intersection and the superscript c means complement of). The μ^0-measurable sets are the abstraction of Lebesgue-measurable sets. The extension is unique, provided the initial measure given on a field $\mathfrak{E}$ is σ-finite; that is, Ω is a countable union of sets of $\mathfrak{E}$ with finite measure.

vals by masses spread on them. He introduced a distribution of masses by means of a nondecreasing function g. It assigns to every interval $[a, b]$ the increment $g(b) - g(a)$ of the function g; if $g(x) = x$ then the increment reduces to the length $b - a$ of the interval. Thus, in Riemann or in Darboux sums, the increments $x_{k+1} - x_k$ are replaced by the general increments $g(x_{k+1}) - g(x_k)$, and the usual passage to the limit of the so-formed Riemann–Stieltjes or Darboux–Stieltjes sums yields Stieltjes integrals. In 1909, in solving a problem set by Jacques-Salomon Hadamard of France, Frigyes Riesz discovered that these Stieltjes integrals are the most general non-negative continuous linear functionals on the space of all continuous functions f defined on the bounded interval $[a, b]$.

Lebesgue–Stieltjes measure

In 1913 Radon, working in N-dimensional Euclidean spaces, combined Lebesgue's ideas with the Stieltjes generalization of ordinary measure and integration, and with Riesz's discovery; thus appeared Lebesgue–Stieltjes measures and integration. The latter measures are defined like the Lebesgue measure, except that the ordinary measure on intervals is replaced by the mass spread on them, as given by the increments of nondecreasing functions. Lebesgue–Stieltjes integrals of non-negative functions are thus the masses spread on the coordinate sets.

In 1915 Fréchet eliminated the topology for the measurable sets and the integrands, while Caratheodory in 1914 and 1918 got rid of topology for the extension of measures. At last, the concepts of measure and of measurability were isolated. The points that summarize the basic concepts of the theory of measure can be outlined briefly (see 55–58).

The basic setup of the history of measure also yields the foundations of probability theory. Deserving mention is the concept of stochastic independence, which was, and to a great extent still is, essential to the evolution of probability theory since its inception in 1654 by Pascal and Fermat. It, too, became measure-theoretic with the creation and abstraction of finite and infinite product measures by Lebesgue and by the U.S. mathematician P.J. Daniell.

Throughout the evolution of the problem of measure, integration appears as a means for finding measures of sets or assigning measures to sets. Thus, seen in historical perspective, measure, integration, and the related concepts of measurable sets and measurable functions ought to have coalesced into one primary concept from which the preceding could be derived. This was done by Carathéodory and Daniell by cleaning up the extension procedure.

Carathéodory's emphasis on measure

Carathéodory, in 1918, emphasized measure as the primary concept. Given a σ-finite measure μ on, for example, a field of sets A, it is extended uniquely to a measure on a σ-field as is done in the basic theory of measure (see 56). It is also used to define a product measure of μ with the ordinary measure on intervals (the product measure of a rectangle, with sides A in Ω and interval I in R is the product of the measures of its sides) and is then similarly extended on the corresponding σ-field in the

(57) The triplet (Ω, A, μ) in which μ is a measure on the σ-field A is a measure space. The integral $\int f d\mu$ of a finite simple function $f = \sum_{k=1}^{n} x_k I_{A_k}$ is defined as the (Eudoxus–Cauchy–Riemann–Stieltjes–Lebesgue–Radon) sum $\sum_{k=1}^{n} x_k \mu A_k$. Every non-negative measurable function f is the limit of a nondecreasing sequence of non-negative simple functions f_n. If one sets $\int f d\mu = \lim_n \int f_n d\mu$, then this limit always exists, but it may be infinite. It is independent of the choice of sequences $f_n \uparrow f$. Every measurable function f can be written as a difference of two non-negative measurable functions: $f = f^+ - f^-$ with $f^+(x) = f(x)$ or 0, and $f^-(x) = 0$ or $-f(x)$, according as $f(x) \geq 0$ or $f(x) < 0$. The integral $\int f du$ is the difference of the integrals of f^+ and of f^-, provided it exists; that is, is not of the form $+\infty - \infty$. The function f is μ-integrable if its integral exists and is finite. Since the absolute value $|f| = f^+ + f^-$, it follows that f is integrable if and only if $|f|$ is integrable. This is the abstraction of the constructive definition of the Lebesgue integral.
The importance and usefulness of this integration is due to its monotone convergence property, which means that for nondecreasing sequences of non-negative measurable functions f_n, passage to the limit and integration commute; in symbols, if $0 \leq f_n \uparrow$ as $n \to \infty$, then $\lim_n \int f_n = \int \lim_n f_n$. All other convergence properties of this integration, in constant use in analysis, follow from it.

(58) The indefinite integral ϕ, on measurable sets A, of a function f whose integral exists, is defined by $\phi(A) = \int_A f d\mu = \int f I_A \, d\mu$. This is the abstraction of the Lebesgue indefinite integral. Its fundamental property, which may serve as a descriptive definition of integration, was first obtained by Lebesgue, then generalized by Radon, and finally abstracted by O. Nikodym. It may be stated as follows: On a σ-finite measure space, a set function ϕ is an indefinite integral of one and only one function f but defined only up to values on a μ-null set if, and only if, ϕ is a difference of two measures and is σ-additive and μ-continuous; that is, $\phi(A) = 0$ whenever $\mu(A) = 0$. This property plays a basic role in the theory of measure and, especially, in the theory of probability.

product space $\Omega \times R$. Thus are formed measurable sets in Ω with the extended measure on them and measurable sets in $\Omega \times R$ with the corresponding product measure on these sets. The ordinate set of a function $f \geqslant 0$ is a set in $\Omega \times R$. The function is measurable when the ordinate set is measurable and the integral of f is the measure of its ordinate set. If f is not of constant sign, one writes $f = f^+ - f^-$ and proceeds as in the basic theory (see 57).

Daniell, in 1918, abstracted W.H. Young's approach of 1911, emphasizing the integral as the primary concept, was followed by the Polish mathematician Stefan Banach in 1937, the U.S. mathematician Marshall Harvey Stone in 1948, and the immortal Nicolas Bourbaki. Integration assigns numbers to functions f; that is, the integral is a functional. The main properties of this functional are that it is linear and non-negative, and, above all, it has the monotone convergence property (if $0 \leqslant f_n \uparrow f$ then the integral of f_n converges to the limit of f). These three properties suffice: A given functional with these properties on a space F of bounded functions that is a vector lattice (closed under linear combinations and undertaking *suprema* and *infima* two by two) may be called an integral on F.

The extension problem is to extend this integral on F to as large a family of functions as possible on which the integral still has the same properties. For example, the Cauchy–Riemann integral on the space F of all continuous functions on a bounded interval has the required properties, and the extension will yield the Lebesgue integral on the Lebesgue-integrable functions.

The extension procedure is as follows: First one extends F to the space U of limits f of nondecreasing sequences of functions f_n in F and then extends the integral by imposing the monotone convergence property so that by definition the integral of f is the limit of the integrals of f_n. This extension is unique. The integral can then be extended to the space $-U$ of negatives $-f$ of functions f, belonging to U, by imposing the linearity property. Finally, an arbitrary function can be approximated by functions g in $-U$ and h in U: $g \leqslant f \leqslant h$. The lower integral of f is defined to be the least upper bound of integrals of functions g and the upper integral of f to be the greatest lower bound of integrals of functions h. When the lower and upper integrals of f coincide and are finite, their common value is taken to be the integral of f, to be called an integrable function. This is primarily the abstraction of the Darboux approach in Lebesgue's context. The space of integrable functions, and the integral on it, has the desired properties. In particular, the integrals of integrable indicator functions of sets (equal to 1 on the set and 0 elsewhere) define a measure on these sets with the usual properties.

Introduction of Banach space

Since about 1930, when integrals of functions with values in a Banach space were introduced by the Polish-born mathematician S. Bochner, a further abstraction began. Measures and functions (hence, integrals) take values in more and more abstract spaces, instead of taking numerical values as above, and various types of convergence are used. Of particular importance are the various stochastic integrals. In fact, the foregoing indications concerning the theory of measure and integration have many variants, even in the numerical case. In the final analysis, their conceptual substance is the same, but it is seen from different points of view in various settings, from Eudoxus and Archimedes, through Leibniz and Riemann, to Lebesgue and the present day. The most important fact is the ever-spreading use of the powerful tools that these concepts provide for such areas as differentiation theory, boundary problems, ergodic theory, probability theory, group theory, potential theory, and harmonic analysis.

INTEGRAL CALCULUS

The definite integral. As mentioned earlier in this article, the integral is a limit concept. The geometrical aspects of this concept will now be explored. First, consider two real numbers, a and b, and a function $f(x)$ defined for all x satisfying $a \leqslant x \leqslant b$. For the sake of simplicity it will be further supposed, for the moment, that the values $f(x)$ are never negative. The basic problem of integral calculus can then be stated, based on Figure 6, as follows. What is the meaning (and numerical value) of the area of the region below the graph of $f(x)$ and above the interval (a, b) on the x-axis?

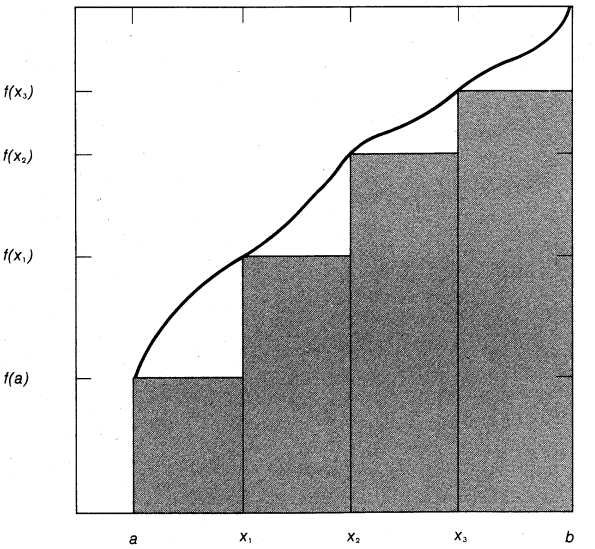


Figure 6: Representation of the area under a curve that underestimates the integral value.

Other basic assumptions are that it is known what is meant by "area" in case the region is a rectangle, and also that, if a region can be subdivided into a finite number of non-overlapping rectangles, then its area is the sum of the areas of these rectangles. Having made these assumptions, the question raised may be answered using the following:

First, the interval (a, b) is to be subdivided into n subintervals by choosing consecutive points $a = x_0, x_1, x_2, \cdots, x_n = b$. In each of the subintervals $I_i = (x_{i-1}, x_i)$, a point ζ_i $(i = 1, 2, \cdots, n)$ is chosen such that the value $f(\zeta_i)$ is smallest among the values assumed by the function in I_i (if such a point ζ_i exists). The rectangle with base I_i and height $f(\zeta_i)$ (see Figure 6, in which $n = 4$ and $\zeta_i = x_{i-1}$) then lies below the graph of $f(x)$. It is reasonable to assume, therefore, that the area in question should be a number exceeding the sum of the areas of the rectangles that have been constructed (that is, the area of the shaded region in Figure 6). This can be expressed algebraically (see 59) in terms of A, in which A denotes the area below the graph of $f(x)$ and above the interval (a, b).

Next, using a similar construction, instead of choosing the point ζ_i in the subinterval I_i, a point μ_i (if it exists) may be chosen such that the value $f(\mu_i)$ is the largest among the values assumed by the function in I_i (see Figure 7, in which $\mu_i = x_i$ and, again, $n = 4$). An inequality between the area A and a sum of rectangles can, again, be expressed algebraically (see 60).

$$(59) \qquad A \geqslant f(\zeta_1)(x_1 - x_0) + f(\zeta_2)(x_2 - x_1) + \cdots + f(\zeta_n)(x_n - x_{n-1})$$

$$(60) \qquad A \leqslant f(\mu_1)(x_1 - x_0) + f(\mu_2)(x_2 - x_1) + \cdots + f(\mu_n)(x_n - x_{n-1})$$

A sum of the type in which the smallest value of $f(x)$ in each subinterval I_i is used is called a lower sum (with respect to the given subdivision); a sum of the type in which the largest value of $f(x)$ in each subinterval is used is called an upper sum. With this terminology, it can be stated that the area in question should be a number, A, less than or equal to any upper sum and greater than or equal to any lower sum. If the notion of area as used here is to have a unique meaning, however, there should exist

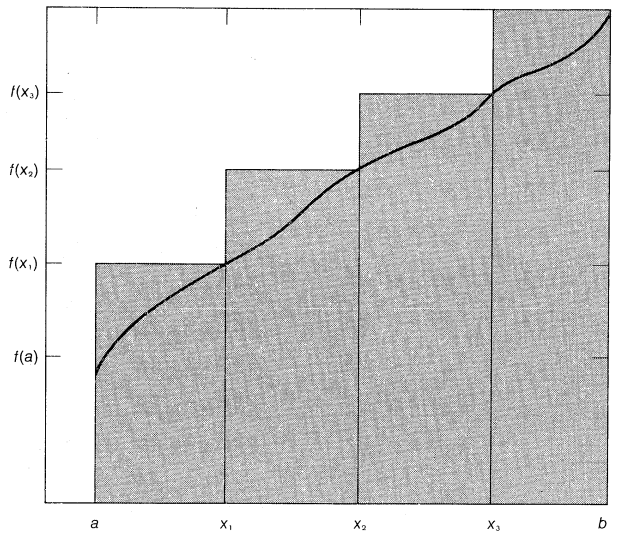


Figure 7: Representation of the area under a curve that overestimates the integral value.

only one number with this property. The following definition, therefore, is essential: If $f(x)$, defined for x satisfying the condition $a \leqslant x \leqslant b$, has the property that there exists exactly one number less than or equal to any upper sum and greater than or equal to any lower sum, then this number is the area of the region below the graph of $f(x)$ and above the interval (a,b).

It will be noted that the restriction $f(x) \geqslant 0$, which was made only in order to obtain a better grasp of the geometrical aspects of the construction, can be dropped without necessitating any change in the above definition. Dropping the restriction, however, implies that regions below the x-axis and above the graph of a negative-valued function will have negative areas.

The Riemann integral

The area that has just been defined is usually called the Riemann integral (or the definite integral, for reasons that will be shown) of the function $f(x)$ over the interval (a, b).

Integrable and nonintegrable functions. It is natural, at this point, to attempt to determine what classes of functions have a Riemann integral. It can be shown that if $f(x)$ is continuous (roughly speaking, this means that whenever x_1 and x_2 are close, then $f(x_1)$ and $f(x_2)$ are close), then it has an integral, or, more briefly, is (Riemann) integrable. In fact, this restriction is not only sufficient but comes close to being also a necessary condition for this type of integrability.

Two simple examples of functions that are not integrable are the following: Suppose $a = 0$, and $b = 1$. Let $f(x) = 1/x$ for $0 < x \leqslant 1$, and let $f(0) = 0$. Then, it can be seen that the upper sums of this function are not defined (for, in the first interval, I_1, of any subdivision, there is no number μ_1 such that the value $f[\mu_1]$ is the largest among the values that may be assumed by the function in I_1).

In the second case, again, let $a = 0$, and $b = 1$, and suppose $f(x) = 0$ if x is rational, but $f(x) = 1$ if x is irrational. It is then easy to check that each upper sum is 1 and each lower sum is 0. In this case there does not exist a unique number less than or equal to any upper sum but greater than or equal to any lower sum.

It is important to realize that the above definition of integrability is equivalent to the following one, which does not involve the concept of upper or lower sum: Subdivide the interval (a, b) as before and, again, form a sum (see 61) except that the only restriction on ζ_i $(i = 1, 2, \cdots, n)$ the point within the ith interval, is that it must belong to the interval I_i. If, as the subdivision is made finer (when the length of each subinterval approaches zero), the sum tends to a limit, then the function is integrable. This limit is the Riemann integral from a to b.

Fundamental theorem of calculus. The connection between the two basic concepts, the derivative and the integral, is perhaps the most important feature of calculus. This connection can be summarized roughly by saying that integration and differentiation are inverse operations; if first one of these operations and then the other is applied to a function, the original function is obtained (this is analogous to the operation of addition, and then subtraction, with a given number).

In order to describe this more precisely, the concept of the indefinite integral must be introduced. Suppose that $f(x)$ is continuous (hence, integrable) in some interval I, and let a and y be two points of this interval. If the number a is fixed, then a function of y, $F(y)$, can be formed by letting $F(y)$ be the integral of f from a to y; if $y < a$, $F(y)$ is equal to the negative of the integral from y to a. The function $F(y)$ is called an indefinite integral of the function $f(x)$.

The connection between differentiation and integration can then be stated in the following way:

The function $F(y)$ (as defined above) is differentiable, and $F'(y) = f(y)$; furthermore, if $G(y)$ is any other differentiable function satisfying $G'(y) = f(y)$ for all y in I, then the difference $F(y) - G(y)$ is constant.

This statement is known as the fundamental theorem of the integral calculus.

The anti-derivative

A function such as $G(y)$, satisfying $G'(y) = f(y)$, is called an anti-derivative of $f(y)$. If a constant is added to an anti-derivative of $f(y)$, another anti-derivative of the same $f(y)$ is obtained (because the derivative of a constant is zero). The fundamental theorem of the integral calculus can then be paraphrased by saying that the most general anti-derivative of $f(y)$ can be obtained by adding a constant function to an indefinite integral of $f(y)$.

$$(61) \qquad f(\zeta_1)(x_1 - x_0) + f(\zeta_2)(x_2 - x_1) + \cdots + f(\zeta_n)(x_n - x_{n-1})$$

$$(62) \qquad \sin y = \int_0^y \cos x \, dx$$

One of the immediate, and most important, consequences of the fundamental theorem is that the integrals of a large class of functions can be obtained merely by inverting the formulas already developed for differentiation. For example, it can be found that $\sin x$ is an anti-derivative of $\cos x$. Thus, $F(y)$ is the integral from 0 to y of the cosine function and must differ from $\sin y$ by a constant; that is, $F(y) = \sin y + C$. The constant C is easily determined, because it follows from the definition of the integral that $F(0) = 0$. Hence, $0 = F(0) = \sin 0 + C = C$. It has been shown, therefore, that the sine is the indefinite integral of the cosine, for all y (see 62).

The power of the fundamental theorem can be appreciated once it is realized that explicit formulas for integration, such as the formula for the integral of the cosine, are by no means easy to obtain by working directly from the definition of the integral. Unlike the case of differentiation, however, there are many functions that can be expressed by simple formulas but that do not have simple expressions for their anti-derivatives.

Conditions for existence of the Riemann integral. The question may now be raised: if a function f is bounded

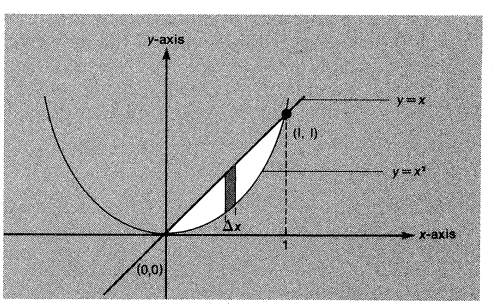


Figure 8: The curve $y = x^2$, the line $y = x$, and an element of area (strip) between them.

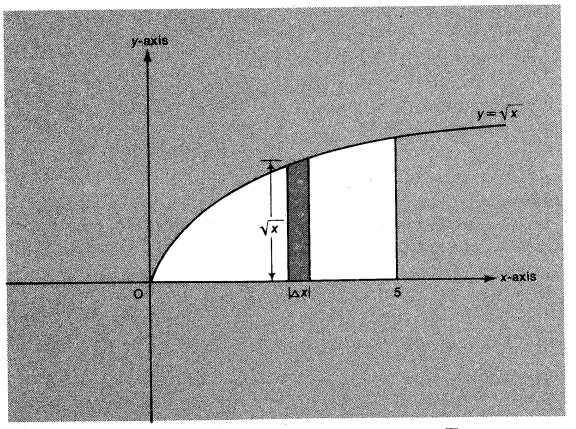


Figure 9: The plane region bounded the curve $y = \sqrt{x}$, the x axis, and the lines $x = 0$ and $x = 5$; when the plane figure is rotated about the x-axis to generate a solid of revolution, the strip of width Δx and height $\sqrt{x}$ generates a thin disk of radius $\sqrt{x}$ and thickness Δx, which is an element of volume.

on an interval $a \leqslant x \leqslant b$, can a criterion be found that will ensure the integrability, or nonintegrability, of f? (A function f is bounded on $a \leqslant x \leqslant b$ if there exists a number M such that $|f(x)| \leqslant M$ for all x in the interval. Boundedness, in effect, prevents the function from assuming arbitrarily large values at any points of the interval.)

This question can be concisely answered in terms of the concept of a set of points of measure zero. A set of points is of measure zero if the points can be covered by intervals the sum of whose lengths can be made arbitrarily small. For example, the set of points $x_0, x_1, x_2, \cdots, x_k, \cdots$, in which $x_0 = 0$, $x_k = 1/k$ $(k = 1, 2, 3, \cdots)$, is of measure zero because for any positive number ϵ (no matter how small) x_0 can be covered by an interval of length $\epsilon/2$; x_1 by an interval of length $\epsilon/2^2, \cdots$; x_k by an interval of length $\epsilon/2^{k+1}$; and by the theory of infinite series it can be shown that the sum of these lengths is precisely ϵ. It should be noted that any finite set of points is of measure zero.

Sets of discontinuity of measure zero

The following criterion provides a necessary and sufficient condition for Riemann integrability: If a function f is bounded on an interval $a \leqslant x \leqslant b$, then the Riemann integral of f on the interval exists if and only if the set of discontinuities of f is of measure zero. The result for piecewise-continuous functions is a special case of this theorem.

The function $f(x) = \sin [1/\sin (\pi/x)]$, for example, is not defined at the points 0, 1, ½, ⅓, $\cdots$. At these points the function can be assigned any values. The resulting function is bounded on the interval $0 \leqslant x \leqslant 1$, however, and its only discontinuities in this interval are the above points, which form a set of measure zero. Hence the integral of f from 0 to 1 exists.

A property that holds at all points of a given set, except for a subset of measure zero, is said to hold almost everywhere on the set. The above theorem therefore states, in effect, that for the Riemann integral to exist, the function f must be continuous almost everywhere on the specific interval. The extension of the concept of an integral to the Lebesgue integral is designed to cover cases that are not so restricted.

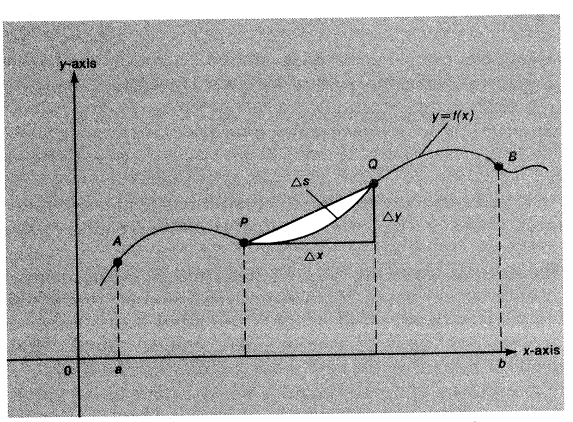


Figure 10: The arc length Δs from P to Q is approximated by the distance from P to Q, if these two points are sufficiently close together.

Improper integrals. In the discussion of the definite integral, the assumption was made that the limits of integration a, b are finite and that the integrand f was continuous or, at worst, bounded in the integral $a \leqslant x \leqslant b$. If f has infinite discontinuities or if a approaches $-\infty$ or if b approaches $+\infty$, the resulting integral is called improper. Examples are easily constructed (see 63, 64).

(63) To evaluate $\int_1^\infty \frac{dx}{x^2}$ it is first noted that $(-1/x)$ is an anti-derivative of $1/x$. Hence

$$\int_1^b \frac{dx}{x^2} = [-1/x]_1^b = -\frac{1}{b} + 1.$$

As b approaches infinity, $1/b$ approaches zero, so that by definition the integral equals

$$\int_1^\infty \frac{dx}{x^2} = \lim_{b\to\infty} \int_1^b \frac{dx}{x^2} = 1.$$

(64) The case of an infinite discontinuity is illustrated by the integral $\int_0^1 \frac{dx}{\sqrt{1-x}}$. The integral is discontinuous at $x = 1$ but is continuous on any interval $0 \leqslant x \leqslant b$ for any positive number $b < 1$. The integral is therefore defined as

$$\lim_{b\to 1} \int_0^b \frac{dx}{\sqrt{1-x}}.$$

An anti-derivative of $1/\sqrt{1-x}$ is $-2\sqrt{1-x}$. Hence the integral becomes

$$\int_0^b \frac{dx}{\sqrt{1-x}} = [-2\sqrt{1-x}]_0^b = -2\sqrt{1-b} + 2.$$

As b approaches one, $1-b$ approaches zero, so that $\int_0^1 \frac{dx}{\sqrt{1-x}} = 2$. It should be noted that, in either of the above cases, if the limits do not exist, then the corresponding integral does not exist.

(65) $$A = \left[\frac{x^2}{2} - \frac{x^3}{3}\right]_0^1 = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}$$

Applications. *Areas.* The method of computing areas under curves has already been indicated in the formal definition of integral. As a further illustration, the area bounded by the curve $y = x^2$ and by the line $y = x$ will be found (see Figure 8). The points of intersection of the curve and the line are (0, 0) and (1, 1). The strip of width Δx and height $x - x^2$ has an area approximated by the product $(x - x^2)\Delta x$. This is the element of area ΔA. Summation of these elements, and subsequent passage to the limit by letting Δx approach zero, leads to the formula for the area as the integral of $x - x^2$. An anti-derivative of $x - x^2$ is the function $x^2/2 - x^3/3$. Hence the integral becomes 1/6 (see 65).

Elements of area

Volumes of solids of revolution. If a plane region is rotated about a line in the plane, the resulting figure is called a solid of revolution. The volume of this solid is readily represented in the form of an integral. As an example, the volume generated by revolving the plane region bounded by the curve $y = \sqrt{x}$, the x-axis, and the lines $x = 0$ and $x = 5$ will be calculated (see Figure 9). When this plane figure is rotated about the x-axis, the strip of width Δx and height $\sqrt{x}$ generates a thin disk (or cylinder) of radius $\sqrt{x}$ and thickness Δx. The volume of this cylinder is approximately given by the area of the base times the thickness. Thus the element of volume of the solid of revolution is $\Delta V = \pi(\sqrt{x})^2\Delta x = \pi x \Delta x$. Sum-

(66) $V = \int \pi x\, dx = \pi\left[\frac{x^2}{2}\right]_0^5 = 12.5\pi$

(67) $V = \int_a^b \pi[f(x)]^2\, dx$

(68) General formula for the length of a curve:
$s = \int_a^b \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx$
The length of the curve $y = \frac{2}{3}x^{3/2}$, between $x = 0$ and $x = 3$, is given by $S = \int_0^3 \sqrt{1 + x}\, dx$. An anti-derivative of $\sqrt{1+x}$ is $\frac{2}{3}(1+x)^{3/2}$, so that $S = [\frac{2}{3}(1+x)^{3/2}]_0^3 = \frac{2}{3}(8-1) = 14/3$.

(69) $S = \int_a^b 2\pi y \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx$

(70) $S = \int_0^3 4\pi\sqrt{x}\sqrt{1 + (1/x)}\, dx$
$= 4\pi \int_0^3 \sqrt{1+x}\, dx = 56\pi/3$

(71) $W = \int_0^2 100x\, dx = 100\left[\frac{x^2}{2}\right]_0^2 = 200$ lb-in.

(72) $M_x = \int\int y\, dx\, dy$

(73) $\int_{\frac{1}{2}(y^2-1)}^{y+1} y\, dx = y \int dx$
$= y\left[x\right]_{\frac{1}{2}(y^2-1)}^{y+1} = y[(y+1) - \frac{1}{2}(y^2-1)]$

(74) $M_x = \int_{-1}^3 y[(y+1) - \frac{1}{2}(y^2-1)]\, dy$

mation of these elements leads to the formula for the volume of the solid: V (see 66). It follows from this example that, in general, if the equation of the curve is $y = f(x)$ and if the plane figure bounded by this curve, the x-axis, and the lines $x = a$ and $x = b$ is rotated about the x-axis, then the volume of the resulting solid can be expressed by an integral (see 67).

Lengths of curves. The arc length Δs from P to Q (see Figure 10) is approximately given by the distance from P to Q if these two points are sufficiently close to each other. Thus, the element of arc length Δs is approximately $\sqrt{(\Delta x)^2 + (\Delta y)^2}$, which can also be written in a form with the common factor Δx. Summation of these arc lengths, and passage to the limit, yields the formula for the length of the curve between the points A, B (see 68), in which a and b are the abscissas of the points A, B. The length of a specific curve is thus obtained by calculation (see 68).

Areas of solids of revolution. If the plane region of Figure 10 is rotated about the x-axis, the lateral area of the thin slab bounded by the segment PQ is given by the circumference of the cross section times the slant height; $2\pi y$ times length of PQ. Since the arc length from P to Q is approximated by the length of the segment from P to Q, the element of surface area of the solid of revolution is given by $\Delta S = 2\pi y \Delta s$. With Δs approximated by $\sqrt{(\Delta x)^2 + (\Delta y)^2}$, subsequent passage to the limit gives the surface area (see 69). For example, if $y = 2\sqrt{x}$, $a = 0$, and $b = 3$, then $s = 56\pi/3$ (see 70).

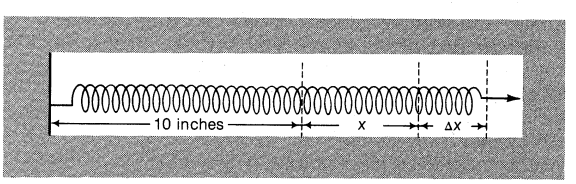


Figure 11: A 10-inch spring displaced to length $10 + x + \Delta x$.

Work done by a force. If a constant force F is directed along the x-axis and acts on a particle, displacing it through a distance x, then the work done by the force is given by the product of the force and the distance: $W = Fx$. If the force is not constant, the formula for the work done is represented by an integral. The element of work ΔW is given by $F(x)\Delta x$, in which $F(x)$ is the value of the force at x, and Δx is a small displacement from x to $(x + \Delta x)$. Summation of the elements of work, followed by passage to the limit, leads to the formula in which W equals the integral from a to b of $F(x)$.

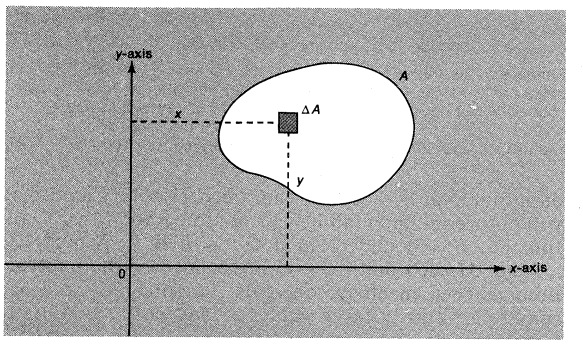


Figure 12: For a plane area A, the moment of an element of area ΔA about the x-axis is $y\Delta A$, and the moment about the y-axis is $x\Delta A$.

Work required to stretch a spring

For example, the force required to stretch a spring is proportional to the amount of stretch: $F(x) = kx$ (see Figure 11). For a certain spring it is found that a force of 50 pounds is required to extend it by 0.5 inch. The constant of proportionality k, therefore, has the value 100, and $F(x) = 100x$. If the unstretched length of the spring is 10 inches the work required to stretch the spring to a length of 12 inches is $W = 200$ lb-in. (see 71).

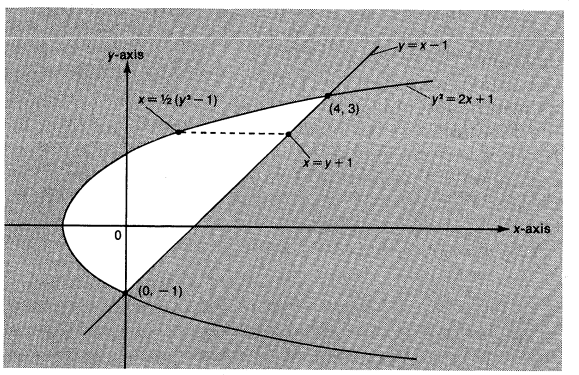


Figure 13: The area bounded by the curve $y^2 = 2x + 1$ and the line $y = x - 1$. The points of intersection are $(0, -1)$ and $(4, 3)$.

Moments and centroids. The moment of an element of area about an axis (perpendicular to the plane of the area) is defined as the product of the element of area by its distance from the axis. In particular, the moment of ΔA about the x-axis is $y\Delta A$, and the moment about the y-axis is $x\Delta A$ (see Figure 12). Summation of these elements over the whole area and a passage to the limit lead to formulas for the moment of the plane area A expressed as double integrals: the moment about the x-axis is the double integral of y with respect to A; the moment about the y-axis is the double integral of x with respect to A.

The evaluation of these integrals is achieved by the computation of successive single integrals whose limits are determined by the boundary of the region. The centroid of the plane area A is defined as the point (x, y) at which $x = M_y/A$, $y = M_x/A$. For example, if the area is bounded by the curve $y^2 = 2x + 1$ and the line $y = x - 1$ (see Figure 13), the points of intersection are $(0, -1)$ and $(4, 3)$. Using the methods previously discussed, the area is found to be $A = 16/3$. With the element of area written as $\Delta x \Delta y$, the moment about the x-axis can be written (see 72) in which the inner integral is taken with respect to x, keeping y fixed. From the diagram it is seen that the lower and upper limits of integration for this integral are $x = \frac{1}{2}(y^2 - 1)$, and $x = y + 1$. Thus, the inner integral is calculable (see 73). The moment is also obtainable (see 74), in which the limits are determined by the points of intersection of the curve and straight line. Evaluation of the integral yields $M_x = 16/3$.

A similar evaluation for the moment about the y-axis gives $M_y = 129/20$. The centroid is therefore located at (387/320, 1). (Hy.K.)

BIBLIOGRAPHY. The history of the calculus is well covered in OTTO TOEPLITZ, *Die Entwicklung der Infinitesimalrechnung* (1949; Eng. trans., *The Calculus*, 1963); CARL B. BOYER, *The Concepts of the Calculus* (1949; reprinted as *The History of the Calculus and Its Conceptual Development*, 1959); and ARTHUR ROSENTHAL, "The History of Calculus," *Am. Math. Mon.*, 58:75–86 (1951). A good introduction may be found in RICHARD COURANT and HERBERT ROBBINS, *What Is Mathematics?* (1941). Many excellent texts at all levels of difficulty are available. Two interesting examples are SHERMAN K. STEIN, *Calculus in the First Three Dimensions* (1967); and MORRIS KLINE, *Calculus: An Intuitive and Physical Approach* (1967). For modern theories of measure and integration, the following may be consulted: STANISLAW SAKS, *Théorie de l'intégrale* (1933; Eng. trans., *Theory of the Integral*, 2nd rev. ed., 1964); ANGUS E. TAYLOR, *General Theory of Functions and Integration* (1965); EDGAR ASPLUND and LUTZ BUNGART, *A First Course in Integration* (1966). These all require a good foundation in calculus. Additional references are: WATSON FULKS, *Advanced Calculus: An Introduction to Analysis*, 2nd ed. (1969); A.N. KOLMOGOROV and S.V. FOMIN, *Measure, Lebesgue Integrals, and Hilbert Space* (1961; orig. pub. in Russian, 1954); MARSHALL E. MUNROE, *Calculus* (1970) and *Introductory Real Analysis* (1965); H.L. ROYDEN, *Real Analysis* (1963); WALTER RUDIN, *Principles of Mathematical Analysis*, 2nd ed. (1964); and G.H. HARDY, *A Course of Pure Mathematics*, 10th ed. (1958).

(Hy.K./Ed.)

Analysis, Vector and Tensor

Many physical quantities, such as temperature, pressure, and mass, have only magnitude. The numbers representing them are called scalars and obey the simple rules of arithmetic. On the other hand, quantities such as velocity, force, or position in two or more dimensions relative to a fixed point have both magnitude and direction and can be represented mathematically by two or more numbers that constitute what is called a vector. Geometrically, a vector is drawn as a line in a specific direction with a length equal to its magnitude (Figure 1).

Resolution of forces by vector parallelogram

The earliest knowledge of how to manipulate vector quantities dates from the time of Galileo, who first formulated the simple parallelogram law for combining forces in the 17th century. If two forces act on a body, their resultant force can be obtained from a parallelogram in which two adjacent sides have lengths and directions representing the forces. The two forces acting together are equivalent to a single force represented, in magnitude and direction, by the diagonal of the parallelogram. This is equivalent to the rule that two or more vectors can be added by putting their geometrical representation end to end, so to speak.

A vector can be multiplied by a positive number by simply multiplying its magnitude by that number (it has the same direction). To multiply a vector by a negative number its direction is reversed and its magnitude multiplied. The process of multiplication by numbers is called scalar multiplication. An example of its application to physical quantities is the process of multiplying the mass of a body by its velocity to give its momentum. The momentum of a body is, thus, a vector quantity having the same direction as the velocity and a magnitude equal to the product of the magnitudes of its mass and its velocity.

Expression of vectors in three-dimensional space

Vectors in three dimensions may be expressed in the following way. A set of three mutually perpendicular vectors each of unit length is chosen, and any vector may be represented uniquely as a combination of these. If these vectors are written **i**, **j**, and **k** (vectors in physical applications are often printed in boldface type), then any vector can be represented by $a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}$. The numbers a_1, a_2, and a_3 are called the components of the vector. By Pythagoras' theorem, the length of $a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}$ is the square root of the sum of the squares of the components. The magnitude of an arbitrary vector **v** is written $\| \mathbf{v} \|$.

Vectors may be multiplied together in two ways. The scalar product of two vectors (also called the dot product) is, in two dimensions only, defined to be the product of their lengths multiplied by the cosine of the angle between their directions. A physical example concerns the work done, if a force **F** moves in a straight line that has magnitude and direction **d**. If the force moves along the direction of **d**, the work is $\| \mathbf{F} \| \, \| \mathbf{d} \|$; if it acts at an angle θ to that direction it is $\| \mathbf{F} \| \, \| \mathbf{d} \| \cos \theta$. If the force acts at right angles to the direction the work done is zero.

In three dimensions two vectors can be defined to give a third. The vector product of **a** and **b** is written as $\mathbf{a} \times \mathbf{b}$, and is also called the cross product. It is a vector perpendicular to **a** and **b** with length $\| \mathbf{a} \| \, \| \mathbf{b} \| \sin \theta$, θ being the angle between them. When an electric charge moves in a magnetic field, for example, it experiences a force that is mutually perpendicular to the field and the direction of motion, hence the principal from which it is inferred that the wire moving in the magnetic field of a generator will carry electrical current.

Many physical quantities take the form of mappings (or correspondences) from one collection of vectors to another. For example, the stress at a point in a fluid is the relation between two vectors. The stress is an example of a tensor, which has a general mathematical definition that will be presented below. It will be seen that the tensor concept depends upon ideas involving more than one vector—in fact, collections of vectors that are called vector spaces. The idea of vectors and tensors was suggested by particular needs in problems of mathematical physics.

Once the concepts of vector and tensor are known from the study of particular physical or mathematical questions, it becomes natural, by the route of mathematical reasoning, to express individual vectors or tensors as members of an appropriate space. By doing so, a setting of generality is established and the properties of these entities can be studied without reference to particular representatives that might have been encountered first historically or that might have been important in special cases. Therefore, with the advantage of historical maturity, discussion will begin here with the treatment of a vector space. (Ed.)

VECTOR ALGEBRA AND ANALYSIS

Vector space: history and basic ideas. The germinal idea of an abstract vector space that is not limited to applications grew out of investigations of the mathematicians Sir William Rowan Hamilton (Irish) and Hermann Günther Grassmann (German) during the early 1840s. In that era mathematicians still regarded complex numbers that involve the square root of -1 with a mixture of mysticism and suspicion, although by 1800 Caspar Wessel of Norway had suggested representing the complex number $a_1 + ia_2$, in which i is the square root of -1, concretely by a point in a plane (specifically, by the point $[a_1, a_2]$ so that a_1 and a_2 become distances to the point measured along straight lines parallel to two rectilinear axes). In 1831 Germany's leading mathematician Carl Friedrich Gauss showed that complex numbers can be regarded as ordered pairs of real numbers for which the operations of addition and multiplication are defined

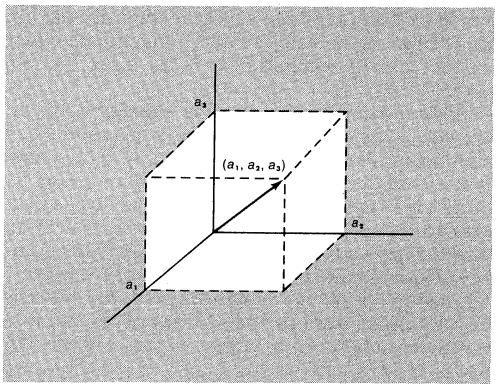


Figure 1: A vector regarded geometrically as a directed line segment (or arrow) from the origin to a point with coordinates a_1, a_2, a_3 (see text).

(see Box, equations 1) in terms of the component parts of the pairs. The resulting algebra of pairs is a perfect replica of the algebra of complex numbers, without any reference to complex numbers or to the symbol i. This algebra proved to be useful in describing rotations in the plane. (An algebra is a collection of entities for which suitable operations are defined.)

Unaware of Gauss's work, Hamilton discovered the same device six years later and continued to search for a comparable algebraic description of rotations in three-dimensional space. His early efforts were thwarted, however, by an algebraic dogma that insisted on certain familiar properties of addition, subtraction and multiplication growing out of centuries of exclusive concern with familiar numbers. Each of the number systems known was an example of an algebraic system that today is called a field. Formally, a field is a set of elements and two operations, called addition and multiplication, satisfying well-defined postulates, as discussed fully in the article ALGEBRAIC STRUCTURES.

Hamilton's modification of the field postulates

Hamilton's eventual success came from his willingness to discard the notion that any useful and consistent algebraic system must satisfy all of the field postulates. His new algebra provided a means of describing rotations in three-dimensional space, by an algebraic system that satisfies all of the field postulates except the commutative law for multiplication, which identifies the products $a \cdot b$ and $b \cdot a$ as equal. In Hamilton's algebra $a \cdot b = -b \cdot a$. Thus Hamilton provided the first example of a noncommutative field, which is now called a division ring. The elements of his algebra were called quaternions.

The realization that useful algebraic systems could be constructed by replacing some traditional postulates proved to be as liberating to algebra as the corresponding discovery had been to geometry only a few years before, when the parallel postulate of Euclidean geometry had been replaced by Gauss and the Hungarian and Russian discoverers of non-Euclidean geometry, János Bolyai and Nikolay Ivanovich Lobachevsky. Mathematical heresy apparently was in the air in those days. In 1844 Grassmann independently developed a theory that was far more imaginative and comprehensive than Hamilton's, providing an algebraic mechanism for n-dimensional space and allowing for various definitions of multiplication. At the same time in England Arthur Cayley invented matrix algebra, which includes the quaternions invented by Hamilton as a very special case.

The contemporary concept of a vector is a natural or closely related abstraction and extension of the observation that an ordered pair of real (not complex) numbers (a_1, a_2) represents a unique point in the Cartesian plane (that for which pairs of numbers are identified with numbers as described above); similarly an ordered triple of numbers (a_1, a_2, a_3) represents a point in three-dimensional Cartesian space (Figure 1), and an ordered n-tuple of numbers $(a_1, a_2, \cdots, a_n)$ represents a point in Cartesian n-dimensional space.

Addition of two n-tuples is defined by the parallelogram law, which states (see 2) that components of the n-tuples are added to form components of a new n-tuple called the sum. Multiplication of an n-tuple by a number is defined by extending the multiplication (by the number) to each component (see 3). The laws of addition and multiplication are illustrated in Figures 2 and 3.

The resulting algebraic system is a particular example of a vector space, and the example is general enough to suggest an axiomatic formulation of the concept of a vector space. To begin with, observe that two types of objects are involved in this example: numbers (which are called scalars and will be denoted here by lower case letters) and n-tuples of numbers (which are called vectors and will be denoted here by capital letters). Normally the set of scalars is assumed to be a field, although a comparable theory of vector spaces exists for the case in which the scalars form only a division ring. It need not be assumed that a vector is an n-tuple of scalars; in general, the set V of vectors can be any set of objects that can be combined according to specific algebraic laws as listed in ALGEBRA, LINEAR AND MULTILINEAR. Because the laws make specific references to a field of scalars, the term vector space over a field is used. Two of the most important fields that appear in this context are the field $\mathfrak{R}$ of real numbers and the field $\mathfrak{C}$ of complex numbers.

In the familiar notation used for many years for vectors in three-dimensional real space $\mathfrak{R}^3$, the vectors along the three mutually perpendicular axes are denoted $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$. Then an arbitrary vector $\mathbf{v}$ is written $\mathbf{v} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$, thus designating the point (x, y, z) in $\mathfrak{R}^3$. This particular basis is called the standard basis for $\mathfrak{R}^3$, and much of the late-19th-century work in vector analysis and vector calculus, developed by the U.S. physicist J. Willard Gibbs and the English physicist Oliver Heaviside, was expressed in this form. One of the great advantages of the abstract approach to vector space algebra is that it emphasizes coordinate-free methods and thereby avoids long arithmetic computations for properties that are intrinsic, and therefore are independent of the choice of coordinates.

The standard basis for $\mathfrak{R}^3$

$$(1)\quad \begin{cases}(a_1, a_2)+(b_1, b_2)=(a_1+b_1, a_2+b_2)\\ (a_1, a_2)\cdot(b_1, b_2)=(a_1b_1-a_2b_2, a_1b_2+a_2b_1)\end{cases}$$

$$(2)\quad (a_1, a_2, \cdots, a_n)+(b_1, b_2, \cdots, b_n) = (a_1+b_1, a_2+b_2, \cdots, a_n+b_n)$$

$$(3)\quad c(a_1, a_2, \cdots, a_n)=(ca_1, ca_2, \cdots, ca_n)$$

It is important to observe that the definition of a vector space describes only the algebraic operations of vector sum and scalar multiple of a vector. Special types of vector spaces can be obtained by superimposing additional operations on a vector space. The most glaring omission up to this point concerns metric concepts or concepts that define distance. In itself a vector space provides no means for measuring distance, or angle, or area, or any of the usual concepts needed for geometry and most applications. Therefore, in particular, the reference in the previous paragraph to mutually perpendicular axes was intuitive and premature.

Products: scalar and vector. To attribute to a vector space $\mathfrak{L}$ the usual metric properties, a way must be found to define the length of a vector and the angle between vectors. Length is by nature a function from $\mathfrak{L}$ into the nonnegative real numbers, in the same sense that to each vector in $\mathfrak{L}$ there is assigned a number, the length. Angle is a function in a similar sense from the space of all ordered pairs of vectors into the real numbers between 0 and π. Therefore, the real numbers are needed for such computations; for simplicity the scalar field can be chosen to be the complex numbers $\mathfrak{C}$, which includes the real scalar field as a special case. Accordingly, the complex conjugate $\bar{z}$ of a complex number $z = x + iy$ is

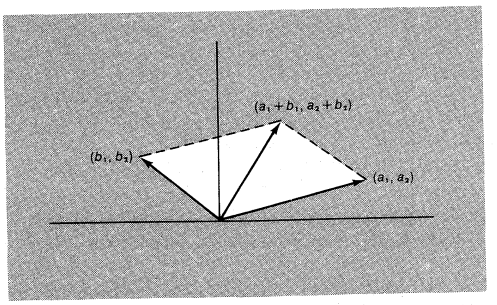


Figure 2: Addition of two n-tuples (see text).

defined by $\bar{z} = x - iy$. It follows that $\overline{z_1 + z_2} = \bar{z}_1 + \bar{z}_2$, and $\bar{z}_1\bar{z}_2 = \bar{z}_1\,\bar{z}_2$. The magnitude $|z|$ of z is defined by $|z| = \sqrt{x^2 + y^2}$.

If a vector space $\mathfrak{L}$ is defined so that vectors multiply with members of the field $\mathfrak{C}$ of complex numbers, a scalar product (or inner product) on $\mathfrak{L}$ is a function p with certain properties that maps pairs of vectors into complex numbers. The properties of the function p are those that define the inner product (see 4). Examples of inner products are the following: First, on the space of n-tuples of complex numbers (see 5), p maps the pair, $(a_1, \cdots, a_n)$, $(b_1, \cdots, b_n)$, into $a_1b_1 + \cdots + a_nb_n$. Second, on the space of n-tuples of real numbers (see 6), the value of the function at $(a_1, \cdots, a_n)$, $(b_1, \cdots, b_n)$ becomes $a_1b_1 + \cdots + a_nb_n$. Third, on the infinite-dimensional space of complex valued functions continuous on the interval of real numbers between a and b (see 7) the pair of functions (f, g), is mapped into the limit of a sum, namely, the value of an integral.

Examples of inner products

The central fact about any scalar product is expressed in the Schwarz inequality (see 8). The length $\| A \|$ of a vector A is defined (see 9) in terms of inner product, and it has the usual properties of Euclidean length (see 10). Distance between points (vectors) A and B is then defined in terms of length (see 11). Distance is symmetric—that is, independent of the order of the vectors in the pair (A, B)—positive if $A \neq B$, and satisfies the triangle inequality (see 12).

In the case of the space of all real n-tuples with the inner product defined as in the examples above, these concepts of length and distance coincide with the usual Euclidean distance. Hence, any real inner product space is called a Euclidean space. Any complex inner product space is called a unitary space.

A major distinction between Euclidean and unitary spaces occurs in the definition of angle. In either case the Schwarz inequality assumes the same form (see 13). In the Euclidean case there is precisely one number $\psi(A, B)$ the cosine of which corresponds to the intuitive notion of the cosine between two directed line segments (see 14). Thus $\psi(A, B)$ is defined to be the angle between vectors A and B. In a unitary space $p(A, B)$ is a complex number, in general, and so the real definition of angle cannot apply. The important case of perpendicularity (orthogonality) of A and B, however, can be defined by $p(A, B) = 0$ in a unitary space, just as in a Euclidean space.

The standard basis in terms of which all other vectors can be expressed for n-dimensional Euclidean or unitary space is a set of n vectors, each of unit length and each pair orthogonal. Such a basis is called orthonormal. Relative to an orthonormal basis, any inner product on an n-dimensional space assumes the first form given in the examples above (5). This form is also called the dot product because of the notation frequently used, $A \cdot B = p(A, B)$, in which $A = (a_1, \cdots, a_n)$ and $B = (b_1, \cdots, b_n)$ are expressed relative to an orthonormal basis for n-dimensional unitary space. The same form is valid in the Euclidean case, because $\bar{b} = b$ whenever b is real. The metric concepts assume familiar forms (see 15).

Quite a different form of product of vectors is used extensively in studying the geometry of Euclidean three-dimensional space and its physical applications. As contrasted with the scalar (or dot) product $A \cdot B$, which is a scalar, the vector (or cross) product $A \times B$ is a vector. Geometrically, $A \times B$ is the vector that, first, is perpendicular to both A and B; second, that is of length $\| A \| \, \| B \| \sin \psi(A, B)$, which numerically equals the area of the parallelogram having A and B as adjacent sides; and, third, that extends from the origin in the direction of travel of a right-hand screw when it is rotated from A to B through ψ, which is the smaller of the two possible angles.

Geometrical meaning of vector cross product

An expression for $A \times B$ in terms of components is essential for computations and precise reasoning (see 16). The basic algebraic properties satisfied by $A \times B$ are important to compare with those of the inner product (see 17). The first of these properties shows that a cross product is anticommutative (or alternating), as first described by Hamilton for quaternions. The last of these properties relates $A \times B$ to the Schwarz inequality.

Unlike the concept of scalar product, the existence of a vector product having the properties listed above depends critically upon the dimension n of the vector space on which it is defined. Except for the trivial cases ($n = 0$ and $n = 1$), such a vector product exists only for $n = 3$ and $n = 7$. This odd fact is closely related to long-stand-

(4) For all A, B in $\mathfrak{L}$ and all c_1, c_2 in $\mathfrak{C}$:
$$\begin{cases} p(c_1A_1 + c_2A_2, B) = c_1p(A_1, B) + c_2p(A_2, B), \\ p(B, A) = \overline{p(A, B)}, \\ p(A, A) \text{ is a positive real number if } A \neq O. \end{cases}$$

(5) $$p((a_1, \cdots, a_n), (b_1, \cdots, b_n)) = a_1\bar{b}_1 + \cdots + a_n\bar{b}_n$$

(6) $$p((a_1, \cdots, a_n), (b_1, \cdots, b_n)) = a_1b_1 + \cdots + a_nb_n$$

(7) $$p(f, g) = \int_a^b f(x)\overline{g(x)}\, dx$$

(8) $$|p(A, B)|^2 \leq p(A, A)p(B, B)$$

(9) $$\|A\| = \sqrt{p(A, A)}$$

(10) $$\begin{cases} \|cA\| = |c| \, \|A\| \\ \|A\| > 0 \text{ if } A \neq O \\ \|A + B\| \leq \|A\| + \|B\| \end{cases}$$

(11) $$d(A, B) = \|A - B\|$$

(12) $$d(A, B) \leq d(A, C) + d(C, B)$$

(13) $$\frac{|p(A, B)|}{\|A\| \, \|B\|} \leq 1 \text{ if } A \neq O \neq B$$

(14) $$\cos\{\psi(A, B)\} = \frac{p(A, B)}{\|A\| \, \|B\|} \text{ for } 0 \leq \psi(A, B) \leq \pi$$

(15) $$\begin{cases} \|A\| = \sqrt{a_1^2 + \cdots + a_n^2} \\ d(A, B) = \sqrt{(a_1 - b_1)^2 + \cdots + (a_n - b_n)^2} \\ A \cdot B = \|A\| \, \|B\| \cos\{\psi(A, B)\} \end{cases}$$

(16) $$(a_1, a_2, a_3) \times (b_1, b_2, b_3) = (a_2b_3 - a_3b_2, a_3b_1 - a_1b_3, a_1b_2 - a_2b_1)$$

(17) $$\begin{cases} A \times B = -B \times A \\ (A + B) \times C = (A \times C) + (B \times C) \\ (cA) \times B = c(A \times B) \\ A \cdot (A \times B) = 0 \\ \|A \times B\|^2 = \|A\|^2 \, \|B\|^2 - (A \cdot B)^2 \end{cases}$$

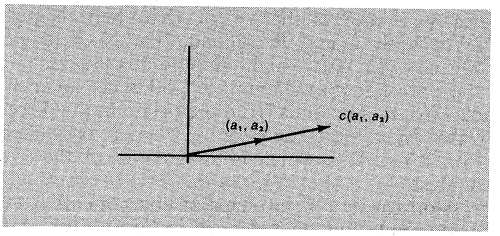


Figure 3: Multiplication of an n-tuple by a number c (see text).

ing problems in algebra, originating in the work of Hamilton and Cayley but only fully solved in the 1960s. To describe those problems it is necessary first to state what is meant by an algebra (see also ALGEBRA, LINEAR AND MULTILINEAR).

Algebras: non-associative, non-commutative, others. An algebra $\mathfrak{A}$ over a field $\mathfrak{F}$ is a finite-dimensional vector space over $\mathfrak{F}$ for which there is also defined a product of vectors such that for all A, B, C in $\mathfrak{A}$ and all a, b in $\mathfrak{F}$, three properties hold (see 18, 19, 20). (In contemporary usage a more general definition of an algebra is obtained by taking $\mathfrak{A}$ to be a module over a commutative ring with unity, with a module and a module product as described in ALGEBRAIC STRUCTURES.) Observe that the product is not assumed to be either commutative or associative. Various special types of algebras can be obtained by imposing additional conditions on the product operation in $\mathfrak{A}$. For example, $\mathfrak{A}$ is called an associative algebra if, and only if, a certain relation holds (see 21), which identifies the product of three vectors, irrespective of the order of operations within the triple product. If the relation (see 21) is not assumed, $\mathfrak{A}$ is called a non-associative algebra, which means that $\mathfrak{A}$ is not necessarily associative. An algebra in which the associative relation (21) is known to be invalid is a not-associative algebra.

An algebra $\mathfrak{A}$, in which another relation holds that identifies the products of two terms irrespective of order (see 22), is called a commutative algebra. If the commutative relation (22) is not assumed, $\mathfrak{A}$ is called a non-commutative algebra. If $\mathfrak{A}$ satisfies still another law involving a change of algebraic sign when the order of the product is reversed (see 23), then $\mathfrak{A}$ is noncommutative and is called an alternating algebra. Thus, three-dimensional Euclidean space, with the vector product $A \times B$, forms an alternating, not-associative algebra.

To see how other examples of such algebras can arise, let $\mathfrak{A}$ be an associative algebra over $\mathfrak{F}$. A new product can be defined on $\mathfrak{A}$ by the formula $[A, B] = AB - BA$. Here the symbol $[A, B]$ denotes the new product of A and B, whereas AB denotes the original product in $\mathfrak{A}$. The vector space $\mathfrak{A}$, together with the product $[A, B]$, forms an algebra that is denoted by $\mathfrak{M}$. Because $[A, B] = -[B, A]$, $\mathfrak{M}$ is an alternating non-associative algebra. Multiplication satisfies an identity attributed to the 19th-century German mathematician Karl Gustav Jacob Jacobi (see 24).

Lie algebras

Any alternating algebra that satisfies the Jacobi identity is called a Lie algebra in recognition of the pioneering work of the 19th-century Norwegian mathematician Sophus Lie. The Poincaré-Birkhoff-Witt theorem demonstrates that every Lie algebra is isomorphic (or equivalent in some sense) to a subalgebra of a Lie algebra obtained in this manner from an associative algebra, thus showing the central role held by algebras obtained this way and by Lie algebras.

Another important class of non-associative algebras was discovered by the 20th-century physicist Pascual Jordan, of Germany, in an investigation of problems in quantum mechanics. A Jordan algebra $\mathfrak{H}$ is a finite-dimensional vector space on which is defined a product $\{A, B\}$ satisfying two identities that lead to the consequence that $\mathfrak{H}$ is commutative (see 25). If one replaces the product AB of an associative algebra by the Jordan product $\{A, B\} = AB + BA$, the resulting system is a Jordan algebra. Unlike Lie algebras, however, not every Jordan algebra can be obtained in this way from an associative algebra.

Still another type of algebra can be obtained by letting $\mathfrak{A}$ be an algebra over $\mathfrak{R}$, and v a function from $\mathfrak{A}$ to $\mathfrak{R}$ satisfying four basic relations that are suggested by the properties of absolute value, as defined on real numbers (see 26). The norm of A is defined by $n(A) = v^2(A)$, and the pair $(\mathfrak{A}, n)$ is called a normed algebra.

The real numbers trivially form a normed algebra over $\mathfrak{R}$ in which $n(a) = a^2$. In particular, $n(a)n(b) = a^2b^2 = (ab)^2 = n(ab)$. For the complex numbers, let $v(A) = \sqrt{a_1^2 + a_2^2}$, in which $A = a_1 + ia_2$. It follows (see 27) that again $n(A)n(B) = n(AB)$, with AB of the form (c_1, c_2) and $n(AB)$ of the form $c_1^2 + c_2^2$. These observations suggest the following problem: given vectors $A = (a_1, \cdots, a_m)$ and $B = (b_1, \cdots, b_m)$ in an algebra of dimension m over $\mathfrak{R}$, does there always exist a vector $C = (c_1, \cdots, c_m)$ such that $AB = C$ and $(a_1^2 + \cdots + a_m^2)(b_1^2 + \cdots + b_m^2) = (c_1^2 + \cdots + c_m^2)$? Hamilton provided an affirmative answer for the case $m = 4$ with his noncommutative normed algebra of quaternions. At almost the same time (1845) Cayley produced an example for $m = 8$, a noncommutative, non-associative algebra over $\mathfrak{R}$, called Cayley numbers.

All of these four examples are normed algebras having a unity element. They also are alternative algebras, meaning that multiplication satisfies a weakened form of the associative law (see 28). In addition, all are division algebras, meaning that for any A, B in $\mathfrak{A}$, there exist vectors X, Y in $\mathfrak{A}$ such that $AX = B$ and $YA = B$. Any normed algebra over $\mathfrak{R}$ is a division algebra.

Division algebras

Attempts to find other solutions to this problem have generated outstanding research in this branch of abstract algebra for more than a century. In 1878 Ferdinand Georg Frobenius of Germany proved that the real numbers, complex numbers, and quaternions are the only finite-dimensional associative division algebras over the reals. Twenty years later Adolph Hurwitz, a Swiss mathematician, showed that the only finite-dimensional normed algebras over $\mathfrak{R}$ are of dimension 1, 2, 4, or 8. This result was strengthened by the U.S. algebraist Abraham Adrian Albert in 1947 to show that the only finite-dimensional normed algebras over $\mathfrak{R}$ are essentially the four known examples. Three years later in the United

(18) AB is in $\mathfrak{A}$

(19) $(aA)B = a(AB) = A(aB)$

(20) $\begin{cases} (aA + bB)C = aAC + bBC \\ C(aA + bB) = aCA + bCB \end{cases}$

(21) $A(BC) = (AB)C$ for all A, B, C in $\mathfrak{A}$

(22) $AB = BA$ for all A, B in $\mathfrak{A}$

(23) $AB = -BA$ for all A, B in $\mathfrak{A}$

(24) $[[A, B], C] + [[B, C], A] + [[C, A], B] = 0$

(25) $\begin{cases} \{A, B\} = \{B, A\} \\ \{\{A, A\}, \{B, A\}\} = \{\{\{A, A\}, B\}, A\} \end{cases}$

(26) $\begin{cases} v(A) > 0 \text{ if } A \neq O \\ v(cA) = |c| v(A) \\ v(AB) = v(A)v(B) \\ v(A + B) \leq v(A) + v(B) \end{cases}$

(27) $\begin{cases} n(A)n(B) = (a_1^2 + a_2^2)(b_1^2 + b_2^2) = (c_1^2 + c_2^2) = n(AB) \\ c_1 + ic_2 = AB \end{cases}$

(28) $A(AB) = (AA)B$ and $(BA)A = B(AA)$ for all A, B in $\mathfrak{A}$

(29) $\mathbf{r} = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k}$

(30) $\dfrac{d\mathbf{r}}{dt} = \dfrac{dx}{dt}\mathbf{i} + \dfrac{dy}{dt}\mathbf{j} + \dfrac{dz}{dt}\mathbf{k} = \mathbf{v}$

(31) $\dfrac{d^2\mathbf{r}}{dt^2} = \dfrac{d^2x}{dt^2}\mathbf{i} + \dfrac{d^2y}{dt^2}\mathbf{j} + \dfrac{d^2z}{dt^2}\mathbf{k} = \mathbf{a}$

(32) $\overrightarrow{PP'} = \Delta\mathbf{r}$

(33) $\dfrac{u(P') - u(P)}{|\Delta\mathbf{r}|}$

(34) $\nabla u = \mathbf{i}\dfrac{\partial u}{\partial x} + \mathbf{j}\dfrac{\partial u}{\partial y} + \mathbf{k}\dfrac{\partial u}{\partial z}$

(35) $\dfrac{\int_S \mathbf{v} \cdot \mathbf{n} d_\sigma}{\tau}$

(36) $\operatorname{div} \mathbf{v} = \dfrac{\partial v_1}{\partial x} + \dfrac{\partial v_2}{\partial y} + \dfrac{\partial v_3}{\partial z}$

(37) $\mathbf{v} = \mathbf{i}v_1 + \mathbf{j}v_2 + \mathbf{k}v_3$

(38) $\operatorname{curl} \mathbf{v}(P) = \lim_{\tau \to 0} \dfrac{\int_S \mathbf{n} \times \mathbf{v} d\sigma}{\tau}$

(39) $\operatorname{curl} \mathbf{v} = \mathbf{i}\left(\dfrac{\partial v_3}{\partial y} - \dfrac{\partial v_2}{\partial z}\right) + \mathbf{j}\left(\dfrac{\partial v_1}{\partial z} - \dfrac{\partial v_3}{\partial x}\right) + \mathbf{k}\left(\dfrac{\partial v_2}{\partial y} - \dfrac{\partial v_1}{\partial x}\right)$

States, Richard Hubert Bruck and Erwin Kleinfeld proved that a certain class of numbers named for Arthur Cayley comprise the only not-associative, finite-dimensional real division algebra. Finally in separate papers on algebraic topology in 1958 by Raoul Bott and John Willard Milnor and by Michel A. Kervaire, it was established that the only finite-dimensional, non-associative division algebras over $\Re$ are dimensions, 1, 2, 4, and 8, and the only such alternative algebras are, once again, the real numbers, complex numbers, quaternions, and Cayley numbers. (D.T.F.)

Field theory. A particle moving in three-dimensional space can be located at each instance of time t by a position vector $\mathbf{r}$ drawn from some fixed reference point O. Because the position of the terminal point of $\mathbf{r}$ depends on time, $\mathbf{r}$ is a vector function of t. Its components in the directions of Cartesian axes, introduced at O, are the coefficients of $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ in the representation that identifies these three vectors with vectors of unit length along the three mutually perpendicular axes (see 29). If the components are differentiable functions, the derivative of $\mathbf{r}$ with respect to t is defined by a formula (see 30) that represents the velocity $\mathbf{v}$ of the particle. The Cartesian components of $\mathbf{v}$ appear as coefficients of $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$. If these components are also differentiable, the acceleration $\mathbf{a} = d\mathbf{v}/dt$ is obtained by differentiation (see 31). It may also be verified that the rules for differentiating products of scalar functions remain valid for derivatives of the dot and cross products of vector functions, and suitable definitions of integrals of vector functions allow the construction of the calculus of vectors. Such calculus has become a basic analytic tool in physical sciences and technology (see ANALYSIS, REAL).

Applications of vector calculus to problems of continuum mechanics (*e.g.*, fluid mechanics, aerodynamics) call for a consideration of scalar and vector functions specified at each point of some region. A region of space with each point of which a scalar function is associated is called a scalar field, while a region in which a vector function is determined is a vector field. Examples of scalar fields are regions at each point of which the temperature or density of a body can be determined. A region in the vicinity of a charged body in which the electric-intensity vector is determined is an example of a vector field. A scalar function $u(P)$ determined at each point P of a scalar field is called a scalar point-function, while a vector function $\mathbf{v}(P)$ specified in a vector field is a vector point-function.

Examples of scalar fields

Gradient of a scalar field. If there is a scalar point-function $u(P)$ at a point P, and another scalar point-function $u(P')$ at a nearby point P', the vector from P to P' being expressed as $\Delta\mathbf{r}$ (see 32), then the ratio (see 33) of the difference between u at P' and u at P' relative to $\Delta\mathbf{r}$ represents the average space rate of change of $u(P)$ in the direction of $\Delta\mathbf{r}$. The limit of this ratio as $|\Delta\mathbf{r}| \to 0$, when this limit exists, represents the space rate of change of $u(P)$ in the direction of $\Delta\mathbf{r}$. The vector in that direction for which the space rate of change of $u(P)$ is a maximum is called the gradient of $u(P)$ and is denoted by grad u or ∇u. It can be shown that in Cartesian coordinates, the vector components of the gradient are partial derivatives of u; that is, derivatives of u, obtained by holding all but one variable fixed, and differentiating with respect to the unrestricted variable (see 34). Thus, with each point of a scalar field, in which ∇u exists, there can be associated a vector field. If $u(P)$ is the temperature, then ∇u gives the direction of the heat-flow vector in the field.

Divergence of a vector field. On the other hand, two important fields can be associated with each continuously differentiable vector point-function $\mathbf{v}(P)$: one is a scalar field and the other a vector field. If $\mathbf{v}(P)$ is defined in some region T bounded by a sufficiently smooth surface S, then the component of $\mathbf{v}(P)$ (which for the sake of concreteness can be thought to represent velocity $\mathbf{v}$ of fluid particles moving in T) in the direction of the unit vector $\mathbf{n}$ that is exterior (directed away from) and normal to the surface S is $\mathbf{v} \cdot \mathbf{n}$. The amount of fluid issuing from S then is given by $\int \mathbf{v} \cdot \mathbf{n} d\sigma$, in which the integration symbol represents the summation of $\mathbf{v} \cdot \mathbf{n}$ over the elements of area $d\sigma$ of the surface S.

The flux of fluid per unit volume is thus equal to a ratio of the integral referred to, relative to τ, in which τ is the volume of T (see 35). The limit of this ratio as $\tau \to 0$, so that S shrinks toward some point P in T, is an important scalar called the divergence of $\mathbf{v}(P)$. Thus, the divergence of $\mathbf{v}(P)$, written div $\mathbf{v}(P)$ represents the rate of fluid flow from P. If div $\mathbf{v}(P)$ is positive at P, then P is a source of fluid; if it is negative, then P is a sink. If div $\mathbf{v}(P) = 0$, then no fluid issues from P.

In Cartesian coordinates div $\mathbf{v}(P)$ is given by a formula (see 36) that is the sum of partial derivatives of vector components of the velocity vector $\mathbf{v}$ (see 37).

Divergence in Cartesian coordinates

Curl of a vector field. Associated with $\mathbf{v}(P)$ is an important vector field in which the vector called curl of $\mathbf{v}(P)$ is defined by a formula that expresses, in the limit of small volume, the ratio of $\int \mathbf{n} \times \mathbf{v} d\tau$ relative to that volume, in which the integral is calculated over the surface S (see 38). This vector provides a measure of the angular velocity of the fluid at any point P in the field. In Cartesian coordinates curl $\mathbf{v}$ is given by an especially convenient formula in which the vector components are differences of partial derivatives (see 39). When curl $\mathbf{v} = 0$ at every point of the region, the field is said to be irrotational, and when $\mathbf{v}(P)$ is such that div $\mathbf{v} = 0$, the field is solenoidal. The importance of these two special fields stems from the fact that every continuously differentiable vector function $\mathbf{v}(P)$ defined in a region τ (subject to mild restrictions) can be expressed as a sum of two vector functions $\mathbf{f}(P)$ and $\mathbf{g}(P)$ such that $\mathbf{f}(P)$ is solenoidal and $\mathbf{g}(P)$ is irrotational. The possibility of such decomposition greatly simplifies the study of many velocity and force fields occurring in physics.

TENSOR ALGEBRA AND ANALYSIS

Tensors, or generalizations of vectors, as characterized at the beginning of this article, were invented to express

geometric entities arising in the study of manifolds, that is, in the study of topological spaces equipped with a family of interrelated local coordinate systems. Because manifolds were originally considered only in the small in terms of coordinate systems, tensors were expressed in terms of functions involving many indices that obey certain transformation laws when the coordinate system is changed. Subsequently, a global study of manifolds resulted in an invariant treatment of tensors and their algebraic rules of combination. This is the subject of tensor or multilinear algebra.

The concept of tensors and the knowledge of some of their properties can be traced back to Gauss and his younger countrymen Bernhard Riemann and Elwin Bruno Christoffel; their algebra and analysis were shaped into a systematic method by the Italians Curbastro Gregorio Ricci and Tullio Levi-Civita, who coined for this powerful branch of mathematics the name absolute differential calculus, now known as tensor analysis. The chief aim of tensor analysis is to construct and discuss relations of laws that are generally covariant; *i.e.*, that remain valid in passing from one to any other system of coordinates. Thus, tensor analysis is of prime importance for differential geometry. It became the object of widespread interest after the advent of general relativity theory, in which a principal requirement is precisely such unrestricted covariance of physical laws. There are useful applications of tensor analysis—both finite and infinite dimensional—to classical mechanics, quantum mechanics, fluid mechanics, and other fields.

Algebra of tensors. Tensors were characterized qualitatively at the beginning of this article. In specific terms, it is necessary to modify a tensor to be of a type that satisfies certain transformation laws or that is represented by a certain number of components. Thus, the term "tensor of type (s, r)" is used. Technically, a tensor of type (s, r) is an element of a vector space $\mathfrak{L}_{s,r}$ associated with a given vector space $\mathfrak{L}$ of dimension d. Any basis of $\mathfrak{L}$ induces a basis of $\mathfrak{L}_{s,r}$ in terms of which every tensor T in $\mathfrak{L}_{s,r}$ can be represented by its components (see 40). The vector space $\mathfrak{L}_{s,r}$ can be constructed from $\mathfrak{L}$: if $\mathfrak{L}^*$ denotes the dual space of $\mathfrak{L}$, that is, the vector space of linear functions of $\mathfrak{L}$, it follows that $\mathfrak{L}_{s,r}$ is the collection of scalar-valued functions of $r + s$ variables, r of them in $\mathfrak{L}^*$ and s of them in $\mathfrak{L}$, which are linear in each variable when the others are left fixed. Thus each T in $\mathfrak{L}_{s,r}$ gives the scalar value $T(v^1, \cdots, v^s, a_1, \cdots, a_r)$ for every $v^1, \cdots, v^s$ in $\mathfrak{L}$ and $a_1, \cdots, a_r$ in $\mathfrak{L}^*$. If $e^1, \cdots, e^d$ is a basis of $\mathfrak{L}$, let $\phi_1, \cdots, \phi_d$ be the basis of $\mathfrak{L}^*$ that is dual to $e^1, \cdots, e^d$ so that $\phi_i(e^j) = 0$ if $i \neq j$ and $\phi_i(e^i) = 1$. Because T is linear in each variable separately, T is completely determined by its values on the $r + s$ basis vectors $(e^{j_1}, \cdots, e^{j_s}, \phi_{i_1}, \cdots, \phi_{i_r})$ (see 41). If $\phi_{j_1} \otimes \cdots \otimes \phi_{j_s} \otimes e^{i_1} \otimes \cdots \otimes e^{i_r}$ is the tensor that gives the value 1 on $(e^{j_1}, \cdots, e^{j_s}, \phi_{i_1}, \cdots, \phi_{i_r})$ and 0 for any other choice of basis vectors, then T is representable as a linear combination of such tensors with the coefficients in the combination being the values of T identified above (see 42).

A basis and its dual

Suppose a different basis $f^1, \cdots, f^d$ of $\mathfrak{L}$ had been chosen and suppose $f^j = \sum_{k=1}^{d} a_k{}^j e^k$. Then the dual basis of $f^1, \cdots, f^d$ is the set of linear functions $\Psi_1, \cdots, \Psi_d$, in which $\Psi_i = \sum_{l=1}^{d} b_i{}^l \phi_l$ and $(b_i{}^l)$ is the inverse of the nonsingular matrix $(a_k{}^j)$. In terms of the new basis, T is represented by its components which are linear combinations of its components relative to the old basis, the expression of which constitutes a law of transformation (see 43).

Thus a tensor of type (s, r) can be described as a function of $r + s$ indices depending upon a basis that obeys such a transformation law when the basis is changed. The advantage of the invariant definition of $\mathfrak{L}_{s,r}$ (*i.e.*, independent of a basis) is that it is applicable when the scalars form only a ring and also vector spaces are infinite. It is also aesthetically more pleasing. The description of T in terms of its components $T_{i_1, \cdots, i_r}^{j_1, \cdots, j_s}$ is useful in explicit computations.

(40) $$T_{i_1, \cdots, i_r}^{j_1, \cdots, j_s}, \quad i_1, \cdots, i_r, j_1, \cdots, j_s = 1, \cdots, d$$

(41) $$T(e^{j_1}, \cdots, e^{j_s}, \phi_{i_1}, \cdots, \phi_{i_r}) = T_{i_1, \cdots, i_r}^{j_1, \cdots, j_s}$$

(42) $$T = \sum_{i,j=1}^{d} T_{i_1, \cdots, i_r}^{j_1, \cdots, j_s} \phi_{j_1} \otimes \cdots \otimes \phi_{j_s} \otimes e^{i_1} \otimes \cdots \otimes e^{i_r}$$

(43) $$T(f^{j_1}, \cdots, f^{j_s}, \Psi_{i_1}, \cdots, \Psi_{i_r}) = \sum_{k,l=1}^{d} a_{k_1}^{j_1} a_{k_2}^{j_2} \cdots a_{k_s}^{j_s} b_{i_1}^{l_1} b_{i_2}^{l_2} \cdots b_{i_r}^{l_r} T_{l_1, \cdots, l_r}^{k_1, \cdots, k_s}$$

(44) $$T\overline{T}(v^1, \cdots, v^s, v^{s+1}, \cdots, v^{s+\bar{s}}, a_1, \cdots, a_r, a_{r+1}, \cdots, a_{r+\bar{r}}) = T(v^1, \cdots, v^s, a_1, \cdots, a_r) \cdot \overline{T}(v^{s+1}, \cdots, v^{s+\bar{s}}, a_{r+1}, \cdots, a_{r+\bar{r}})$$

(45) $$T_{i_1, \cdots, i_r}^{j_1, \cdots, j_s} T_{i_{r+1}, \cdots, i_{r+\bar{r}}}^{j_{s+1}, \cdots, j_{s+\bar{s}}}$$

Construction and properties of tensor algebra. Two tensors can be multiplied to give another tensor. If T is a tensor of type (s, r) and T of type $(\bar{s}, r)$, then TT is a tensor of type $(s + s, r + r)$, in which in the law of multiplication (see 44), the first s vectors in the argument of T and the first $\bar{s}$ vectors in the argument of T combine to give the first $s + \bar{s}$ vectors in the argument of TT. In terms of a basis the components of TT are expressible as products of components of T and T (see 45). From this multiplication an associative algebra called the tensor algebra $\mathfrak{N}$ can be constructed over the vector space $\mathfrak{L}$. The elements of $\mathfrak{N}$ are the scalars together with the sum of all the spaces $\mathfrak{L}_{0,r}$ for $r = 1, 2, 3, \cdots$. Two tensors of type $(0, r)$ and $(0, r)$ can be multiplied by the rule above to give a tensor of type $(0, r + r)$. This rule together with the distinctive law determines the multiplication in $\mathfrak{N}$. The tensor algebra $\mathfrak{N}$ has the following properties, which characterize it completely:

1. It contains $\mathfrak{L} = (\mathfrak{L}_{0,1})$ and a unity element, the scalar 1. The smallest algebra in $\mathfrak{N}$ containing $\mathfrak{L}$ and 1 is $\mathfrak{N}$.
2. If $\mathfrak{S}$ is any associative algebra with unity, any linear mapping of $\mathfrak{L}$ into $\mathfrak{S}$ can be extended to a map of $\mathfrak{N}$ into $\mathfrak{S}$ that preserves multiplication.

Tensor product of vector spaces. The multiplication of two tensors is a special case of the tensor product of vector spaces. If $\mathfrak{L}_1, \cdots, \mathfrak{L}_k$ are k vector spaces, their tensor product $\mathfrak{L}_1 \otimes \cdots \otimes \mathfrak{L}_k$ is the vector of scalar-valued functions in k variables in $\mathfrak{L}_1^*, \cdots, \mathfrak{L}_k^*$, respectively, which are linear in each variable separately. If v^i lie in $\mathfrak{L}_i$, $i = 1, \cdots, k$, then their tensor product $v^1 \otimes \cdots \otimes v^k$ has the value $v^1(\alpha_1)v^2(\alpha_2) \cdots v^k(\alpha_k)$ on $(\alpha_1, \cdots, \alpha_k)$ in $\mathfrak{L}_1^*, \cdots, \mathfrak{L}_k^*$ respectively. The space $\mathfrak{L}_{s,r}$ is the tensor product of $\mathfrak{L}^*$ taken s times and $\mathfrak{L}$ taken r times.

Tensor fields. The tensors that arise in the study of manifolds are obtained by attaching to each point of a smooth d-dimensional manifold a d-dimensional vector space, its tangent space or space of directional derivatives. The tensors considered are those associated with each tangent space. A tensor field of type (s, r) on a manifold is a function that assigns to each point of the manifold a tensor of type (s, r) associated with the tangent space at that point. If a coordinate system with coordinate functions $x_1, \cdots, x_d$ is chosen about some point m of the manifold, then the directional derivatives $\frac{\partial}{\partial x_1}, \cdots, \frac{\partial}{\partial x_d}$ form a basis of the tangent spaces of points near m. In terms of this basis, a tensor field of type (s, r) has components $T_{i_1, \cdots, i_r}^{j_1, \cdots, j_s}$ that are functions of the coordinates $x_1, \cdots, x_d$. On the whole, tensor analysis studies the analytical properties of tensor fields, while tensor algebra deals only with the algebraic properties of tensors associated with a fixed vector space.

Tangent spaces

Associated algebras. Closely related to the tensor algebra $\mathfrak{N}$ over a vector space $\mathfrak{L}$ are several other algebras. The most important for geometric and topological applications are the Grassmann or exterior algebra over $\mathfrak{L}$. In this algebra, the elements to be multiplied are skew-symmetric tensors. A tensor T of type $(0, r)$ is skew-symmetric if $T(\alpha_{\pi(1)}, \alpha_{\pi(2)}, \cdots, \alpha_{\pi r}) = \varepsilon(\pi)T(\alpha_1, \alpha_2, \cdots, \alpha_r)$; $\varepsilon(\pi)$ is $+1$ or -1, depending upon whether the permutation π of the integers $1, 2, \cdots, r$ is even or odd. Any tensor T of type $(0, r)$ can be made skew-symmetric by alternating it; *i.e.*, $\text{ALT}_r(T)$ is skew-symmetric, in which it is a sum of terms of the type $(r!)^{-1}\ \varepsilon(\pi)T(\alpha_{\pi(1)}, \alpha_{\pi(2)}, \cdots, \alpha_{\pi(r)})$, summed over all permutations π of the integers $1, \cdots, r$ (see 46). The exterior algebra $\mathfrak{G}$ over $\mathfrak{L}$ is the space of skew-symmetric tensors in $\mathfrak{N}$ in which the multiplication is the product of two tensors made skew-symmetric. Thus if T and $\overline{T}$ are skew-symmetric tensors of type $(0, r)$ and $(0, \bar{r})$, respectively, $T \wedge \overline{T} = \text{ALT}_{r+\bar{r}}(T\overline{T})$ is a skew-symmetric tensor of type $(0, r+\bar{r})$. Under this multiplication $\mathfrak{G}$ becomes an associative algebra with unity of dimension 2^d that contains $\mathfrak{L} = (\mathfrak{L}_{0,1})$ and in which $v \wedge v = 0$ for each v in $\mathfrak{L}$.

Grassmann and Clifford algebras

An algebra with the symmetric tensors can be formed in a similar manner. A tensor T of type $(0, r)$ is symmetric if $T(\alpha_{\pi(1)}, \cdots, \alpha_{\pi(r)}) = T(\alpha_1, \cdots, \alpha_r)$ for every permutation π of the integers $1, \cdots, r$. Any tensor of type $(0, r)$ can be made symmetric; *i.e.*, $\text{SYM}_r(T)$ is a symmetric tensor of type $(0, r)$ where it is a sum of terms of the type $(r!)^{-1}\ T(\alpha_{\pi(1)}, \cdots, \alpha_{\pi(r)})$, the sum (see 47) taken over all permutations π of the integers $1, \cdots, r$. The polynomial algebra over $\mathfrak{L}$ is the space of symmetric tensors in $\mathfrak{N}$ where the multiplication is the product of two tensors made symmetric. Thus if T and $\overline{T}$ are symmetric tensors of type $(0, r)$ and $(0, \bar{r})$, respectively, then $T \times \overline{T} = \text{SYM}_{r+\bar{r}}(T\overline{T})$.

A generalization of Grassmann algebras called Clifford algebras is important for the study of certain representations of orthogonal groups. If B is a symmetric tensor of type (2, 0), over $\mathfrak{L}$, the orthogonal group of B is the group of nonsingular linear transformations l on $\mathfrak{L}$ for which $B(l(v), l(v)) = B(v, v)$ for all v in $\mathfrak{L}$. The Clifford algebra of B, which contains $\mathfrak{L}$ and in which $v \cdot v = B(v, v)1$, is useful for studying the spin representations of this group. The Clifford algebra of B is constructed from $\mathfrak{N}$ thus: Let $\mathfrak{J}_B$ denote the ideal in $\mathfrak{N}$ generated by all elements of the form $v^2 - B(v, v)1$ for v in $\mathfrak{L}$; *i.e.*, $\mathfrak{J}_B$ is the set of all linear combinations of elements of $\mathfrak{N}$ of the form $T(v^2 - B(v, v)1)$ or $(v^2 - B(v, v)1)T$. The Clifford algebra of B is the algebra $\mathfrak{N}/\mathfrak{J}_B$; *i.e.*, $\mathfrak{N}$ with the elements $\mathfrak{J}_B$ set equal to zero. If the field of scalars is not of characteristic two, and if $e^1, \cdots, e^d$ is a basis of $\mathfrak{L}$, then $e^{i_1} \cdot e^{i_2} \cdot e^{i_3} \cdot \cdots \cdot e^{i_k}$ $(i_1 < i_2 < \cdots < i_k)$ is a basis of the Clifford algebra, so that the Clifford algebra has dimension 2^d. If B is identically zero, the Clifford algebra is the previously constructed Grassmann algebra (see also ALGEBRA, LINEAR AND MULTILINEAR).

(46) $$\text{ALT}_r(T)(\alpha_1, \cdots, \alpha_r) = \frac{1}{r!}\sum \varepsilon(\pi)\ T(\alpha_{\pi(1)}, \cdots, \alpha_{\pi(r)})$$

(47) $$\text{SYM}_r(T)(\alpha_1, \cdots, \alpha_r) = \frac{1}{r!}\sum T(\alpha_{\pi(1)}, \cdots, \alpha_{\pi(r)})$$

(48) $$dx'_\iota = \frac{\partial x'_\iota}{\partial x_\alpha}\, dx_\alpha$$

(49) $$A'_\iota = \frac{\partial x'_\iota}{\partial x_\alpha} A^\alpha$$

(50) $$A'_\iota = \frac{\partial x_\alpha}{\partial x'_\iota} A_\alpha$$

(51) $$A'_{\iota\kappa} = \frac{\partial x_\alpha}{\partial x'_\iota}\frac{\partial x_\beta}{\partial x'_\kappa} A_{\alpha\beta}$$

(52) $$A'^{\iota\kappa} = \frac{\partial x'_\iota}{\partial x_\alpha}\frac{\partial x'_\kappa}{\partial x_\beta} A_{\alpha\beta}$$

(53) $$A'^{\kappa}_{\iota} = \frac{\partial x_\alpha}{\partial x'_\iota}\frac{\partial x'_\kappa}{\partial x_\beta} A^\beta_\alpha$$

(54) $$\left(A^{ik\cdots}_{\iota\kappa\cdots}\right)' = \frac{\partial x_\alpha}{\partial x'_\iota}\frac{\partial x_\beta}{\partial x'_\kappa}\cdots\frac{\partial x'_\iota}{\partial x_a}\frac{\partial x'_\kappa}{\partial x_b}\cdots A^{ab\cdots}_{\alpha\beta\cdots}$$

Tensor analysis. Consider a continuous n-dimensional manifold or space S_n (see TOPOLOGICAL GROUPS AND DIFFERENTIAL TOPOLOGY), the element or point $P(x_\iota)$ of which is determined by assigning the values of n real independent variables or coordinates x_ι. Let $Q(x_\iota + dx_\iota)$ be another point of S_n. Then the ordered point-pair P, Q or the set of differentials dx_ι is a vector. To begin with, the idea of size or length is foreign to this concept, as S_n thus far is a nonmetrical manifold. If the x_ι be transformed into any other system of n coordinates x_ι', the former being continuous functions of the latter with continuous derivatives $\partial x_\iota/\partial x_\kappa'$ and non-vanishing, finite Jacobian $\mathrm{J} = \dfrac{\partial x_\iota}{\partial x_\kappa'}$; then dx_ι are transformed into differentials dx_ι' by a linear transformation in which the coefficients are partial derivatives (see 48). (The convention will be adopted that every term in which an index occurs twice, is to be summed over all its values.) Any set of n magnitudes A^ι, functions of the x_ι, that are transformed by this rule, that is, by the rule of transformation that governs differentials (see 49), form a contravariant tensor of rank 1, the A^ι being its n components. For contravariant tensors upper indices are used, except for dx_ι, the prototype of all such tensors. Next, any n magnitudes A_ι that are transformed as the differentiators $\partial/\partial x_\iota$, which are arrived at by formally inverting the partial derivatives that constitute the coefficients (see 50), form a covariant tensor of rank 1; Lower indices are used for such tensors. These two kinds of tensors, or rank 1, are also termed vectors; *e.g.*, three-vectors, four-vectors (such as the relativistic four-velocity or four-potential, in space-time, S_4) and so on, according as $n = 3, 4, \cdots$. Similarly, any n^2 magnitudes $A_{\iota\kappa}$ transformable by a natural generalization of the rank 1 covariant tensor transformation (see 51) form a covariant, and any A^ι transformable by a natural generalization of the rank 1 contravariant tensor transformation (see 52) form a contravariant tensor of rank 2. Again, n^2 magnitudes $A_\iota{}^\kappa$ transformable by a linear transformation in which coefficients are products of coefficients from rank 1 covariant and contravariant tensor transformations (see 53) are said to form a mixed tensor of rank 2, covariant in ι and contravariant in κ. The extension to any rank is obvious. Any $n^{r_1+r_2}$ magnitudes $A^{\iota\kappa\cdots}_{\iota\kappa\cdots}$, with r_1 lower and r_2 upper indices, which are transformed according to a generalized rule (see 54), form a mixed tensor of rank $r = r_1 + r_2$. This is the most general concept of a tensor. A tensor of rank zero, called also a scalar, is a single function of the x_ι, invariant with respect to any transformations of coordinates, $f' = f$. $A_{\iota\kappa}$ is symmetrical if $A_{\iota\kappa} = A_{\kappa\iota}$, and antisymmetrical or a skew tensor if $A_{\iota\kappa} = -A_{\kappa\iota}$, implying $A_{\iota\iota} = 0$. Similarly for $A^{\iota\kappa}$. Analogous definitions hold for mixed tensors and for higher ranks. Symmetry and antisymmetry are invariant properties.

Tensor transformation and rank

Because the transformed tensor components are linear homogeneous functions of the original ones, the sums of corresponding components of tensors of same rank and kind form again a tensor. Thus, as A_ι, B_ι are covariant vectors, so is $A_\iota + B_\iota = C_\iota$. Similarly, $A_{\iota\kappa} + B_{\iota\kappa} = C_{\iota\kappa}$, and so on. The addends, functions of x_ι, must be taken at the same point of S_n. If a tensor vanishes in one coordinate system, it will vanish also in any other. Consequently, any tensor equation, if valid in one, holds also in any other coordinate system. This is the chief reason tensors are important in pure geometry and physics.

The importance of tensors in geometry and physics

The outer product of two tensors of ranks r and s (*i.e.*, the array of n^{r+s} products of their components) is again a tensor, of rank $r+s$. Thus, $A_\iota B_\kappa = C_{\iota\kappa}$, $A_{\iota\kappa}B_\lambda = C_{\iota\kappa\lambda}$, $A_{\iota\kappa}B^\lambda = C^\lambda{}_{\iota\kappa}$. The contraction, an operation of almost magical efficiency, applicable to any mixed tensor, consists in equating one of its r_1 lower to one of its r_2 upper indices and summing over it from 1 to n. The result is again a tensor, with $r_1 - 1$ covariant and $r_2 - 1$ contravariant indices; *e.g.*, a contraction of $A^\lambda{}_{\iota\kappa}$ gives $A^\kappa{}_{\iota\kappa} = A_\iota$, a vector; $B^{\lambda\mu}{}_{\iota\kappa}$ contracted once becomes $B^{\lambda\kappa}{}_{\iota\kappa} = B^\lambda{}_\iota$, and this yields $B_\iota{}^\iota = B$, a scalar.

The inner product is the outer product (supposed mixed) contracted once or more. Thus, the inner product of A_ι, B^κ is $A_\kappa B^\kappa$, a scalar; that of $A_{\iota\kappa}$, B^λ gives $A_{\iota\kappa}B^\kappa = C_\iota{}^\lambda$ a mixed tensor, and after a second contraction, $A_{\iota\kappa}B^{\iota\kappa} = C_\iota{}^\iota = C$, a scalar. Unlike $A_\kappa B^\kappa$, $A_\kappa B_\kappa$ is not invariant: it has no tensor character, and likewise for $A^\kappa B^\kappa$. Conversely, if $A_\kappa B^\kappa$ is a scalar for any covariant vector A_κ, B^κ is a contravariant vector. Similarly, if $A_{\iota\kappa}B^{\iota\kappa}$ is a scalar for any covariant $A_{\iota\kappa}$, $B^{\iota\kappa}$ is contravariant. This is an efficient method of establishing the tensor character of a set of n, n^2, $\cdots$ magnitudes.

Differentiation. The differentiation of tensors with respect to the coordinates yields, in certain circumstances, further tensors. If f is a scalar function of the x, or scalar field, $\partial f/\partial x_\iota = f_\iota$ is a covariant vector, the gradient of f (but $\partial^2 f/\partial x_\iota \partial x_\kappa$ is not a tensor; again, if du is an invariant, dx_ι/du is, but d^2x_ι/du^2 is not a vector). Further, if A_ι is a vector, $B_{\iota\kappa} = \partial A_\iota/\partial x_\kappa - \partial A_\kappa/\partial x_\iota$ is a skew tensor, the rotation of A_ι. Finally, if $A_{\iota\kappa}$ is antisymmetric, $B_{\iota\kappa\lambda} = \partial A_{\iota\kappa}/\partial x_\lambda + \partial A_{\kappa\lambda}{}^+/\partial x_\iota + \partial A_{\lambda\iota}/\partial x_\kappa$ is again such a tensor, the expansion of $A_{\iota\kappa}$.

In what precedes only such properties of tensors were treated as are independent of any metrical considerations; the space S_n thus far is a nonmetrical, amorphous manifold. Its metrics may now be fixed by laying down the line element, a quadratic differential form (see 55) to be considered as invariant and to serve as the measure of the (squared) size or length of the vector dx_ι. In the differential form there are coefficients $g_{\iota\kappa} = g_{\kappa\iota}$ that are prescribed functions of x. It follows that because dx_ι is contravariant, $g_{\iota\kappa}$ is covariant. Equivalently, it may be said that a certain symmetrical tensor $g_{\iota\kappa}$ is impressed upon S_n as the fundamental or metrical tensor, converting it into a metrical manifold, called a Riemannian space (after Bernhard Riemann).

The minors $g^{\iota\kappa}$ of $g = |g_{\iota\kappa}|$, divided by g, form again a symmetrical tensor, the contravariant metrical tensor, to be used along with $g_{\iota\kappa}$. The outer product and contraction can be used in raising and lowering tensor indices. Thus $g^{\iota\kappa}A_\kappa = A^\iota$ and $g_{\iota\kappa}B^\kappa = B_\iota$.

The angle θ between two (copunctal) vectors A^ι and B^ι is defined by the invariant $\cos\theta = g_{\iota\kappa}A^\iota B^\kappa/AB$; where A, B defines their sizes. The unit vector $\dot{x}_\iota = dx_\iota/ds$ determines, locally, a direction in S_n. The angle between two directions $\dot{x}_\iota, \dot{y}_\iota$ is given by $\cos\theta = g_{\iota\kappa}\dot{x}_\iota\dot{y}_\kappa$.

Integrals. The integral $\int dx_1, \cdots, dx_n$, briefly $\int dx$, extended over a region of S_n, is transformed into $\int J dx'$, and the determinant g into $g' = J^2 g$. Consequently, $\int\sqrt{g}dx$ is an invariant metrically impressed upon that region, its size or volume (area, if $n = 2$). This concept is readily extended to any sub-manifold of S_n characterized by x_ι as functions of $m < n$ parameters p_a, the role of metrical tensor is taken over by a linear transformation of it (see 56). If, by a proper choice of the coordinate system, all components $g_{\iota\kappa}$ are reducible to constants, the metrical space S_n is Euclidean or homaloidal (flat). This is but a very special case of Riemannian space. In general such a reduction is not possible, and S_n is non-Euclidean.

Differentiation of tensors by metrics

Differentiation of tensors, aided by metrics, yields an unlimited number of new tensors. The oldest of such is Riemann's set of four-index symbols of 1861. The simplest metrically differential tensor, however, was discovered in 1869 by Christoffel. This is the covariant derivative of a vector A, (see 57) in which certain terms $[\iota\kappa, \mu]$ are Christoffel's symbols (not forming a tensor) (see 57). Again, the generalization of this is a mixed tensor (see 58). Detailed symbolism is necessary to express the covariant derivatives of $C^{\iota\kappa}$ and $C_{\iota\kappa}$ (see 59, 60). The covariant derivative of $g_{\iota\kappa}$ itself vanishes identically. Noteworthy for their applications are the vector divergence of a skew $A^{\iota\kappa}$ and the scalar divergence of A^κ (see 61); *e.g.*, one group of the 19th-century Scottish physicist James Clerk Maxwell's equations is represented by $\mathrm{Div}\,(F^{\iota\kappa}) = C^\iota$, where C^ι is the four-current and $F^{\iota\kappa}$ the electromagnetic skew tensor (six vector); the second group is expressed by equating to zero the expansion of $F_{\iota\kappa} = g_{\iota\alpha}g_{\kappa\beta}F^{\alpha\beta}$.

Geodesics, null lines. The most characteristic curves of the metrical S_n are its null or minimal lines, $ds = 0$, and shortest curves or geodesics defined by $\delta\int ds = 0$, with fixed terminals. The equations of a geodesic are second-order differential equations expressed in terms of Christoffel symbols and $\dot{x}$, in which $\dot{x} = dx/ds$ (see 62). They determine—*e.g.*, in the case of space-time—the motion of a free particle in the metrical, and gravitational, field $g_{\iota\kappa}$. The null lines are imaginary or real according as the form (55) expressing the square of the differential of the line element is a definite or nondefinite form. The definite form holds in the case of spaces proper as contemplated by the pure geometer, and the nondefinite form characterizes space-time with one positive and three negative $g_{\kappa\kappa}$. The null lines of space-time represent light propagation. The definiteness or nondefiniteness of the quadratic form and its index of inertia (number of negative $g_{\kappa\kappa}$ terms) are invariant properties.

Parallel displacement of vectors. A fruitful contribution to tensor analysis due to Levi-Civita is the concept of parallel displacement of a vector with respect to a curve. The metrical S_n can in general be imagined as imbedded locally in a Euclidean space E of $\frac{1}{2}n(n+1)$ dimensions. Levi-Civita defined parallel displacement in S_n as a consequence of a parallel displacement in E. In spite of the ultraspatial construction, parallelism is intrinsic, expressible by the S_n metrics alone.

If $x_\iota = f_\iota(s)$ are the parametric equations of any chosen curve C, then there are differential equations along C that with the initial vector $A^\iota(S_o) = A_o{}^\iota$ determine uniquely a contravariant vector A^ι at each point of the curve (see 63). The vector A^ι at the point corresponding to $s = s$ is said to be obtained by a parallel displacement of the vector $A_o{}^\iota$ with respect to the given curve C and is said to be parallel to $A_o{}^\iota$. Any two vectors generated

(55) $$ds^2 = g_{\iota\kappa}dx_\iota dx_\kappa$$

(56) $$h_{ab} = g_{\iota\kappa}(\partial x_\iota/\partial p_a)(\partial x_\kappa/\partial p_b)$$

(57)
$$\begin{cases} A_{\iota,\kappa} = \dfrac{\partial A_\iota}{\partial x_\kappa} - \{\iota\kappa, \lambda\}A_\lambda \\ \{\iota\kappa, \lambda\} = g^{\lambda\mu}[\iota\kappa, \mu] = \{\kappa\iota, \lambda\} \\ [\iota\kappa, \mu] = \frac{1}{2}(\partial g_{\mu\iota}/\partial x_\kappa + \partial g_{\kappa\mu}/\partial x_\iota - \partial g_{\iota\kappa}/\partial x_\mu) \end{cases}$$

(58) $$B^\lambda_{\iota\kappa} = \partial B^\iota/\partial x_\kappa + \{\alpha\kappa, \iota\}B^\alpha$$

(59) $$C^{\iota\kappa}_{,\lambda} = \partial C^{\iota\kappa}/\partial x\lambda + \{\lambda\alpha, \iota\}C^{\alpha\kappa} + \{\lambda\alpha, \kappa\}C^{\iota\alpha}$$

(60) $$C_{\iota\kappa,\lambda} = \partial C_{\iota\kappa}/\partial x^\lambda - \{\iota\ , \alpha\}C_{\alpha\kappa} - \{\kappa\lambda, \alpha\}C_{\iota\alpha}$$

(61)
$$\begin{cases} A^\iota = \mathrm{div}(A^{\iota\kappa}) \equiv \dfrac{1}{\sqrt{g}}\dfrac{\partial}{\partial x_\kappa}(\sqrt{g}A^{\iota\kappa}) \\ \mathrm{div}(A^\kappa) \equiv \dfrac{1}{\sqrt{g}}\dfrac{\partial}{\partial x_\kappa}(\sqrt{g}A^\kappa) \end{cases}$$

(62) $$\ddot{x}_\iota + \{\alpha\beta, \iota\}\dot{x}_\alpha\dot{x}_\beta = 0$$

(63) $$\frac{dA^\iota}{ds} + \{\alpha\beta, \iota\}A^\alpha\dot{x}^\beta = 0$$

$$(64)\qquad X_{\iota,\beta,\alpha}-X_{\iota,\alpha,\beta}=R^{\kappa}_{\iota\alpha\beta}X_{\kappa}$$

$$(65)\qquad R^{\kappa}_{\iota\alpha\beta}=\frac{\partial}{\partial x_{\alpha}}\{\iota\beta,\kappa\}-\frac{\partial}{\partial x_{\beta}}\{\iota a,\kappa\}+\{a\gamma,\kappa\}\{\iota\beta,\gamma\}-\{\beta\gamma,\kappa\}\{\iota a,\gamma\}$$

$$(66)\qquad K=\frac{(\iota\kappa,\lambda\mu)\xi^{\iota}\eta^{\kappa}\xi^{\lambda}\eta^{\mu}}{(g_{\iota\lambda}g_{\kappa\mu}-g_{\iota\mu}g_{\kappa\lambda})\xi^{\iota}\eta^{\kappa}\xi^{\lambda}\eta^{\mu}}$$

as above, one at each point, are parallel with respect to C. If it be required to construct a curve the tangents of which follow from each other by parallel displacement along the curve itself, the result will be the equations of a geodesic (62), so that the curve is a geodesic. The latter is thus throughout parallel to itself. Accordingly, a vector transferred along a geodesic remains equally inclined to it.

If a vector X^{ι} be carried by a parallel displacement along two different paths a,b from O to O', the two vectors at O', equal in size, generally differ in direction. One is parallel to X^{ι} via a, the other via b. In short, parallelism depends on the route of transfer or is nonintegrable (unless S_n is essentially Euclidean). Thus also, if the vector is carried around a circuit, it returns at O with its direction changed.

The covariant derivative of the vector is the tensor $X_{\iota\alpha}$. Its covariant derivative $X_{\iota,\alpha,\beta}$ is not symmetric in α and β. A straightforward calculation shows that the difference of the two forms of this covariant derivative involves the important tensor $R_{\iota\alpha\beta}{}^{\kappa}$ (see 64). This latter tensor is the Riemann-Christoffel or curvature tensor (see 65) the vanishing of which is the necessary and sufficient condition for the reducibility of $g_{\iota\kappa}$ to a constant tensor, or the criterion of a homaloidal space.

The Riemann-Christoffel curvature tensor

Riemann's four-index symbols and curvature. Riemann's own set of four-index symbols is the covariant tensor $(\iota\mu,\ \lambda\kappa) \equiv R_{\iota\mu\lambda\kappa} = g_{\mu\alpha}R_{\iota\kappa\lambda}{}^{\alpha}$. Conversely, $R_{\iota\kappa\lambda}{}^{\alpha} = g^{\mu\alpha}(\iota\mu,\ \lambda\kappa)$. Like equation (9), the symbols are antisymmetric in κ, λ. Three more linear relations hold between them. This reduces the number of independent symbols to $n^2(n^2-1)/12$; *e.g.*, six for a three-space, 20 for an S_4, and but one for a surface, say (12, 12). This symbol divided by g is a scalar of the surface, its Gaussian curvature K. If a vector undergoes parallel displacement around a circuit on a surface, it returns with its direction changed by an amount equal to $\int K d\sigma$, the Gaussian curvature of the surface integrated over the surface area enclosed by the circuit. As a consequence, the excess (over π) of the angle sum in any geodesic triangle is $\int K d\sigma$, a theorem proved by Gauss.

When there are three or more dimensions, the curvature properties can no longer be expressed by a single magnitude, but require for their description the knowledge of the whole curvature tensor or the associated Riemann symbols. The concept of Gaussian curvature is now replaced by that of Riemannian curvature; *i.e.*, at any point $O(x)$ of S_n, the set of Gaussian curvatures K of geodesic surfaces of all possible orientations laid through O. If $h_{\alpha\beta}$ be the metrical tensor of such a surface as submanifold, $K = 1/h(12,\ 12)_k$, the symbol is calculated with $h_{\alpha\beta}$. This and the determinant h can be expressed in terms of the tensor $g_{\iota\kappa}$ of the manifold and the vector pair, ξ^{ι}, η^{ι}, fixing the orientation (see 66). In general, K will depend on position and on orientation. In other words, with regard to curvature, S_n may be nonhomogeneous and anisotropic (*e.g.*, space-time within or around matter); but if K is everywhere isotropic, it is also constant throughout S_n (Schur's theorem). By the expression (66) for K in terms of the tensor $g_{\iota\kappa}$ and the vector pair, ξ^{ι}, η^{ι}, the necessary and sufficient condition for isotropy of Riemannian curvature becomes $(\iota\lambda,\ \kappa\mu) = K(g_{\iota\kappa}g_{\lambda\mu} - g_{\iota\mu}g_{\lambda\mu})$ or $R_{\iota\kappa\lambda}{}^{\alpha} = K(\delta_{\kappa}{}^{\alpha}g_{\iota\lambda} - \delta_{\lambda}{}^{\alpha}g_{\iota\kappa})$ with constant K. See also PHYSICAL THEORIES, MATHEMATICAL ASPECTS OF.

BIBLIOGRAPHY. The following five works are textbooks on the algebra of vectors given in approximate order from elementary to advanced: DANIEL ZELINSKY, *A First Course in Linear Algebra* (1968); DANIEL T. FINKBEINER, *Introduction to Matrices and Linear Transformations*, 2nd ed. (1966); KENNETH HOFFMAN and RAY KUNZE, *Linear Algebra*, 2nd ed. (1971); PAUL R. HALMOS, *Finite-Dimensional Vector Spaces*, 2nd ed. (1958); NATHAN JACOBSON, *Lectures in Abstract Algebra*, vol. 2 (1953). A very readable account of recent research with additional bibliography is contained in CHARLES W. CURTIS, "The Four and Eight Square Problem and Division Algebras," ERWIN KLEINFELD, "A Characterization of the Cayley Numbers," and LOWELL J. PAIGE, "Jordan Algebras," in ABRAHAM A. ALBERT (ed.), *Studies in Modern Algebra*, pp. 100–186 (1963). The following three books are of historical and biographical interest: ERIC T. BELL, *The Development of Mathematics*, 2nd ed. (1945); J. WILLARD GIBBS, *Vector Analysis*, ed. by EDWIN B. WILSON (1901); WILLIAM R. HAMILTON, *Lectures on Quaternions* (1853). The following four works deal with vector analysis: LOUIS BRAND, *Vector and Tensor Analysis* (1947); HARRY LASS, *Vector and Tensor Analysis* (1950); IVAN S. SOKOLNIKOFF and R.M. REDHEFFER, *Mathematics of Physics and Modern Engineering*, 2nd ed. (1966); S. SIMONS, *Vector Analysis for Mathematicians, Scientists and Engineers* (1964). CLAUDE CHEVALLEY, *The Construction and Study of Certain Important Algebras* (1955); NICOLAS BOURBAKI, *Algèbre multilinéaire* (1948); and A.I. BORISENKO and I.E. TARAPOV, *Vector and Tensor Analysis with Applications*, rev. ed. (1968; orig. pub. in Russian, 3rd ed., 1966), are works that may be consulted by the reader interested in tensor algebra. The following books are references on tensor analysis and its applications: HENRY D. BLOCK, *Introduction to Tensor Analysis* (1962); DEREK F. LAWDEN, *An Introduction to Tensor Calculus and Relativity*, 2nd ed. (1967); IVAN S. SOKOLNIKOFF, *Tensor Analysis*, 2nd ed. (1964); and LEON BRILLOUIN, *Les Tenseurs en mécanique et en élasticité*, 2nd ed. (1960; Eng. trans., *Tensors in Mechanics and Elasticity*, 1964).

(D.T.F./Ed.)

Analytic and Linguistic Philosophy

The methods that have dominated British philosophy for most of the 20th century and American philosophy since somewhat more recently have been called Linguistic and Analytic because language and the analysis of the concepts expressed by language have been a central concern. Though Australia and the Scandinavian countries have also contributed to this movement, it has won a very limited following elsewhere. Although there is a unity of outlook in the tradition, individual philosophers and movements within it have differed, often radically, about the goals and methodology of philosophy. A leading figure prior to the mid-20th century, Ludwig Wittgenstein, an Austrian-born Cambridge philosopher, for example, may have been unique in the history of philosophy in having engaged in two periods of profoundly influential philosophical productivity of which the later work was in large part a renunciation and a sustained argument against the earlier. Yet both the early Wittgenstein (represented by his *Tractatus Logico-Philosophicus*, 1922) and the later (represented by the *Philosophical Investigations*, 1953) are central examples of Analytic philosophy.

Different approaches to the study of language

Moreover, the aims assigned to the philosophical study of language have often been different. Some philosophers, among them Bertrand Russell and the early Wittgenstein, have thought that the underlying structure of language mirrors that of the world—that from an analysis of language a philosopher can grasp important truths about reality. This so-called picture theory of language, though influential, is generally repudiated by current Analytic philosophers. Another important dispute concerns whether everyday language is defective, vague, misleading, and even, at times, contradictory. Some Analytic philosophers have thus proposed the construction of an "ideal" language: precise, free of ambiguity, and clear in structure. The general model for such a language has been symbolic logic, the growth of which in the 20th century has played a central role in Analytic philosophy. An ideal language, it was thought, would resolve many traditional philosophical disputes that have arisen from the misleading structure of natural languages. At the other pole, some philosophers have thought that many philosophic

problems have come from paying too little attention to what men say in everyday language about various situations.

Despite such disagreements, Analytic philosophers have much in common. Most of them, for example, have concentrated on particular philosophical problems, such as that of induction, or have examined specific concepts, such as those of memory or of personal identity, without attempting to construct any grand metaphysical schemes—an attitude that has roots as ancient as those of the Socratic method exemplified in Plato's dialogues. Almost invariably Plato began with specific questions such as "What is knowledge?" or "What is justice?" and pursued them in a way that can be viewed, without undue strain, as philosophical analysis in the modern sense.

Theory of definite descriptions

Ideally, a philosophical analysis illuminates some important concept and helps to answer philosophical questions involving the concept. A famous example of such analysis is contained in Bertrand Russell's theory of definite descriptions. In a simple subject–predicate statement such as "Socrates is wise," he said, there seems to be something referred to (Socrates) and something said about it (that he is wise). If, instead of a proper name, however, a "definite description" is substituted, as in the statement "The president of the United States is wise," there is apparently still something referred to and something said about it. But a problem arises when nothing fits the description, as in the statement "The present king of France is wise." Though there is apparently nothing for the statement to be about, one nevertheless understands what it says. Consequently, a pre-World War I philosopher, Alexius Meinong, celebrated for his *Gegenstandstheorie* ("theory of objects"), felt forced by such examples to distinguish between things that have real existence and things that have some other sort of existence; for such statements could not be understood unless they were about something.

In Russell's view, philosophers such as Meinong were misled by surface grammatical form into thinking that such statements are simple subject–predicate statements. In reality they are complex; in fact, an analysis of the foregoing example shows that the definite description, "the present king of France," is not an independent unit in the statement at all. Upon analysis, the statement is a complex conjunction of statements: (1) "There is a present king of France"; (2) "There is at most one present king of France"; and (3) "If anyone is a present king of France, he is wise." But, more importantly, each of the three components is a general statement and is not about anything or anyone in particular. There is no phrase in the complete analysis equivalent to "the present king of France," which shows that the phrase is not an expression, like a proper name, that refers to something as the thing that the whole statement talks about. There is no need, therefore, to make Meinong's distinction between things that have real existence and things that have some other kind of existence.

GENERAL VIEWPOINT OF ANALYTIC PHILOSOPHY

Nature, role, and method of analysis. Analytic philosophy is concerned with the close and careful examination of concepts.

Status of philosophy in the Empiricist tradition. In spirit and style Analytic philosophy has strong ties with the Empiricist tradition, which stresses the data received through the senses and which, except for brief periods, has characterized British philosophy for some centuries, distinguishing it from the more Rationalistic trends of continental European philosophy. It is not surprising, therefore, that Analytic philosophy should find its home mainly in the Anglo-Saxon countries. In fact, the beginning of modern Analytic philosophy is generally dated from the time when two of its major figures, Bertrand Russell and G.E. Moore, both Cambridge philosophers, rebelled against an anti-Empiricist Idealism that had temporarily captured the English philosophical scene. The most famous British Empiricists—John Locke, George Berkeley, David Hume, and John Stuart Mill—had many interests, doctrines, and methods in common with contemporary Analytic philosophers. Although many of their particular doctrines are favorite targets of attack by Analytic philosophers today, one feels that this is more the result of a common interest in certain problems than any difference in general philosophical outlook.

Most Empiricists, though admitting that the senses fail to yield the certainty requisite for knowledge, hold nonetheless that it is only through observation and experimentation that justified beliefs about the world can be gained; *i.e.*, a priori reasoning from self-evident premises cannot reveal how the world is. This view has resulted in a sharp dichotomy among the sciences: between the physical sciences, which ultimately must verify their theories by observation, and the deductive or a priori sciences—*e.g.*, mathematics and logic—the method of which is the deduction of theorems from given axioms. Thus, the deductive sciences cannot give justified beliefs, much less knowledge, of the world. This consequence was one of the cornerstones of two important movements within Analytic philosophy, logical atomism and Logical Positivism (see below). In the Positivist's view, for example, the theorems of mathematics are merely the result of working out the consequences of the conventions that have been adopted for the use of its symbols.

Physical versus a priori sciences

The question then arises whether philosophy itself is to be assimilated to empirical or to a priori sciences. Early Empiricists assimilated philosophy to the Empirical sciences. They were less self-reflective about its methods than contemporary Analytic philosophers are. Being preoccupied with epistemology (theory of knowledge) and the philosophy of mind, and holding that fundamental facts can be learned about these subjects from individual introspection, they took their work to be a kind of introspective psychology. Analytic philosophers in the 20th century, on the other hand, have been less inclined to appeal ultimately to direct introspection. Moreover, the development of rigorous methods in formal logic seemed to promise help in solving philosophical problems—and logic is as a priori as a science can be. It seemed, then, that philosophy must be classed with mathematics and logic.

Function and methodology of philosophy

Conceptual, linguistic, and scientific analysis. The question remained, however, what philosophy's function and methodology are. For a great many Analytic philosophers who do philosophy in the minute and meticulous manner of G.E. Moore and, in particular, for those who have made Oxford the centre of Analytic philosophy (see below, *Recent trends in England: Oxford philosophers*), its business is the analysis of concepts. For them, philosophy is an a priori discipline because the philosopher in some sense already possesses the concept in which he is interested and needs no observations in order to analyze it.

Philosophy can be seen either as conceptual or as linguistic analysis. In the analysis of the concept of seeing, for example, the philosopher is not expressing purely linguistic concerns—with, say, the English verb "to see"—though an investigation of what can be said using that verb may be relevant to his conclusions. For a concept is independent of any particular languages; a concept is something that all languages, insofar as they are capable of expressing the concept, have in common. Thus, philosophers who stress that it is concepts that they analyze attempt to rebut the charge that their problems and solutions are merely verbal.

In contrast, other Analytic philosophers have been concerned with how expressions are used in a particular, nontechnical, everyday language. Thus, the term ordinary language philosophy has been applied by critics as a term of opprobrium to such philosophers. An influential study, *The Concept of Mind* (1949), by Gilbert Ryle, a prominent Oxford Analyst, is an example of a work that some critics took to depend in large part on a trivial appeal to how English speakers talk; but many of Ryle's arguments could equally well have been given by Analytic philosophers who would look upon the term ordinary language with horror.

The problem of perception illustrates how Analytic

Philosophy as compared with science

philosophers who do conceptual analysis think of the goal of philosophy as both different from and complementary to sicence. Physiologists, psychologists, and physicists—through experiments, observations, and testable theories—have also contributed to man's understanding of perception. There is in the sciences, however, a strong tendency to advance beyond earlier positions, which seems to be absent from philosophy. In philosophy, for example, the account of perception given by such 20th-century Analytic philosophers as G.E. Moore and the Positivist A.J. Ayer has a close connection with that of Locke in the 17th century.

The difference between philosophy and science is that, whereas the scientist investigates an actual occurrence, such as seeing, the philosopher investigates a concept that he already possesses quite independently of what he might discover through the occurrence. Whereas the scientist begins by supposing that he can recognize examples of seeing and is already exercising the concept, the philosopher wants to know what is involved in seeing in the sense of what conditions one can use to classify cases as examples of seeing. He may want to know, for example, whether certain conditions are necessary or sufficient. In testing the philosophical theory that, for an observer to see an object, the object must cause a visual experience in him (the causal theory of perception), one does not set up a scientific experiment. It would be of no use to set up situations in which various physical objects are not causing any visual experiences in order to see whether they still can be seen. For if the theory is correct, no such experimental situation will be an instance of seeing; and if it is wrong, merely describing a hypothetical situation would suffice. The question is one about how situations are classified, and for that purpose hypothetical situations are as good as real ones.

Therapeutic function of analysis. For some philosophers in the Analytic tradition, especially those influenced by Wittgenstein, the analysis of concepts has therapeutic value beyond the intrinsic enjoyment of doing it. Even scientists and laymen in their philosophical moments generate problems by not understanding the proper analyses of the concepts that they employ. They are then tempted to formulate theories to explain these difficulties, when instead they should be sorting out the roles of the concepts, which would show them that there was no problem to begin with. Thus, the failure to see how psychological concepts—sensations, emotions, and desires—are employed has led philosophers to such problems as how one can know what is going on in another's mind or how desires and emotions can produce physical changes in the body, and vice versa. Analysis of the concepts involved would, in this way of looking at philosophy, "dissolve" rather than solve the problems, for philosophers would come to see that their formulations of the problem rest on mistakes about the concepts involved.

Dissolving of problems by analysis

This way of looking at philosophy has often been criticized as making it merely a clearing up of the confusions of other philosophers and therefore a sterile enterprise. The confusions, however, need not be only those of other philosophers. Scientists, for example, can also generate philosophical theories that affect how they design their experiments, which may, thus, be subjects for philosophical therapeutics. Behaviorism in psychology—which views emotions, desires, and attitudes as being dispositions to behave in certain ways—seems to be a philosophical theory and perhaps to be based on a confusion about the analysis of psychological concepts. Yet Behaviorism has influenced psychologists in their approach to the science. Thus, in this view, philosophy can have a therapeutic value beyond the sphere of philosophical games.

Philosophy, in spite of its abstractness, has traditionally been concerned with human needs, and the therapeutic model may even fulfill this ideal. Laymen, as well as philosophers, for example, are bothered by the thought that their actions are determined not by themselves but by prior conditions. This is a problem that, if the therapeutic view is correct, rests on the misunderstanding of such concepts as causation, responsibility, and action, which need clarification.

Formal versus ordinary language. The role of language as a central concern of Analytic philosophers is the dimension most involved in disputes about the methodology employed. Philosophers outside the Analytic movement tend to think that its preoccupation with language is a departure from philosophy as classically conceived. Yet Plato and Aristotle, medieval philosophers, the Empiricists—and, in fact, most of the philosophers whose works have been considered important—have found it essential to talk about language. There are serious differences, however, about what role language should play. One such difference concerns the importance of formal languages (in the sense employed in symbolic logic) for philosophical problems.

Development of mathematical logic. Since the time of Aristotle, logic has been allied to philosophy. Until the late 19th century, however, logic was largely confined to formulating elaborate rules for one fairly simple form of argument—the syllogism; and there was a lack of systematic development of the subject along lines that had been taken in mathematics since early times.

Almost from the beginning, mathematicians had rigorously exploited two important techniques: (1) the use of the axiomatic method (as in Euclid's geometry) in developing the subject; and (2) the use of schematic letters or variables for stating general truths in the subject (thus, one can write "$A + B = B + A$," in which any names or numbers whatsoever can be substituted for A and B, and the result will still be true).

Axiomatic method and use of symbols

It is surprising that logicians through the ages failed to grasp the power of the use of schematic letters. When they finally began to employ these and other mathematical techniques, they made great contributions to man's understanding of the subject.

Among the developments that occurred in the 19th century, primarily through the work of mathematicians, those of the Englishman George Boole, creator of Boolean algebra, and of Georg Cantor, the Russian-born creator of set theory, are especially important inasmuch as they gave promise of bringing logic and mathematics closer together. The one figure who was both a mathematician and a philosopher and so might be credited with the marriage of logic as a philosophical subject with the techniques of mathematics was Gottlob Frege (died 1925), of the University of Jena in Germany. Historically, Frege, whose works are now appreciated in their own right, was important principally for his influence on Bertrand Russell, whose monumental work, *Principia Mathematica* (1910–13), written in collaboration with Alfred North Whitehead, together with Russell's earlier *Principles of Mathematics* (1903), awakened philosophers to the fact that the use of mathematical techniques in logic might prove to be of great importance for philosophy. Its symbolism had the advantage of being closely connected with ordinary language, whereas its rules can be precisely formulated. Moreover, work in symbolic logic has produced many distinctions and techniques that can be applied to ordinary language.

Divergence of ordinary language from formal logic. Ordinary language, however, seems to differ from the artificial language of symbolic logic in more respects than its lack of precisely stated rules. On the surface, it often appears to violate the rules of symbolic logic. In the English statement "If this is gold [symbolized by *p*], then this will dissolve in aqua regia [symbolized by *q*]," for example, which in symbolic logic is expressed in a form known as the material conditional, $p \supset q$ (in which $\supset$ means "If . . . then . . ."), one of the rules is that the statement is true whenever "This is gold" is false. In ordinary language, on the contrary, one would not count the statement as true merely on formal logical grounds but only if there were some real connection in the world of chemical reactions between being gold and dissolving in aqua regia—a connection that plays no role in symbolic logic.

Divergences between logic and language

Among Analytic philosophers the existence of many such apparent divergences between symbolic logic and

ordinary language has generated attitudes ranging from complete mistrust of symbolic logic as relevant to nonartificial languages to the position that ordinary language is not a proper vehicle for the rigorous statement of scientific truths.

Interpretations of the relation of logic to language. Symbolic logic has been viewed by many Analytic philosophers as providing the framework for an ideal or perfect language. This statement can be taken in two ways:

1. Russell and the early Wittgenstein thought of logic as revealing, in a precise fashion, the real structure of any language. Any seeming departure from this structure in ordinary language must therefore be attributed to the fact that its surface grammar fails to reveal its real structure and is apt to be misleading. As a corollary, philosophers who have held this view have often explained philosophical problems as arising from being taken in by the surface features of the language. Because of the similarity of sentences such as "Tigers bite" and "Tigers exist," for example, the verb "to exist" may seem to function, as other verbs do, to predicate something of the subject. It may seem, then, that existence is a property of tigers just as their biting is. In symbolic logic, however, the symbolic equivalent of the two sentences would be quite different; existence would not be represented by a symbol for a predicate but by what is called the existential quantifier, $(\exists x)$, which means "There exists at least one x such that. . . ."

2. The other sense in which symbolic logic has been seen as the framework of an ideal language is exemplified in the work of Rudolf Carnap, a 20th-century semanticist, who was concerned with what the best language—especially the best for the purposes of science—is.

Nonassertoric uses of language

One distinctive feature of the formal language of *Principia Mathematica* is that it becomes, when interpreted, a language of true-or-false statements. In ordinary language, on the contrary, one is not restricted to statements of truths; in it one can also issue commands, ask questions, make promises, express beliefs, give permission, and assert necessities and possibilities. Consequently, many philosophers have developed nonstandard logics that incorporate the nonassertoric features of language. Thus, various systems of logic have been formulated and studied (see LOGIC, APPLIED).

On the other side of the coin, many philosophers—most notably the later Wittgenstein and those influenced by him—have thought that attempting to put language into the straitjacket of a formal system is to falsify the way that language works. Language performs a multitude of tasks, and even among expressions that seem to be alike in the way they function—those sentences, for example, that one might think are used simply for expressing facts—examination of their actual use reveals many differences: differences, for instance, in what is counted as showing them to be true or false and in their relationships to other parts of language. Formal systems, according to this view, at best oversimplify and at worst can lead to philosophical problems generated by supposing that all language operates strictly according to a simple set of rules. Accordingly, far from settling philosophical disputes by getting underneath the misleading exterior of ordinary language, formal systems add their own share of confusion.

EARLY HISTORY OF ANALYTIC PHILOSOPHY

Reaction against Idealism. During the last decades of the 19th century, English philosophy was dominated by an absolute Idealism that stemmed from the German philosopher G.W.F. Hegel. For English philosophy this represented a break in an almost solid tradition of Empiricism. The seeds of modern Analytic philosophy were sown when two of the most important figures in its history, Bertrand Russell and G.E. Moore, broke with Idealism at the turn of the 20th century.

Idealism and science

Absolute Idealism was avowedly metaphysical in the sense that its adherents thought of themselves as describing, in a way not open to scientists, certain very fundamental truths about the world. Indeed, what pass for truths in the sciences, were, in their view, not really truths at all; for the scientist must, perforce, treat the world as composed of distinct objects and can only describe and state the relationships supposedly holding among them. But the Idealists held that to talk about reality as if it were a multiplicity of objects is to falsify it; in the end only the whole, the absolute, has reality.

In their conclusions and, most importantly, in their methodology, the Idealists were decidedly not on the side of commonsense intuition. Thus, a Cambridge philosopher, J.M.E. McTaggart, argued that the concept of time is inconsistent and cannot therefore be exemplified in reality. British Empiricism, on the other hand, had always thought of common sense as an ally and science as the model of the way in which to find out about the world. Even when their views might seem out of step with common sense, the Empiricists were generally concerned to reconcile the two.

One can hardly claim that Analytic philosophers have universally thought of themselves as on the side of common sense and much less that metaphysical conclusions (on the ultimate nature of reality) are absent from their writings. But there is in the history of the Analytic movement a strong antimetaphysical strain, and its exponents have generally assumed that the methods of science and of everyday life are the authentic ways of finding out the truth.

Founding fathers: Moore and Russell. The first break from the Idealist view that the physical world is really only a world of appearances occurred when Moore, in a paper, "The Nature of Judgment" (1899), argued for a theory of truth that implies that the physical world has the independent existence that, apart from philosophical theories, it is naïvely supposed to have. Though the theory was soon abandoned, it did represent a return to common sense.

Comparison of the two founders of Analytic philosophy

The influences on Russell and Moore—and thus their methods of dealing with problems—soon diverged, and their different approaches became the roots of two broadly different methodologies in the Analytic tradition.

Russell was a major influence on those who approached philosophical problems armed with the technical equipment of formal logic, who saw the physical sciences as the only means of gaining knowledge of the world, and who regarded philosophy—if a science at all—as a deductive and a priori enterprise on a par with mathematics. Russell's contributions to this side of the Analytic tradition have been important and, in great part, lasting.

Moore, on the other hand, never found much need to employ technical tools nor to turn philosophy into a science. His dominant themes were (1) the defense of commonsensical views about the nature of the world against esoteric, skeptical, or grandly metaphysical views and (2) the conviction that the right way to approach philosophical puzzles is to ask exactly what the question is that generated the puzzle before trying to solve it. Philosophical problems, he thought, are often intractable because philosophers have not stopped to formulate precisely what is at issue.

G.E. Moore. Because of these two themes, Moore enlisted much more sympathy among Analytic philosophers from the 1930s and onward who were followers of Wittgenstein's later writings, of Gilbert Ryle's postwar *The Concept of Mind*, and of John Austin's work (see below: *Oxford philosophers*). These philosophers, like Moore, saw little hope in advanced formal logic as a means of solving traditional philosophical problems and believed that philosophical skepticism about the existence of an independent external world or of other minds—or, in general, about what men label as common sense—must be wrong. The followers of Wittgenstein also shared with Moore the belief that it is often more important to look at the questions that philosophers pose than at their proposed answers. Thus, unlike Russell, who was important for his solutions in formal logic and ideal models of language, it was more the spirit of Moore's conception of philosophy than its lasting contributions that makes him a seminal influence.

The Idealists were given to arguing for what, in Moore's

Moore's attack on Idealism

eyes, were outrageous positions. Thus, in his essay "A Defence of Common Sense" (1925), as in others, his defense was not only against such Idealist doctrines as the unreality of time but also against any of the forms of skepticism—about the existence of other minds or of a material world—that philosophers have espoused. The skeptic, he pointed out, usually has some argument for his conclusion. Instead of examining such arguments, however, Moore pitted against the skeptic's premises quite everyday beliefs, such as, for example, that he had breakfast that morning (thus time cannot be unreal) or that he does in fact have a pencil in his hand (thus there must be a material world). His challenge to the skeptic is to show that the premises of the skeptic's argument are more certain than the everyday beliefs that form Moore's premises.

Although some commentators have seen Moore as an early practitioner of the appeal to "ordinary language," his appeal was really not to what it is proper to say but rather to the beliefs of common sense. His rejection of anything that offends against common sense, however, was influential not only in the release that it afforded from the metaphysical excesses of absolute Idealism but also in its impact on the continuing attitudes of most Analytic philosophers—even though they may have given it a linguistic turn.

Moore was also important for his vision of the proper business of philosophy—analysis. He was puzzled about what is the proper analysis of "*X* sees *Y*," in which *Y* designates a physical object (*e.g.*, a pencil). There must be a special sense of "see," in which one does not see the pencil but only part of its surface. And finally—and most importantly—there is also a sense in which what is directly perceived is not even the surface of the pencil but, rather, what Moore called "sense data" and which earlier Empiricists had called "visual sensations" or "sense impressions." Moore's problem was to discern the relationships among these various elements in perception and, in particular, to discover how a person can be justified, as Moore fully believed he is, in his claims to see physical objects when what he immediately perceives are really sense data. The idea that sense impressions form the immediate objects of perception has played a large role in Analytic philosophy, showing once again its Empiricist roots. It later became an important source of division, however, among the Logical Positivists (see below). Most post-World War II Oxford philosophers, however, together with those closely influenced by Wittgenstein's later work, have found sense data to be as unpalatable and unwarranted as Moore had found McTaggert's doctrine of the unreality of time to be.

Moore's notion of "sense data"

Bertrand Russell. One of the recurring themes in philosophy is the idea that the subject needs to be given a new methodology. Among Empiricists this has often meant making it more scientific. From an early date, Russell enunciated this viewpoint (which was not shared by Moore), finding in the techniques of symbolic logic a measure of reassurance that philosophy might be put on a new basis. Russell did not see the philosopher, however, as merely a logician. Symbolic logic might provide the framework for a perfect language, but the content of that language is something else. The job of the philosopher is—for Russell, as it was for Moore—analysis. But the purpose is somewhat different. In most of Russell's work, analysis has the task of uncovering the necessary assumptions—especially about the kinds of things that exist—for a description of the world as it is. For the most part this description is the one that science gives and is therefore realistic. Thus, Russell's use of analysis was openly metaphysical.

The question then arises of how philosophical analysis, which is concerned with how men talk about the world, can presume to give any answers about how the world is. The search for an answer begins with the above-mentioned theory of descriptions—a theory that seems to be closely tied to linguistic concerns. It will be recalled that Russell considered that such definite descriptions as "the author of 'On Denoting'" are not really expressions used to refer to things in the world but that, instead, they make the statements in which they occur into quite general propositions about the world, to the effect that one and only one thing of a certain sort exists and that it has a certain property. Because there must be some way, however, of directly speaking of the things in the world, Russell turned his attention to proper names. The name, Aristotle, for example, does not seem to carry any descriptive content. But Russell argues, on the contrary, that ordinary names are really concealed definite descriptions ("Aristotle" may simply mean "The student of Plato who taught Alexander, wrote the *Metaphysics*, etc."). If a name had no descriptive content, one could not sensibly ask about the existence of its bearer, for one could then not understand what is expressed by a statement involving it. If "Bosco" were a name in this sense (without any descriptive content), then merely to understand the statement that Bosco exists or the statement that Bosco does not exist presupposes that one already knows what the name Bosco refers to. But then there cannot be any genuine question about Bosco's existence, for just to understand the question one must know the thing to which the name refers. Ordinary proper names, however—Russell, Homer, Aristotle, and Santa Claus—as Russell pointed out, are such that it makes sense to question the existence of their bearers. Thus, ordinary names must be concealed descriptions and cannot be the means of directly referring to the particular things in the world.

Use of the theory of descriptions

Names in the strict logical sense, then, are very rare; Russell, in fact, suggests that in English the only possible candidates are the demonstrative pronouns, this and that. Yet, if men are ever to talk about the actual things in the world directly, there must be the possibility of such demonstrative expressions underlying their language—in their private thoughts about the world if not in their public language.

To this point, Russell had concluded that things in the world can be talked about only through the medium of a special kind of name; in particular, one about which no question can arise whether it names something or not. At this point there was a transition from questions about the nature of language to results about the nature of the world. Russell asked what sort of thing it is that can be named in the strict logical sense, that can be known and talked about, and that can tell a man something about the world. The important restriction is that no question can arise about whether it exists or not. Ordinary physical objects and other people seem not to fit this requirement.

The question of what can be known and talked about

In his search for something whose existence cannot be questioned, Russell hit upon present experience and, in particular, upon sense data: one can question whether he is really seeing some physical object—whether, for example, there is a desk before him—but a person cannot question that he has had visual impressions or sense data; thus, what a man can name in the strict logical sense and what things he can actually talk about turn out to be the elements of his present experience. Russell therefore made a distinction between what can be known by acquaintance and what can be known only by description; *i.e.*, between those things the existence of which cannot be doubted and those about which, at least theoretically, doubt can be raised. What is novel about Russell's conclusion is that it was arrived at from a fairly technical analysis of language: to be directly acquainted with something is to be in a position to give it a name in the strict logical sense, and to know something only by description is to know only that something uniquely fits the description.

Russell was not constant in his view about physical objects. At one point he thought that the observer must infer their existence as the best hypothesis to explain his experience. Later he argued that they could be taken as logical constructions out of sense data.

Logical atomism: Russell and the early Wittgenstein. The next important development in Analytic philosophy was initiated when Russell published a series of articles entitled "Philosophy of Logical Atomism" (1918–19), in which he acknowledged a debt to Wittgenstein, who had

studied with Russell before the war. Wittgenstein's own work, the *Tractatus Logico-Philosophicus* (1922), which can also justly be said to present a logical atomism, turned out to be not only tremendously influential on developments in Analytic philosophy but also such a deep and difficult text that it has generated a growing body of scholarly interpretation.

Russell's choice of the words logical atomism to describe this viewpoint was, in fact, particularly apt. By using the word logical Russell meant to sustain the position, described earlier, that through analysis—particularly with the aid of the ideal structure provided by symbolic logic—the fundamental truths about how any language functions can be revealed and that this disclosure, in turn, would show the fundamental structure of that which the language is used to describe. And by using the word atomism Russell highlighted the particulate nature of the results that his analyses and those of Wittgenstein seemed to yield.

Simple atomic and molecular propositions

On the linguistic level, the atoms in question are atomic propositions, the simplest statements that it is possible to make about the world; and on the level of what language talks about, the atoms are the simplest atomic facts, those expressible by atomic propositions. More complex propositions, called molecular propositions, can then be built up out of atomic propositions via logical connectives—such as "either . . . or . . .," "both . . . and . . .," and "not . . ."—the truth-value of the molecular proposition being in each case a function of the truth values of its component atomic propositions.

Language, then, must break down, upon analysis, into ultimate elements that cannot be analyzed into any other component propositions; and, insofar as language mirrors reality, the world must then be composed of facts that are utterly simple. Atomic propositions are composed, however, of strings of names understood, as Russell had explained it, in the strict logical sense; and atomic facts are composed of simple objects, the things that could be thus named.

The details of the Russell–Wittgenstein view have fascinated philosophers by the way in which they not only formed a coherent view but also seemed to follow inexorably from the central assumptions. There are close connections between this period, which was perhaps the most metaphysical in contemporary Analytic philosophy, and traditional Empiricism. The breakdown of language and the world into atomic elements had been one of the prominent features in the classical Empiricists, John Locke, George Berkeley, and David Hume. There was also a view of the connection between language and the world—adumbrated in Russell but fully evident in the *Tractatus*—which has been important and influential, viz., the picture theory, which holds that the structure of language mirrors that of the world. Analysis is important because ordinary language does not show immediately, for example, that it is founded on the atomic–molecular proposition model. Another theme is that the deductive sciences—mathematics and logic—are based solely on the way that language operates and cannot reveal any truths about the world, not even about a world of entities called numbers. Finally, logical atomism, in Wittgenstein's thought as opposed to Russell's, was at one and the same time metaphysical—in the sense of conveying via pure reasoning something about how the world is—and antimetaphysical. Wittgenstein's *Tractatus* is unique in the history of Empiricism in its acceptance of the fact that it is itself a metaphysic and that part of its metaphysics is that metaphysics is impossible: the *Tractatus* says of itself that what it says cannot be coherently said. Only empirical science can tell a man anything about the world as it is. Yet the *Tractatus* apparently tells him, for example, about the relationship between language and the facts of the world. For Wittgenstein, the solution of this seeming paradox lies in his distinction between what can be said and what can only be shown. There are certain things that can somehow be seen to be so—in particular, the ways in which language is connected with the world. The *Tractatus* could not straightforwardly tell its readers about these matters—metaphysics cannot be a body of facts expressible in any language—but the attempt to say these things, done in the right way, can show them what it cannot coherently express.

Logical atomism as an antimetaphysical metaphysic

Logical Positivism: Carnap and Schlick. Wittgenstein's *Tractatus* was both a landmark in the history of contemporary Analytic philosophy and perhaps its most aberrant example. It not only contained the most highly sophisticated metaphysics but also was an important influence on the most antimetaphysical of the positions taken by Analytic philosophers, viz., that of Logical Positivism, which was mainly developed by a group of philosophers, scientists, and logicians who were centred in Vienna and came to be known as the Vienna Circle. Among these, Rudolf Carnap and Moritz Schlick have perhaps had the most influence on Anglo-American philosophy, although it was an English philosopher, A.J. Ayer—whose *Language, Truth and Logic* (1936) is still the most widely read work of the movement in America and England—who introduced the ideas of Logical Positivism to English philosophy. Its main tenets have struck sympathetic chords in the Analytic philosophers and are still important today, even if in repudiation.

Above all else, Logical Positivism was antimetaphysical; nothing can be learned about the world, it held, except through the methods of the empirical sciences. The Positivists sought a method for showing both (1) when a theory that seemed to be about the world was really metaphysical and (2) that such a theory was, in fact, meaningless, and this they found in the principle of verification. In its positive form, the principle said that the meaning of any statement that is really about the world is given by the methods employed for verifying its truth or falsity—the only allowable methods being, ultimately, those of observation and experiment. In its negative form, the principle said that no statement could both be a statement about the world and have no method of verification attached to it. Its negative form was the weapon used against metaphysics and for the vindication of science as the only possible source of knowledge about the world. The principle would, thus, class as meaningless many philosophical and religious theories that purport to say something about the world but provide no way of testing the truth of the statements; for example, in religion it would render suspect the statement that God exists, which, being metaphysical, would be, strictly speaking, meaningless.

Principle of verification

The principle of verification ran almost immediately into difficulties, most of which were first raised by the Positivists themselves. The attempt to work out these difficulties belongs to a more detailed study of the movement (see POSITIVISM AND LOGICAL EMPIRICISM). It is sufficient to note here that these problems were sufficient to make most subsequent Analytic philosophers wary of appealing directly to the principle. It has, however, influenced philosophical work in more subtle ways.

With the principle of verification in hand, the Positivists thought that they could show a great many theories to be nonsense. There were several areas of discourse, however, which failed the test of the principle but which were simply impossible to rule out as concealed nonsense. Foremost among these disciplines were mathematics and ethics. Mathematics (and logic) could hardly be written off as nonsense. Yet their theorems are not verifiable by observation and experiment; they are known, in fact, by pure a priori reasoning alone. The answer seemed to be provided in Wittgenstein's *Tractatus*, which held that the propositions of mathematics and logic are, in Kantian terms, analytic; *i.e.*, true—like the statement "All bachelors are unmarried"—in virtue of the conventions that lie behind the use of the symbols involved.

Emotive theory of ethics

About ethics or, more precisely, about any statements involving value judgments, the Positivist view was different, yet still of lasting importance. In this view, value judgments are not, like mathematical truths, necessary adjuncts to science. But they can not be put off as nonsense; nor, obviously, are they true by definition or linguistic convention. The usual view of the Positivists, called emotivism, is that what look like statements of fact (*e.g.*, that one should not tell lies) are really expres-

sions of one's feelings toward a certain action; thus, value judgments are not really true or false. The Positivist's position was that neither mathematical nor ethical statements could be dismissed, as were metaphysical propositions. Both had then to be exempted from the principle of verification; and this was done by arguing that their statements are not really about the world: mathematical truths are conventions, and ethical statements are merely expressions of feelings. The divorce of ethics from science, once again, reflects an old Empiricist theme, to be seen, for example, in David Hume's dictum that from matters of fact one cannot derive a conclusion about what ought to be nor vice versa.

LATER HISTORY OF ANALYTIC PHILOSOPHY

"Philosophical Investigations": the later Wittgenstein. A crucial turn that initiated developments that were destined to have a lasting and profound effect on much of contemporary Analytic philosophy occurred in 1929, when Wittgenstein, after some years in Austria during which he was not philosophically very active, returned to England and established his residence at Cambridge. There, the direction of his thought soon shifted radically away from his *Tractatus*, and his views became in many ways diametrically opposed to those of logical atomism. Because he published none of the materials of this period, his influence on other English philosophers—and ultimately on those in all of the countries associated with Analytic philosophy—spread by way of his students and those who heard him in the small groups to whom he spoke at Cambridge. His style, too, changed from the semi-rigorous and formally organized propositions of the *Tractatus* to sets of loosely connected paragraphs and remarks in which the ideas are often conveyed more by suggestion and example than discursively. The result has been that one of the major splits within the ranks of Analytic philosophy is that between those who derive their methods from the later Wittgenstein and those who have followed the *Tractatus*.

Although Wittgenstein's thoughts ranged over almost the entire field of philosophy, from the philosophy of mathematics to ethics and aesthetics, their impact has been felt most, perhaps, where it has concerned the nature of language and the relationship between the mental and the physical.

Language and following rules. In logical atomism, as shown above, language was conceived as having a certain necessary and fairly simple underlying structure that it was the job of philosophy to expose. Wittgenstein began to tear away at this assumption. Language, he now thought, is like an instrument that can be used for an indefinite number of purposes. Hence, any effort to codify how it must operate by giving some small set of rules would be like supposing that there is some rigid necessity that a screwdriver (for instance) can be used only to drive screws and forgetting that screwdrivers are also, quite successfully, used to open jars and to jimmy windows. Language is a human institution that is not bound by an outside set of rules—only by what men consider to be correct and incorrect. And that, in turn, is not really a matter for a priori theories to consider.

Language as a multi-purpose conventional instrument

The notion of a rule and what it means to follow a rule was especially prominent in his writings. Several concerns made this point of particular interest to Wittgenstein. In mathematics and logic, emphasis was being placed on the rules for manipulating the symbolism. As has been seen, symbolic logic has also been a model for the underlying structure of language. If this fact is coupled with the fact that Russell and the Wittgenstein of the *Tractatus* saw language as reflecting these rules and with the general Empiricist tradition that explains how language operates by each person following internal rules and standards for the use of his words, the picture of the system that Wittgenstein thought mistaken then emerges, and it becomes clear why he placed the notion of a rule so centrally.

Natural languages, however, are significantly different in that one does not first learn the rules and then use the language; indeed, prior to learning the language, one would not know what to do with rules. Mathematics and logic are, in this sense, bad models for language because they aim at setting out beforehand the rules and principles that are subsequently to be used. They encourage the belief that language must have a rigid structure and that, without rules, no language would be possible. The "rules" that one might plausibly discern in the language that he speaks are not, as rules, already there, in a ghostly way, guiding what he says; they are either generalizations from the finite data of what is counted as correct or incorrect, or they are rules that, as Wittgenstein metaphorically expressed it, one puts away in the archives—he adopts the rule but only after the fact.

Post-factum status of rules

Following a rule, however, was a concept that Wittgenstein saw as wrongly analyzed in many classical views about language. Thus, he cast irrevocable doubt on the prevalent theory—typified best, perhaps, in John Locke's *Essay Concerning Human Understanding* (1690)—that to use an expression meaningfully is to have in one's mind a standard or a rule for applying it correctly. Against this theme, Wittgenstein's point was that a rule by itself is dead—it is like a ruler in the hands of someone who has never learned to use it, a mere stick of wood. Rules cannot compel nor even guide a person unless he knows how to use them; and the same is true about mental images, which have often been thought to provide the standard for using linguistic expressions. But if rules themselves do not give life to words but require a similar explanation for what gives them life, then there is a useless regress and no (philosophical) explanatory value in the whole apparatus of internal rules and standards.

Relation between mental and physical events. In some respects, Wittgenstein made some significant breaks with the Empiricist tradition—in his views about language and the explanation of the rigour of the deductive sciences. His treatment of the relationship between mental events and physical events also represents an important departure. Empiricists generally have started from the important assumption that what a person is immediately acquainted with is his own sensations, ideas, and volitions, and that these are mental and not physical; and, most importantly, that the things he knows immediately are essentially private and inaccessible to others. For both Moore and Russell there then arose the problem of how, in view of the privacy stressed by the sense-datum theory, the world of physical objects could be known. Wittgenstein's attack on this viewpoint, which has come to be known as "the private language" argument, has become well-known, partly because it was in this area that Wittgenstein presented what could most easily be picked out as a more or less formal argument—one that could then be analyzed and criticized in an analytic manner. Even in this case, however, his style of writing was such that his precise formulation of the argument has become a main source of controversy. Wittgenstein argued that the notion of an utterly private experience would imply: (1) that what goes on in the mental life of a person could be talked about only in a language that that person alone whose mental life it was could understand; (2) that such a private language would be no language at all (this has been the main source of controversy); and (3) that the widely held doctrine that there are absolutely private mental events cannot be intelligibly stated, because to do so would be to suppose that one can publically say something about what the doctrine itself says cannot be mentioned in a language accessible to more than one person.

Arguments against private language

The fact that Wittgenstein's argument against private language depends essentially on the question, "What is it to follow a rule?" illustrates a common characteristic of his writings, viz., that themes developed in one area of philosophy continually emerge in apparently quite divorced areas. His extraordinary ability to see a common source of difficulty in philosophical problems that seem to be unrelated helps to explain his style of writing, which seems at first sight to be a somewhat chaotic arrangement of ideas.

Analytic philosophy has also been attracted to a be-

Problem of behaviorism

havioristic view of mental phenomena that holds that such apparently private events as the feeling of fear are not only not really private but also that they can be identified with publicly observable patterns of behaviour. The disposition toward empirical science, with observation as its foundation, united with the observation that the evidence men have of what goes on in the mental lives of other people must come from what they see of their behaviour, has often warred against the other inclination of Empiricism to regard the starting point of all knowledge of the world, for each person, as being essentially private sense experience. Wittgenstein has had tremendous influence, however, in suggesting that these two extremes are not the only alternatives. Yet attempts to state how Wittgenstein could deny the privacy of experience without espousing some form of behaviorism have not been very successful. Sympathetic interpreters have taken up the notion of "criteria," used, but not developed in any detail, by Wittgenstein. For mental states such as fear, outward behaviour (*e.g.*, running away, blanching, or cringing) does not constitute what it is to be in that state, as behaviorism would have it, but neither is it merely evidence of some completely private event. The problem has been to characterize the relation between behaviour and mental states so that the two are neither identical nor evidence one for the other, while still acknowledging that a knowledge of the person's characteristic behaviour is essential to understanding the notion of a certain mental state.

Recent trends in England. Those philosophers who might fairly be labelled "Wittgensteinians," who follow the methods that Wittgenstein employed in his later period, should be distinguished from those who have been influenced more indirectly by the general trends and philosophical atmosphere that arose in large part from Wittgenstein's work.

Wittgensteinians. Close students of his ideas have tended to work chiefly on particular concepts that lie at the core of traditional philosophical problems. As an example of such an investigation, a monograph entitled *Intention* (1957), by G.E.M. Anscombe, an editor of Wittgenstein's posthumous works, may be cited as an extended study of what it is for a person to intend to do something and of what the relationship is between his intention and the actions that he performs. This work has occupied a central place in a growing literature about human actions, which in turn has influenced views about the nature of psychology, of the social sciences, and of ethics. And, as an extension of this British influence into the United States, one of Wittgenstein's students, Norman Malcolm of Cornell University, has investigated such concepts as knowledge, certainty, memory, and dreaming. As these topics suggest, Wittgensteinians have tended to concentrate on Wittgenstein's ideas about the nature of mental concepts and to work in the area of philosophical psychology. Typically, they begin with classical philosophical theories and attack them by arguing that they employ some key concept, such as that of knowledge, in a manner incongruous with the way in which the concept would actually be employed in various situations. Their works thus abound with descriptions of hypothetical, though usually homely, situations and with questions of the form, "What would a person say if . . .?" or "Would one call this a case of *X*?" In doing so, they are following out Wittgenstein's advice that, instead of trying to capture the essence of a concept by an abstract analysis, the philosopher should look at how it is employed in a variety of situations.

Concentration on philosophical psychology

Oxford philosophers. After World War II, Oxford University was the centre of extraordinary philosophical activity; and, although Wittgenstein's general outlook on philosophy—his turning away, for example, from the notion of formal methods in philosophical analysis—was an important ingredient, many of the Oxford philosophers could not be called Wittgensteinians in the strict sense. The method employed by many of these philosophers has often been characterized—especially by critics—as an "appeal to ordinary language," and they were thus identified as belonging to the school of "ordinary language" philosophy. Exactly what this form of argument is supposed to be and what exemplifies it in the writings of these philosophers has been by no means clear. Gilbert Ryle, Moore's successor as editor of a leading journal, *Mind*—and especially in his *The Concept of Mind*—was among the most prominent of those analysts who were regarded as using ordinary language as a philosophical tool. Ryle, like Wittgenstein, pointed out the mistake of regarding the mind as what he called "a ghost in a machine"—to defeat the radical dualism of mind and body that has characterized much of philosophical thinking—by investigating how people employ a variety of concepts, such as memory, perception, and imagination, that designate "mental" properties. He tried to show that, when philosophers carry out such investigations, they find that, roughly speaking, it is the way people act and behave that leads to attributing these properties to them, and that there is no involvement of anything internally private. He also attempted to show how philosophers have come to dualistic conclusions—usually from having a wrong model in terms of which to interpret human activities. A dualistic model may be constructed, for example, by wrongly supposing that an intelligently behaving person must be continually utilizing knowledge of facts—knowledge that something is the case. Ryle contended, on the contrary, that much intelligent behaviour is a matter of knowing how to do something and that, once this fact is acknowledged, there is no temptation to explain the behaviour by looking for a private internal knowledge of facts. Though Ryle's objectives were similar to those of Wittgenstein, his results have often seemed more behavioristic than Wittgenstein's.

Gilbert Ryle's *The Concept of Mind*

It is true that Ryle did ask, in pursuit of his method, some fairly detailed questions about when a person would say, for example, that someone had been imagining something; but it is by no means clear that he was appealing to ordinary language in the sense that his was an investigation into how, say, speakers of English use certain expressions. In any case, the charge, often voiced by critics, that this style of philosophizing trivializes and perverts philosophy from its traditional function would probably also have to be levelled against Aristotle, who frequently appealed to "what we would say."

A powerful philosophical figure among postwar Oxford philosophers was John Austin, who was White's professor of moral philosophy until his death, in 1960. Austin felt that many philosophical theories derive their plausibility from overlooking distinctions—often very fine—between different uses of expressions, and he also thought that philosophers too frequently think that any one of a number of expressions will do just as well for their purposes. (Thus, ignoring the difference between an illusion and a delusion, for example, gives credence to the view that what one immediately perceives are not physical objects, but sense data.) Austin's work was, in many respects, much closer to the ideal of philosophy as comprising the analysis of concepts than was that of Ryle or Wittgenstein. He was also much more concerned with the nature of language itself and with general theories of how it functions. This novel approach, as exemplified in *How to Do Things with Words* (1962), set a trend that has been followed out in a growing literature in the philosophy of language. Austin took the total speech act as the starting point of analysis, which allowed him to make distinctions based not only upon words and their place in a language but also upon such points as the speaker's intentions in making the utterance and its expected effect on the audience. There was also in Austin's approach something of the program of Russell and the early Wittgenstein for laying bare the fundamental structure of language.

The work of John Austin

Recent trends in the United States. Although the Oxford philosophers and the posthumous publication of Wittgenstein's writings have produced a revolution in Anglo-American philosophy, the branch of Analytic philosophy that emphasized formal analyses by means of modern logic has by no means been dormant. Since the appearance of *Principia Mathematica*, striking new find-

ings have emerged in logic, many of which, though requiring for their understanding a high level of mathematical sophistication, are nevertheless important for philosophy.

W.V.O. Quine's use of symbolic logic

Among those philosophers for whom symbolic logic occupies a central position, W.V.O. Quine, Pierce professor of philosophy at Harvard University, has been especially important. Symbolic logic represented for him, as it did for many earlier Analytic philosophers, the framework for the language of science. There were two important themes in his work, however, that represent significant departures from, say, the positions of the logical atomists and the Logical Positivists. In the first place, Quine rejected the distinction between those statements in which their truth or falsity depends upon the meaning of the terms involved and those in which their truth or falsity is a matter of empirical and observable fact—a distinction that had played an essential role in Logical Positivism and was thought by most Empiricists to be the basis for a division between the deductive and the empirical sciences. Quine, in "Two Dogmas of Empiricism" (1951) and subsequent writings, argued that the sort of distinction intended by philosophers is impossible to draw. In the course of his argument, a similar doubt was cast upon concepts traditional not only to philosophy but also to linguistics—in particular, the concept of synonymy or sameness of meaning. Quine's attack has been a threat not only to some long-held doctrines of the Analytic tradition but also to its conception of the nature of philosophy, which has generally depended upon contrasting it with the empirical sciences.

The second important departure of Quine's philosophy has been his attempt to show that science can be successfully conducted without what he calls "intentional entities." In contrast to "extensional," used above as an essential feature of standard symbolic logic, intentional entities include many of the common items that Analytic philosophers often assume that they can talk about without difficulty, such as the meanings of expressions, propositions, or the property of certain statements (such as those of mathematics) of being necessarily true. Quine's program—as exemplified by *Word and Object* (1960)—is intended in part to show that science can say everything that it needs to say without using concepts that cannot be expressed in the extensional language of standard logic. Quine's work, though by no means widely accepted, has made Analytic philosophers at least wary of uncritically accepting certain of their standard distinctions.

Noam Chomsky's generative grammar

Since the mid-20th century, there has been an interaction between the science of linguistics and Analytic philosophy. This did not occur before because Analytic philosophers had almost always considered their study of language to be a priori and unconcerned with empirical facts about particular languages. Recently, however, a book by Noam Chomsky, a U.S. generative grammarian, entitled *Syntactic Structures* (1957), has produced a theory of grammar that not only has profoundly affected the course of linguistics but also bears striking resemblances to philosophical analysis. At first, some Analytic philosophers saw in Chomsky's theory a technique that could be applied to philosophy. It was subsequently considered, however, that, whereas the possibility of looking at grammar in Chomsky's way had contributed valuable concepts for philosophers, the possibility that it would become a methodology for Analytic philosophy had receded. The interchange between linguists and philosophers, however, has continued.

Analytic philosophy today. It is not possible to forecast in any detail the future trends of Analytic philosophy in Anglo-American and Scandinavian countries. It seems relatively certain, however, that the two conceptions of the subject that stem from Moore and Russell will both continue.

Analytic philosophers, mainly influenced by Oxford philosophy, and those for whom symbolic logic is a touchstone analyze many of the same problems and benefit from each other's work. Analysis in the more rigorous sense that Russell's theory of definite descriptions represents is more frequently an aim, despite the doubts of Wittgenstein and many of the Oxford philosophers. The general idea that the only ultimate explanations of the world are the scientific ones and the usual corollary that philosophy is in the service of science—which was a central idea for Russell, for the Logical Positivists, and (in recent times) for Quine—has apparently lost nothing of its vigour. The opposing tendencies, noted above, among Empiricists in general, and present also in Analytic philosophy, toward behaviorism or Materialism, on the one hand, and toward an Idealism of a phenomenalistic sort (such as that of the Irish bishop George Berkeley), on the other, are not present in the same form—mainly because of the sustained criticisms of Wittgenstein, of his followers, and of the Oxford philosophers. The battleground has shifted to a more subtle level. A substantial number of Analytic philosophers who are styled Materialists or physicalists have proposed a novel technique for reducing mental events and states to physical states. They avoid the well-exposed difficulties of older attempts in which it was held that, when one apparently talks about a separate realm of the mind—speaking of such things as thoughts, emotions, and sensations—the proper analysis of its meaning would be in terms of physical properties and events (usually observable behaviour). The novel idea, on the contrary, is that there is, in fact, an identity between so-called mental events and certain physical events, particularly those occurring in the brain, an identity that it is eventually the task of science to specify—in a way modelled after that in which science discovered that lightning is identical with an electrical discharge. The opposition against this new brand of scientific Materialism does not set up against it a view of the mind as a separate realm coexisting with the physical nor as an essentially private collection of nonphysical events and objects. Rather, the issue has been joined on the question whether the language (or perhaps the concepts) of the psychological and the physical are such as to allow for a scientifically discovered identity between items of the one and items of the other. That there still remains a division among Analytic philosophers concerning the problem of the mental and the physical (though in much altered form) shows both the continuity of the movement and the changes that have occurred.

The New Materialism

BIBLIOGRAPHY. G.J. WARNOCK, *English Philosophy Since 1900* (1958), is a very readable account covering much of the history. J.O. URMSON, *Philosophical Analysis* (1956), is especially good on logical atomism and its demise. BERTRAND RUSSELL, *The Problems of Philosophy* (1959), is a nontechnical account of his views just prior to the period of logical atomism. A collection of Russell's papers, *Logic and Knowledge*, ed. by R.C. MARSH (1956), contains his "Lectures on Logical Atomism," together with more technical papers. G.E. MOORE, *Some Main Problems of Philosophy* (1953), gives a good introduction to his methods and philosophical concerns. Two key papers by Moore are "A Defence of Common Sense," in *Contemporary British Philosophy*, Second Series (1925), and "Proof of an External World," in *Proceedings of the British Academy*, 25:273–300 (1939). Among anthologies are those of A.J. AYER (ed.), *Logical Positivism* (1959), which also includes papers critical of the movement and has an excellent bibliography; BERNARD WILLIAMS and ALAN MONTEFIORE (eds.), *British Analytical Philosophy* (1966); RICHARD RORTY (ed.), *The Linguistic Turn* (1967); J.A. FODOR and J.J. KATZ (eds.), *The Structure of Language* (1964); and LEONARD LINSKY (ed.), *Semantics and the Philosophy of Language* (1952). The 1961 translation of WITTGENSTEIN'S *Tractatus Logico-Philosophicus* by D.F. PEARS and B.F. MCGUINESS is superior to the original English version. Wittgenstein's *Philosophical Investigations* (1953) is the best known and historically the most influential of the posthumously published works from his later period. A.J. AYER, *Language, Truth and Logic*, 2nd ed. (1946), remains the best introduction to Logical Postivism. GILBERT RYLE, *The Concept of Mind* (1949), is the most famous example of what has been called "ordinary language" philosophy. J.L. AUSTIN, *How To Do Things with Words* (1962), contains the most discussed features of his views. W.V.O. QUINE, *From a Logical Point of View* (1953), is a collection of fairly nontechnical essays that provide good examples of the continuing influence of formal logic.

(K.S.D.)

Anarchism

Anarchism is a term describing a cluster of doctrines and attitudes whose principal uniting feature is the belief that government is both harmful and unnecessary. Derived from a Greek root signifying "without a rule," the terms anarchism, anarchist, and anarchy are used to express both approval and disapproval. In early contexts all these terms were used pejoratively: during the English Civil War of the 17th century the opponents of the radical Levellers referred to them as "Switzerising anarchists," and during the French Revolution the Girondin leader Jacques-Pierre Brissot accused his most extreme rivals, the Enragés, of being the advocates of "anarchy."

> Laws that are not carried into effect, authorities without force and despised, crime unpunished, property attacked, the safety of the individual violated, the morality of the people corrupted, no constitution, no government, no justice, these are the features of anarchy.

These words uttered by the leader of the French Revolutionary moderates in 1793 could serve as a model for the denunciations delivered by all opponents of the anarchists. The latter, for their part, would admit many of Brissot's points. They deny man-made laws, regard property as a means of tyranny, and believe that crime is merely the product of a society based on property and authority. But they would argue that their denial of constitutions and governments leads not to "no justice" but to the real justice inherent in the free development of man's sociality, his natural inclination, when unfettered by laws, to live according to the principles and practice of mutual aid.

ANARCHIST THINKERS

The first man who willingly called himself an Anarchist was the French political writer and pioneer Socialist Pierre-Joseph Proudhon. In 1840, writing his controversial study of the economic bases of society, *Qu'est ce que la propriété?* (*What is Property?*), Proudhon set out to shock his readers into attention by declaring: "I am an anarchist!" He went on to explain that in his view the real laws of society have nothing to do with authority but stem from the nature of society itself; he foresaw the eventual dissolution of authority and the emergence of a natural social order.

> As man seeks justice in equality, so society seeks order in anarchy. Anarchy—the absence of a sovereign—such is the form of government to which we are every day approximating.

The essential elements of the philosophy to which Proudhon in 1840 gave the name of Anarchism had already been developed by various earlier thinkers. There is a tradition of the rejection of political authority going back to classical antiquity, to the Stoics and the Cynics, and recurring through Christian history in dissenting sects such as the medieval Catharists and certain factions of Anabaptists during the Reformation. With such groups—often mistakenly claimed as ancestors by modern anarchist writers—the rejection of political government is merely one aspect of a retreat from the material world into a realm of spiritual grace; it becomes part of the search for individual salvation and as such is hardly compatible with the sociopolitical doctrine of Anarchism that in all its forms consists of (1) a fundamental criticism of the existing power-based order of society, (2) a vision of an alternative libertarian society based on cooperation as opposed to coercion, and (3) a method of proceeding from one order to the other.

The Diggers

English anarchist thought. The first sketch of an anarchist commonwealth in this sense was developed in the years immediately following the English Civil War by Gerrard Winstanley, a dissenting Christian who identified God with reason; he founded the minute Digger movement. In his pamphlet of 1649, *Truth Lifting Up Its Head Above Scandals*, Winstanley laid down what later became basic principles among the anarchists: that power corrupts; that property is incompatible with freedom; that authority and property are between them the begetters of crime; and that only in a society without rulers, where work and its products are shared, can men be free and happy, acting not according to laws imposed from above but according to their consciences. Winstanley was not only the pioneer of anarchist theory but also the forerunner of anarchist activism. He held that only by their own deeds can the people bring an end to social injustice, and in 1649, calling upon the people "to manure and work upon the common lands," he led a band of his followers in occupying a hillside in southern England, where they set about cultivation, established free communism among themselves, and offered passive resistance to the hostile landlords.

The Digger experiment was destroyed by local opposition, and Winstanley himself vanished into such obscurity that the place and date of his death are unknown. But the principles he defended lingered on in the traditions of English Protestant sects and reached their ultimate flowering in the masterpiece of a former dissenting minister, William Godwin, who in 1793 published his *Political Justice*—of which Sir Alexander Gray has said that in it "Godwin sums up, as no one else does, the sum and substance of anarchism, and thus embodies a whole tradition" (*The Socialist Tradition*, 1946, p. 134). This is essentially true, since Godwin not only presents the classic anarchist argument that authority is against nature and that social evils exist because men are not free to act according to reason but also sketches out a decentralized society in which the small autonomous community (the parish) is the essential unit. In his community, democratic political procedures are dispensed with as far as possible, because even the rule of the majority is a form of tyranny, and such procedures as voting dilute the responsibility of the individual; Godwin also condemns "accumulated property" as a source of power over others and envisages a loose economic system in which men will give and take according to their needs. Godwin was a prophet of technological progress; he believed that industrial development would eventually reduce the necessary working time to half an hour a day, provided men lived simply, and that this would facilitate the transition to a society without authority.

Godwin's classic argument

Godwin enjoyed great celebrity in the 1790s and influenced such varied writers as Percy Bysshe Shelley (whose *Queen Mab* and *Prometheus Unbound* are virtually anarchist poems), William Wordsworth, William Hazlitt, and Robert Owen, but he was almost forgotten by the time of his death in 1836. Though his ideas were to have, through Owen, a subterranean influence on the British labour movement, it is only recently that professed Anarchists have recognized his affinities with them. His *Political Justice* had virtually no effect on the quasi-political movement on the continent of Europe during the mid-19th century.

French anarchist thought. It was Pierre-Joseph Proudhon who laid the theoretical foundations of this movement. A brewer's son of peasant stock from the Franche-Comté, he started life (like many later anarchists) as a printer, but by the revolutionary year of 1848 he had already become a polemicist and a radical journalist with two books to his credit, *What is Property?* and *Économic Contradictions*. These established him among the leading theoreticians of Socialism, a term that in the early 19th century embraced a wide spectrum of attitudes. In Paris during the 1840s, Proudhon associated with Karl Marx and the Russian Mikhail Bakunin, and, out of the experiences of the Revolution of 1848 (when he served in the Constituent Assembly and voted against the constitution "because it is a constitution"), he developed the libertarian theories that he discussed in later works such as *Du principe fédératif* (*The Federal Principle*) and *De la capacité politique des classes ouvrières* (*The Political Capability of the Working Classes*).

The work of Proudhon

Proudhon was a complex and voluminous writer who remained obstinately independent, refusing to consider himself the founder of either a system or a party. Yet he was justly regarded by Bakunin, Peter Kropotkin, and other leaders of organized Anarchism as their true ancestor, for he had adumbrated their philosophy.

Mutualism, federalism, and direct action were the es-

sential doctrines Proudhon taught. By mutualism he meant the organization of society on an egalitarian basis. He declared that "property is theft," but this did not mean that he advocated communism. He attacked the use of property as a means of exploiting the labour of others, but he regarded "possession"—the right of a worker or group of workers to control the land or tools necessary for production—as an essential bulwark of liberty. He therefore envisaged a society formed of independent peasants and artisans, with factories and utilities run by associations of workers, all united by a system of mutual credit founded on people's banks. In place of the centralized state—the enemy of all Anarchists—Proudhon suggested a federal system of autonomous local communities and industrial associations, bound by contract and mutual interest rather than by laws, with arbitration replacing courts of justice, workers' management replacing bureaucracy, and integrated education replacing academic education. Out of such a network would emerge a natural social unity compared with which the existing order would appear as "nothing but chaos, serving as a basis for endless tyranny."

Proudhon remained all his life an independent polemicist, but in his final, posthumously published work, *De la capacité politique des classes ouvrières*, with its insistence that the liberation of the workers must be the task of the workers themselves organized in industrial associations, he laid the intellectual foundations of a movement that would reject democratic and parliamentary politics in favour of various forms of direct action. Unlike their master, Proudhon's working class followers of the 1860s did not accept the title Anarchist (though in 1850 an independent revolutionary, Anselme Bellegarrigue, had founded a short-lived magazine entitled *L'Anarchie*); they preferred to call themselves Mutualists, after a working class secret society bearing the same name to which Proudhon had belonged in Lyons during the 1830s. In 1864, shortly before Proudhon's death, a group of them joined with British trade unionists and European Socialists exiled in London to found the International Workingmen's Association (the First International). The Mutualists became the first opposition within the International to Karl Marx and his followers, who advocated political action and the seizure of the state in order to create a proletarian dictatorship. Marx's most formidable opponents, however, were not the Mutualists but the followers of Mikhail Bakunin, a Russian nobleman turned revolutionary who entered the International in 1868 after a long career as a political conspirator.

Russian anarchist thought. Bakunin had begun as a supporter of nationalist revolutionary movements in various Slav countries. In the 1840s he had come under the influence of Proudhon, and by the 1860s, when he entered the International, he had not only founded his own proto-Anarchist organization, the Social Democratic Alliance, with a considerable following in Italy, Spain, Switzerland, and the Rhône Valley, but had modified the Proudhonian teachings into the doctrine that became known as collectivism. Bakunin accepted Proudhon's federalism and his insistence on the need for working class direct action, but he argued that the modified property rights Proudhon allowed were impractical and instead suggested that the means of production should be owned collectively, though he still held that each worker should be remunerated only according to the amount of work he actually performed. The second important difference between Bakunin and Proudhon lay in their concepts of revolutionary method. Proudhon believed it was possible to create within existing society the mutualist associations that could replace it; he therefore opposed violent revolutionary action. Bakunin, declaring that "the passion for destruction is also a creative urge," refused to accept the piecemeal approach; a violent revolution, sweeping away all existing institutions, was in his view the necessary prelude to the construction of a free and peaceful society.

The work of Bakunin

Though the individualism and nonviolence implicit in Proudhon's vision have survived in the peripheral currents of the anarchist tradition, it was Bakunin's stress on collectivism and violent revolutionary action that dominated the mainstream from the days of the First International down to the destruction of anarchism as a mass movement at the end of the Spanish Civil War in 1939.

The First International was itself destroyed by the conflict between Marx and Bakunin, a conflict rooted as much in the contradictory personalities of the two leaders as in their rival doctrines—revolution by a disciplined party versus revolution by the spontaneous insurgence of the working class. When the international finally broke apart at the Hague congress in 1872, Bakunin's followers were left in control of the working class movements in the Latin countries—Spain, Italy, southern France, and French-speaking Switzerland—and these were to remain the principal bases of Anarchism in Europe. In 1873 the Bakuninists set up their own International, which lasted as an active body until 1877; during this period its members finally accepted the title of Anarchist rather than collectivist.

Bakunin died in 1876. His ideas had been developed in action rather than in writing, for he was the hero of many barricades, prisons, and meetings. His successor as the ideological leader, Peter Kropotkin (who renounced the title of prince when he became a revolutionary in 1876), is more celebrated for his writing than for his actions, though in his early years he led an eventful career as a revolutionary militant, which he described in a fine autobiography, *Memoirs of a Revolutionist* (1899). Under the influence of the French geographer Elisée Reclus (a former disciple of the Utopian Socialist Charles Fourier), Kropotkin developed the variant of anarchist theory known as anarchist communism. Kropotkin and his followers went beyond Bakunin's collectivism, since they argued not only that the means of production should be owned cooperatively but also that there should be complete communism in terms of distribution; this revived the scheme Sir Thomas More had sketched out in his 16th-century *Utopia* of common storehouses from which everyone should be allowed to take whatever he wished on the basis of "From each according to his means, to each according to his needs." In *La Conquête du pain* (1892) Kropotkin sketched the vision of a revolutionary society organized as a federation of free communist groups. He reinforced the vision by writing *Mutual Aid: A Factor in Evolution* (1902), in which he endeavoured to prove by means of biological and sociological evidence that cooperation is more natural and usual among both animals and men than competition. In his *Fields, Factories and Workshops* (1899) he put forward ideas on the decentralization of industry appropriate to a nongovernmental society.

The work of Kropotkin

ANARCHISM AS A MOVEMENT

Kropotkin's writings completed the theoretical vision of the Anarchist future, and little new has been added since his time. But this work was of less immediate importance than the emergence among the Italian anarchists of the theory of "propaganda of the deed." In 1876 Enrico Malatesta expressed the belief of the Italian Anarchists that "the *insurrectionary deed*, destined to affirm socialist principles by acts, is the most efficacious means of propaganda." The first acts were rural insurrections, intended to arouse the illiterate masses of the Italian countryside. After the insurrections failed, Anarchist activism tended to take the form of individual deeds of protest by terrorists, who would attempt to kill ruling figures in the hope of demonstrating the vulnerability of the structure of authority and inspiring the masses by their self-sacrifice. In this way, between 1890 and 1901, a series of symbolic murders were enacted; the victims included King Umberto I of Italy, the empress Elizabeth of Austria, President Carnot of France, President McKinley of the United States, and Antonio Cánovas del Castillo, the prime minister of Spain. This brief but dramatic series of terrorist acts established the image of the Anarchist as a mindless destroyer; after 1901, however, the Anarchists continued to practice widespread terrorism only in such countries as Spain and Russia, where the general political atmosphere was conducive to violence.

Anarchism and terrorism

During the 1890s, especially in France, Anarchism was adopted as a philosophy by avant-garde painters and writers. Gustave Courbet had already been a disciple of Proudhon; among those who in the 1890s accepted an Anarchist philosophy were Camille Pissarro, Georges Seurat, Paul Signac, Paul Adam, Octave Mirbeau, Laurent Tailhade, and, at least as a strong sympathizer, Stéphane Mallarmé. At the same time in England, Oscar Wilde declared himself an Anarchist and, under Kropotkin's inspiration, wrote his libertarian essay, "The Soul of Man under Socialism" (1891).

The artists were attracted by the individualist spirit of Anarchism. By the mid-1890s, however, the more militant Anarchists in France began to realize that an excess of individualism had tended to detach them from the workers they sought to liberate. Anarchists, indeed, have always found it difficult to reconcile the claims of general human solidarity with the demands—equally insistent—of the individual who desires freedom. Some Anarchist thinkers, such as the German Max Stirner, who published *The Ego and His Own* in 1845, have refused to recognize any limitation on the individual's right to do as he will or any obligation to act socially; and even those who accepted Kropotkin's socially oriented doctrines of Anarchist communism have in practice been reluctant to create forms of organization that threatened their freedom of action or seemed likely to harden into institutions.

In consequence, although a number of international Anarchist congresses were held (the most celebrated being those of London in 1881 and of Amsterdam in 1907), no effective worldwide organization was created, even though by the end of the century the Anarchist movement had spread to all continents and was united by informal links of correspondence and friendship between the leading figures. National federations were weak even in countries where there were many Anarchists, such as France and Italy, and the typical unit of organization was the small group dedicated to propaganda by deed or word. Such groups engaged in a wide variety of activities; in the 1890s many of them concentrated on setting up experimental schools and communities that attempted to live out Anarchist principles.

Revolutionary Syndicalism. In France, where individualist trends had been most pronounced and public reaction to terrorist acts had imperilled the very existence of the movement, an effort was made to acquire a mass following. The Anarchists infiltrated the trade unions. They were particularly active in the *bourses du travail* ("labour exchanges"), local groupings of unions, originally set up to find work for their members, that appealed to the Anarchist ideal of decentralization. In 1892 a national confederation of *bourses du travail* was formed, and by 1895 the Anarchists, led by Fernand Pelloutier, Émile Pouget, and Paul Delesalle, had gained effective control and were developing the theory and practice of working class activism that became known as Anarcho-Syndicalism, or Revolutionary Syndicalism.

Anarchism and the workers

The Anarcho-Syndicalists argued that the traditional function of trade unions—to struggle for better wages and working conditions—was not enough. The unions should become militant organizations dedicated to the destruction of capitalism and the state. They should aim to take over factories and utilities, which would then be operated by the workers. In this way the union or syndicate would have a double function—as an organ of struggle under the present dispensation and as an organ of administration after the revolution. To sustain militancy, an atmosphere of incessant conflict should be induced, and the culmination of this strategy should be the general strike. Many of the Syndicalists believed that such a massive act of noncooperation would bring about what they called "the revolution of folded arms," resulting in the collapse of the state and the capitalist system. But, although partial general strikes, with limited objectives, were undertaken in France and elsewhere with varying success, the millennial general strike aimed at overthrowing the social order in a single blow was never attempted. The Anarcho-Syndicalists acquired great prestige among the workers of France and, later, of Spain and Italy, because of their generally tough-minded attitude at a time when working conditions were bad and employers tended to respond brutally to union activity. After the great French trade-union organization, the Confédération Générale du Travail (CGT), was founded in 1902, their militancy enabled the Anarchists to retain control of the organization until 1908 and to wield considerable influence on its activities until after World War I.

Like Anarchism, Revolutionary Syndicalism proved attractive to certain intellectuals, notably Georges Sorel, whose *Réflexions sur la violence* (1908; Eng. trans., *Reflections on Violence*, 1914 and 1950) was the most important literary work to emerge from the movement. He argued the importance of the general strike as a social myth. The more purist Anarchist theoreticians were disturbed by the monolithic character of Syndicalist organizations, which they feared might create powerful interest structures in a revolutionary society. At the International Anarchist Congress at Amsterdam in 1907, a crucial debate on this issue took place between the young Revolutionary Syndicalist Pierre Monatte and the veteran Anarchist Enrico Malatesta. It defined a division of outlook that still lingers in what remains of the historic Anarchist movement, which has always included individualist attitudes too extreme to admit any kind of large-scale organization.

Revolutionary Syndicalism did transform Anarchism, for a time at least, from a tiny minority current into a movement with considerable mass support, even though most members of Syndicalist unions were sympathizers and fellow travellers rather than committed Anarchists. In 1922 the Syndicalists set up their own International with its headquarters in Berlin, taking the historic name of the International Workingmen's Association; it still survives, with headquarters in Stockholm. When it was established, the organizations that formed it could still boast considerable followings. The Unione Sindicale Italiana had 500,000 members; the Federación Obrera Regional Argentina, 200,000 members; the Portuguese Confederação General de Trabalho, 150,000 members; and the German Freie Arbeiter Union, 120,000 members. There were smaller organizations in Chile, Uruguay, Denmark, Norway, Holland, Mexico, and Sweden. In Britain, the influence of Syndicalism was shown most clearly in the Guild Socialist movement that flourished briefly in the early years of the present century. In the United States, Revolutionary Syndicalist ideas were manifested in the Industrial Workers of the World (IWW), which in the years immediately before and after World War I played a vital part in organizing American miners, loggers, and unskilled workers; but only a small minority of the IWW militants were ever avowed Anarchists.

Syndicalist organizations

Anarchism in Spain. The reconciliation of Anarchism and Syndicalism was most complete and most successful in Spain; for a long period the Anarchist movement in that country remained the most numerous and the most powerful in the world. The first known Spanish Anarchist, Ramón de la Sagra, a disciple of Proudhon, founded the world's first Anarchist journal, *El Porvenir*, in La Coruña in 1845; it was quickly suppressed. Mutualist ideas were later publicized by Pi y Margall, a federalist leader and the translator of many Proudhon books; during the Spanish revolution of 1873, Pi y Margall attempted to establish a decentralist, or "cantonalist," political system on Proudhonian lines. In the end, however, the influence of Bakunin was stronger. In 1868 his Italian disciple, Giuseppe Fanelli, visited Barcelona and Madrid, where he established branches of the International. By 1870 they had 40,000 members, and in 1873 the movement numbered about 60,000, organized mainly in working men's associations. In 1874 the Anarchist movement in Spain was forced underground, a phenomenon recurring often in subsequent decades. It flourished, nevertheless, and Anarchism became the favoured type of radicalism among two very different groups, the factory workers of Barcelona and other Catalan towns and the impoverished peasants who toiled on the absentee-owned estates of Andalusia.

As in France and Italy, the movement in Spain during

the 1880s and 1890s was inclined toward insurrection (in Andalusia) and terrorism (in Catalonia). It retained its strength in working class organizations because the courageous and even ruthless Anarchist militants were often the only leaders who would stand up against the army and the employers, who hired squads of gunmen to engage in guerrilla warfare with the Anarchists in the streets of Barcelona. The workers of Barcelona were finally inspired by the success of the French CGT to set up a Syndicalist organization, Solidaridad Obrera (Workers' Solidarity). Established in 1907, Solidaridad Obrera quickly spread throughout Catalonia, and in 1909, when the Spanish army tried to conscript Catalan reservists to fight against the Riffs in Morocco, it called a general strike. "La Semana Tragica," "the Tragic Week" of largely spontaneous violence that followed (with hundreds dead and 50 churches and monasteries destroyed), ended in violent repression. Tortures of Anarchists in the fortress of Montjuich and the execution of the internationally celebrated advocate of free education Francisco Ferrer led to worldwide protests and the resignation of the conservative government in Madrid. These events also resulted in a congress of Spanish trade unionists at Seville in 1910, which founded the Confederación Nacional del Trabajo (CNT).

The CNT, which included the majority of organized Spanish workers, was dominated throughout its existence by the Anarchist militants; these in 1927 founded their own activist organization, the Federación Anarquista Iberica (FAI). While there was recurrent conflict within the CNT between moderates and FAI activists, the atmosphere of violence and urgency in which radical activities were carried on in Spain ensured that the more extreme leaders, such as Garcia Oliver and Buenaventura Durutti, tended to wield the decisive influence. The CNT was a model of Anarchist decentralism and antibureaucratism: its basic organizations were not national unions but *sindicatos unicos*, which brought together the workers of all trades and crafts in a certain locality; the national committee was elected each year from a different locality to ensure that no individual served more than one term; and all delegates were subject to immediate recall by the members. This enormous organization, which claimed 700,000 members in 1919, 1,600,000 in 1936, and more than 2,000,000 during the civil war, employed only one paid secretary. Its day-to-day work was carried on in their spare time by workers chosen by their fellows. This meant that the Spanish Anarchist movement was not dominated by the *déclassé* intellectuals and self-taught printers and shoemakers who were so influential in other countries. One consequence was that the Spanish movement contributed nothing original to the ideological literature of Anarchism; like Bakunin, whom they favoured most among the classic libertarian thinkers, the Spanish Anarchists developed their attitudes in action rather than in writing.

Moderates and activists

The CNT and the FAI, which remained clandestine organizations under the dictatorship of Primo de Rivera, emerged into the open with the abdication of King Alfonso XIII in 1931. Their antipolitical philosophy led them to reject the republic as much as the monarchy it had replaced, and between 1931 and the military rebellion led by Francisco Franco in 1936 there were several unsuccessful Anarchist risings. In 1936 the Anarchists, who over the decades had become expert urban guerrillas, were mainly responsible for the defeat of the rebel generals in both Barcelona and Valencia, as well as in country areas of Catalonia and Aragon; and for many early months of the Civil War they were in virtual control of eastern Spain, where they regarded the crisis as an opportunity to carry through the social revolution of which they had long dreamed. Factories and railways in Catalonia were taken over by workers' committees, and in hundreds of villages in Catalonia, Levante, and Andalusia the peasants seized the land and established libertarian communes like those described by Kropotkin in *La Conquête du pain*. The internal use of money was abolished, the land was tilled in common, and village products were sold or exchanged on behalf of the community in general, with each family receiving an equitable share of food and other necessities. An idealistic Spartan fervour characterized these communities, which often consisted of illiterate labourers; intoxicants, tobacco, and sometimes even coffee were renounced; and millenarian enthusiasm took the place of religion, as it has often done in Spain. The reports of critical observers suggest that at least some of these communes were efficiently run and more productive agriculturally than the villages had been previously.

Anarchists in the Spanish Civil War

The Spanish Anarchists failed during the Civil War largely because, expert though they were in spontaneous street fighting, they did not have the discipline necessary to carry on sustained warfare; the columns they sent to various fronts were unsuccessful in comparison with the Communist-led International Brigades. In December 1936 four leading Anarchists took posts in the Cabinet of Francisco Largo Caballero, radically compromising their antigovernmental principles. They were unable to halt the trend toward left-wing totalitarianism encouraged by their enemies the Communists, who were numerically far fewer but politically more influential. In May 1937 bitter fighting broke out in Barcelona between Communists and Anarchists. The CNT held its own on this occasion, but its influence quickly waned. The collectivized factories were taken over by the central government, and many agricultural communes were destroyed by Franco's advance into Andalusia and by the hostile action of General Lister's Communist army in Aragon. In January 1939 the Spanish Anarchists were so demoralized by the compromises of the Civil War that they were unable to mount a resistance when Franco's forces marched into Barcelona; the CNT and FAI became phantom organizations in exile.

Decline of the Anarchist movement. By this time the movement outside Spain had been either destroyed or greatly diminished as a result of two developments: the Russian Revolution and the rise of right-wing totalitarian regimes. Though the most famous Anarchist leaders, Bakunin and Kropotkin, had been Russian, the Anarchist movement had never been strong in Russia, partly because the more numerous Socialist Revolutionary Party (the Narodniki) had adopted Bakuninist ideas while remaining essentially a constitutional party. After the 1917 revolution the small anarchist groups that emerged in Petrograd (now Leningrad) and Moscow were powerless against the Bolsheviks, and Kropotkin, who returned from exile, found himself without influence. Only in the south did N.I. Makhno, a peasant Anarchist, raise an insurrectionary army that by brilliant guerrilla tactics held a large part of the Ukraine from both the Red and the White armies; but the social experiments developed under Makhno's protection were rudimentary, and when he was driven into exile in 1921 the Anarchist movement became extinct in Russia.

The competition of Communism

In other countries, the prestige of the Russian revolution enabled the new Communist parties to win much of the support formerly given to the Anarchists, particularly in France, where the CGT passed permanently into Communist control. The large Italian Anarchist movement was destroyed by Benito Mussolini's Fascist government in the 1920s, and the small German Anarchist movement was smashed by the Nazis in the 1930s. In Japan, Anarchism had emerged during the Russo-Japanese war of 1904–05, when the Socialist leader Shusui Kotoku became converted by reading Kropotkin in prison; Kotoku and other Anarchists were executed in 1911 for their involvement in a plot against the Emperor, but, after World War I, new Anarchist organizations appeared, including a Black Federation and a Syndicalist federation. After the Japanese invasion of Manchuria in 1931, the imperial government began to suppress all left-wing groups, and the Anarchist movement was finally destroyed in 1935 after a secret society, the Anarchist Communist Party, had been accused of plotting armed insurrection.

Anarchism in the Americas suffered similar reverses. In the United States, a native and mainly nonviolent tradition developed during the 19th century in the writings of

Henry David Thoreau, Josiah Warren, Lysander Spooner, and Benjamin Tucker (editor of *The Anarchist*, a highly individualist journal published in Boston). Activist Anarchism in the U.S. was mainly sustained by immigrants from Europe, including Johann Most (editor of *Die Freiheit*), Alexander Berkman, and Emma Goldman, whose *Living My Life* gives a picture of radical activity in the United States at the turn of the century. Anarchism appeared as a dramatic element in American life in 1886, when seven policemen were killed in the Haymarket bombing in Chicago and four Anarchist leaders were executed—unjustly, as later investigations revealed. In 1901 a Polish Anarchist, Leon Czolgosz, assassinated President McKinley. In 1903 the U.S. congress passed a law to bar foreign Anarchists from the country and to deport alien Anarchists found within it. In the repressive mood that followed World War I the Anarchists, like the IWW, were suppressed; Alexander Berkman, Emma Goldman, and many others were imprisoned and deported.

In Latin America, strong Anarchist elements were involved in the Mexican Revolution. The Syndicalist teachings of Ricardo Flores Magon influenced the peasant revolutionism of Emiliano Zapata. After the deaths of Zapata in 1919 and Flores Magon in 1922, the revolutionary image in Mexico, as elsewhere, was taken over by the Communists. In Argentina and Uruguay, considerable Anarcho-Syndicalist movements existed early in the 20th century, but they too were greatly reduced by the end of the 1930s through intermittent repression and the competition of Communism.

CONTEMPORARY CURRENTS

In the 1970s the theories of Anarchism aroused more interest and sympathy than at any time since the Russian revolution. The same cannot be said of the Anarchist movement itself. After World War II, Anarchist groups and federations re-emerged in almost all countries where they had formerly flourished—the notable exceptions being Spain and the Soviet Union—but these organizations wielded little influence compared to that of the broader movement inspired by libertarian ideas.

Wider influence of anarchist ideas

Such a development is not inappropriate, since Anarchists have never stressed the need for organizational continuity, and the cluster of social and moral ideas that are identifiable as Anarchism has always spread beyond any clearly definable movement. The Russian writer Count Leo Tolstoy refused to call himself an Anarchist because those who used the title at that time in Russia were terrorists; nevertheless, he had developed out of his rational Christianity a form of pacifist radicalism that rejected the state and all forms of government, called for the simplification of life in the name of moral regeneration, and sought to replace property by free communism. An impressive example of the breadth of Anarchist influence is Mahatma Gandhi, who based his strategy of nonviolent civil disobedience in South Africa and India on the teachings of nonviolent Anarchists such as Tolstoy and Thoreau and who remembered his reading of Kropotkin when he devised for India the plan of a decentralized society based on autonomous village communes. Gandhi's village India has not come into being, but the movement known as Sardovaya, led by Vinova Bhave and Jayaprakash Narayan, has been working toward it through *gramdan*—community ownership of land. By 1969 a fifth of the villages of India had declared for *gramdan*, and, while this remained largely a matter of unrealized gestures, it represented perhaps the most extensive commitment to basic Anarchist ideas in the contemporary world.

Anarchism's appeal to intellectuals

In the West the appeal of Anarchism has been strongest, at least since 1917, among the intellectuals. Kropotkin's arguments for decentralization have wielded their strongest influence among writers on social planning such as Patrick Geddes, Lewis Mumford, and Paul Goodman—himself a declared Anarchist. Anarchist treatises from Godwin onward have contributed to the progressive movement in education, and the most influential of the many books written by the libertarian critic Sir Herbert Read was his *Education Through Art* (1943). But in the 1960s and 1970s Anarchism became popular among rebellious students and the left in general because of the values it opposed to those of the increasingly technological cultures of western Europe, North America, and Japan.

Here the mediating thinker was undoubtedly Aldous Huxley, who anticipated many elements of the "counterculture" of the 1960s and 1970s. In *Brave New World* (1932) Huxley had presented a warning vision of the kind of mindless, materialistic existence a society dominated by technology might produce. In his "Foreword" to the 1946 edition, he stated his belief that only by radical decentralization and simplification and by a politics that was "Kropotkinesque and co-operative" could the dangers implicit in modern society be avoided. In his later writings he explicitly accepted the validity of the Anarchist critique of existing society.

The emergence of Anarchist ideas in a wider frame of reference began in the American civil rights movements in the 1950s, with their recognition of the need to resist injustice through other than legal channels in the name of a morality different from that recognized by constituted authority. By the end of the 1950s the new radicalism in the United States, Europe, and Japan had begun to take a course separating it from the narrower issues of the civil rights movement; it shifted toward a general criticism of the elitist structure and materialist goals of modern industrial societies—Communist as well as capitalist. Within this movement there was a limited revival of traditional Anarchism—exemplified in the sudden popularity of the American writer Paul Goodman and the emergence in Britain of a sophisticated monthly, *Anarchy*, that applied Anarchist ideas to the whole spectrum of modern life.

Anarchism and contemporary rebellion

In all of this, however, Anarchism was only one strand in what can be described as a climate of rebellion rather than an ideology. Anarchist ideas were mingled with strains of Leninism, early Marxism, unorthodox psychology, and often with elements of mysticism, neo-Buddhism, and Tolstoyan Christianity. None of the leaders of the student rebellions in the United States and Germany or the militants of Zengakuren in Japan could be called in any literal and complete sense an Anarchist, although they had read Bakunin as well as Marx and "Che" Guevara. In Paris, the leaders of the Anarchist Federation admitted that they had no influence at all on the strikes and street fighting in 1968 that seemed to threaten the very existence of the French state.

To all these movements, which rejected the old parties of the left as strongly as they did the existing political structure, the appeal of Anarchism was strong and understandable. The Anarchist outlook, in its insistence on spontaneity, on theoretical flexibility, on simplicity of life, on love and anger as complementary and necessary components in both social and individual action, appeals to those who reject the impersonality of institutions and the calculations of political parties. The Anarchist rejection of the state, the insistence on decentralism and local autonomy, found strong echoes among those who talked of participatory democracy. The recurrence of the theme of workers' control of industry in so many manifestos of the 1960s, notably during the Paris insurrection of 1968, showed the enduring relevance of Anarcho-Syndicalist ideas.

The Anarchist insistence on direct action had an almost universal appeal to the radicals of the 1960s and 1970s, who advocated extraparliamentary action and confrontation. Some student groups in France, the United States, and Japan accepted Bakunin's pan-destructionism, holding that existing society was so corrupt that it had to be swept completely away. They were fond of quoting a sentence of Bakunin:

> There will be a qualitative transformation, a new living, life-giving revelation, a new heaven and a new earth, a young and mighty world in which all our present dissonances will be resolved into a harmonious whole.

This kind of secular apocalypticism, which envisaged total violence as the paradoxical way to total reconcilia-

tion, linked many of the radicals with the Anarchists who took to terrorism in the 1880s and 1890s. Now, as then, violence proceeded from theory to action, as the upsurge of terrorist bombings and urban-guerrilla activity demonstrated at the end of the 1960s.

Anarchism as an ideal. But there were others who believed with Tolstoy and Gandhi that a moral change must come first and who sought revolution by evolution. Paul Goodman, with his proposals for urban reform as a step toward a freer society, was one of them. Some turned to the creation of models of a different kind of society in the form of utopian communities. This has been a recurrent feature of radical movements, appearing in the United States during the mid-19th century, in Britain as a form of constructive pacifist protest during World War II, and in the United States in the 1970s as a way of manifesting one's rejection of normal life-styles.

Community reform

Experience suggests that utopian experiments and radical actions are not likely to achieve that wholly nongovernment society of which libertarians have dreamed. The true value of their vision was stated by the Anarchist poet Herbert Read in his last book, *The Cult of Sincerity* (1968):

> My understanding of the history of culture has convinced me that the ideal society is a point on a receding horizon. We move steadily towards it but can never reach it. Nevertheless we must engage with passion in the immediate strife . . .

It is precisely as an ideal, as a touchstone to judge the existing world, that the Anarchist vision is useful. It corresponds to a recurrent and necessary strain in human thought—the revulsion against regimentation, against large organizations, against complexity and luxury. The insights of the Anarchist are likely to find their maximum usefulness in urban and rural planning, in the development of community relations based on full participation, in integrated education, and perhaps most of all in encouraging what Read calls a "process of individuation, accomplished by general education and personal discipline." Anarchism is a moral and social doctrine before it is a political one; it stands as a permanent reminder of the perils of national and corporate giantism and of the virtues of local interests and loyalties. It teaches the vigilance by which man may be able to avoid such bleak utopias as those of Aldous Huxley and George Orwell.

BIBLIOGRAPHY. The principal anarchist classics are mentioned in the text. In recent years many have been reprinted, notably WILLIAM GODWIN, *Enquiry Concerning Political Justice and Its Influence on Morals and Happiness*, 3 vol. (1793, reprinted 1946); PETER KROPOTKIN, *Memoirs of a Revolutionist*, 2 vol. (1899, reprinted 1968), and *Mutual Aid* (1890–96, reprinted 1939); and PIERRE JOSEPH PROUDHON, *Qu'est-ce que la propriété?* (1840; Eng. trans., *What Is Property?*, 2 vol., 1898–1902, reprinted 1966). G.P. MAXIMOFF (ed.), *The Political Philosophy of Bakunin: Scientific Anarchism* (1953), is a good anthology of Bakunin's writings; and VERNON RICHARDS (ed.), *Errico Malatesta: His Life and Ideas* (1965), contains a collection of writings by Malatesta, many of which were previously unavailable in English. TOLSTOY'S principal writings in the anarchist vein, *The Kingdom of God Is Within You*, *The Slavery of Our Time*, and *What I Believe*, are available in standard editions. HERBERT EDWARD READ'S anarchist writings are mainly collected in *Anarchy and Order* (1954, reprinted 1971) and *The Cult of Sincerity* (1968); his *Education Through Art* (1963) should also be consulted. The best general accounts of anarchist thought and history are GEORGE WOODCOCK, *Anarchism* (1962); JAMES JOLL, *The Anarchists* (1965); DANIEL GUERIN, *L'Anarchisme, de la doctrine à l'action* (1965; Eng. trans., *Anarchism: From Theory to Practice*, 1970); and APRIL CARTER, *The Political Theory of Anarchism* (1971). A useful collection of essays on contemporary trends is DAVID E. APTER and JAMES JOLL, *Anarchism Today* (1970); two recent anthologies that contain writings drawn from the whole libertarian tradition are LEONARD I. KRIMERMAN and LEWIS PERRY (eds.), *Patterns of Anarchy* (1966); and IRVING L. HOROWITZ (ed.), *The Anarchists* (1964). P. ELTZBACHER, *Der Anarchismus* (1900; Eng. trans., 1908), is a classic account of leading anarchist thinkers.

On anarchism in Russia: PAUL AVRICH, *The Russian Anarchists* (1967); FRANCO VENTURI, *Il populismo russo* (1952; Eng. trans., *Roots of Revolution: A History of the Populist and Socialist Movements in Nineteenth Century Russia*, 1960); and DAVID FOOTMAN, *Civil War in Russia* (1961). On anarchism in Spain: GERALD BRENAN, *The Spanish Labyrinth: An Account of the Social and Political Background of the Civil War*, 2nd ed. (1950, paperback 1960); BURNETT BOLLOTEN, *The Grand Camouflage: The Communist Conspiracy in the Spanish Civil War* (1961); and HUGH THOMAS, *The Spanish Civil War* (1961). On anarchism in Italy: RICHARD HOSTETTER, *The Italian Socialist Movement*, vol. 1 of a projected 3-volume work (1958). On anarchism in the United States: JAMES J. MARTIN, *Men Against the State: The Expositors of Individualist Anarchism in America, 1827–1908* (1953); and RUDOLF ROCKER, *Pioneers of American Freedom* (1949). On anarchism in France: EUGENIA W. HERBERT, *The Artist and Social Reform: France and Belgium, 1885–1898* (1961).

(G.W.)

Anatolia, Ancient

The name Anatolia has come to be applied to the whole Asiatic territory of the Turkish Republic, an approximate rectangle of largely mountainous country, extending latitudinally over a thousand miles from the Aegean coast to the borders of Russian Armenia and Iranian Azerbaijan. Its western half is a huge peninsula connecting Asia with Europe. It consists of a central plateau with an average altitude of 3,000 feet (900 metres), separated from the sea on its north and south sides by higher mountain barriers and to the west by fingerlike ridges with river valleys between. The eastern half of Anatolia is comprised primarily of a complex of mountain ranges with high peaks and an alpine climate. To the south, however, it also includes a belt of piedmont steppe country, traversed by the upper courses of the Tigris and Euphrates.

This article covers the history of ancient Anatolia until it came under Roman rule in the 1st century BC. The main divisions of the article are as follows:

I. Anatolia to *c.* 1180 BC
- The Neolithic and Chalcolithic ages
  - The Neolithic
  - The Chalcolithic
- The Early Bronze Age
  - Early Bronze Age I
  - Early Bronze Age II
  - Early Bronze Age III
- The Middle Bronze Age
- The Hittites
  - The Hittite occupation of Anatolia
  - The Old Hittite Kingdom
  - The Middle Kingdom
  - The Hittite Empire to *c.* 1180 BC

II. Anatolia from *c.* 1180 BC until the Roman conquest
- From the fall of the Hittite Empire until the Achaemenian period
  - Phrygia from *c.* 1180 to 700 BC
  - The neo-Hittite states (from *c.* 1180 to 700 BC)
  - The Cimmerians, Lydia, and Cilicia, *c.* 700–547 BC
  - Greek colonies on the Anatolian coasts, *c.* 1180–547 BC
- Anatolia in the Achaemenian and Hellenistic periods
  - The Anatolian Greeks in the Achaemenian period
  - Caria, Lycia, and Cilicia in the Achaemenian period
  - Anatolia in the Hellenistic Age (334–*c.* 30 BC)

I. Anatolia to c. 1180 BC

THE NEOLITHIC AND CHALCOLITHIC AGES

The earliest evidence of human occupation in Anatolia is at present confined geographically to a comparatively small area inland from Antalya on the Mediterranean coast. Of special significance are rock engravings of animals on the walls of caves, which suggest a relationship with the Upper Paleolithic art of western Europe. Associated with these are rock shelters, the stratified occupational debris of which may, after further study, finally clarify the transitional phases between cave-dwelling society and the Neolithic economy of the first agricultural communities. In the technical terms applied by archaeologists to these and subsequent phases of Near Eastern prehistory, there is at present a confusing lack of uniformity, but for the present purposes they may be simplified as follows: the first indications of the beginning of the Neolithic Revolution (from food-gathering to food producing) are found in approximately 9000 BC; the two millennia that followed are often termed Proto-Neolithic to

Chronological summary

distinguish them from the Mesolithic or Terminal Paleolithic that preceded them; the true Neolithic begins in 7000 BC, by which time farming and stockbreeding were well established, and lasts until 5600 BC; the Neolithic is succeeded by the Chalcolithic, during which metal weapons and tools gradually took their place beside their stone prototypes, and painted pottery came generally into use; the Chalcolithic ends in the final centuries of the 4th millennium BC, when the great dynastic civilizations of Egypt and Mesopotamia were already foreshadowed by the invention of writing.

The Neolithic. It was long understood that the origins of agriculture and stockbreeding should be sought in those areas of the Near East where the wild ancestors of modern food grains and the natural habitats of domesticable animals were to be found. This line of inquiry pointed to the well-watered uplands around the fringe of the Fertile Crescent: Iraqi Kurdistan, northern Syria, and the Levant. There, indeed, the first discoveries were made of Neolithic farming communities. To the north in Anatolia, they extended to the great coastal plain of Cilicia, connected by easy routes with northern Syria, and to the eastern slopes of the Amanus Mountains. The plateau itself and the western provinces beyond Taurus, however, at first made little contribution to this investigation; and it was long thought that they had remained unoccupied until the beginning of the Chalcolithic period. Knowledge of the Anatolian Neolithic, then, depended for the time being upon sequences of pottery and small artifacts from exiguous soundings at such peripheral sites as Mersin, Tarsus, and Sakcagöz. Since the 1960s, however, the picture has been changed by British-sponsored excavations at two plateau sites: Hacilar, near Lake Burdur, and Çatalhüyük in the Konya Plain. Neither of these dated back as far as the Proto-Neolithic period, of which such striking evidence had been found at Jericho in Jordan. Hacilar, however, showed an earliest occupation of about 7000 BC by a people still unfamiliar with the use of pottery, though living in mud-brick houses with plastered walls and floors, painted and burnished like those in contemporary Jericho. Afterward abandoned for well over a thousand years, Hacilar was reoccupied in the latest phase of the Neolithic by villagers of a far more sophisticated culture. The houses were symmetrically arranged, and a striking collection of seminaturalistic figurines threw new light on Neolithic art and symbolism. The gap between these two widely separated periods was filled by the discovery (1961–63) at Çatalhüyük of a Neolithic settlement covering 32 acres and occupied from the mid-7th to the mid-6th millennia.

Hacilar and Çatalhüyük

The discoveries at Çatalhüyük not only amplified but transformed the whole conception of human behaviour in Neolithic times. In the town, houses were built of sun-dried brick, closely contiguous like the cells of a honeycomb, but each had several rectangular rooms similarly planned and was accessible only by a wooden ladder from its flat roof. The contiguous roofs provided space for the communal life of the inhabitants. Some of these buildings appear to have been religious shrines, elaborately ornamented with heads or horns of animals, either real or imitated in plaster. The walls were decorated with coloured murals, repeatedly repainted after replastering, and the designs closely resembled the cave paintings of the Paleolithic era. As a source of information about early man's activities, appearance, dress, and even religion, these paintings are of great significance. Other arts and crafts were well attested. Human and animal figurines were carved in stone or modelled in clay. Bone was used for tools and implements, sometimes with finely carved ornamentation. Weapons included polished maces, arrows, and lances with tanged obsidian heads. Impressions of mats and baskets were found, as well as implements used in spinning and weaving. Miraculously, fragments of actual textiles were recovered and preserved. The presence of Mediterranean shells and of metal ores and pigments not locally available suggested extensive trade. Undecorated pottery was in use throughout the life of the settlement, its shapes often imitating those of wooden vessels, examples of which were found intact.

Agriculture evidently formed the main basis of the economy at Çatalhüyük. The site, on a river subject to regular flooding, suggested that irrigation may have been practiced. No proof was found of domesticated animals, but the bones of wild cattle, deer, and boar confirmed the implications of the wall paintings that hunting was still widespread. Generally speaking, nothing suggests that this precocious culture had its origin elsewhere than in Turkey, and the peoples of the Anatolian plateau may well have played a leading part in the Neolithic Revolution.

The Anatolian origin of the Neolithic

The Chalcolithic. The transition from the Neolithic to the Chalcolithic phase of cultural evolution is thought to have taken place gradually in the mid-6th millennium BC. At most sites where its progress can be traced, no perceptible break occurs in the continuity of occupation, and there is little reason to assume any major ethnographic upheaval. Archaeologically, the most conspicuous innovation is the decoration of pottery with coloured paint, the first clumsy attempts at which can be seen at Mersin. At Hacilar, however, expertise in this craft seems to have been born fully fledged, and the last four occupations of the village (designated Levels V-I) are characterized by some of the most competently and attractively decorated pottery in prehistoric Anatolia. Village architecture at this period is undistinguished, but it provides evidence for the necessity of communal defense. A line of contiguous houses created a peripheral fortification. Their heavy ground-floor chambers had no doorways and were evidently entered by ladders from a more fragile upper story. Metallurgy was beginning to be understood, and copper was used for pins and simple implements. For the later Chalcolithic, additional evidence is available from a deep sounding at Beycesultan, which produced a cache of more sophisticated copper tools and a silver ring, the earliest known use of that metal. For improvements in architecture at this period, however, one must return to Mersin, where one of its latest phases (Level XVI) is represented by a neatly planned and constructed fortress. The steeply revetted slope of the mound was crowned by a continuous defensive wall, pierced by slit windows and entered through a gateway protected by flanking towers. Inside, there was formally arranged accommodation for the garrison and other evidence of military discipline as conceived in 4500 BC.

Thus, in Anatolia, prehistoric man had emerged from the Neolithic having made much progress in social organization. He cultivated crops, herded livestock, lived in brick houses, made vessels of clay, stone, wood, or basketwork, and fashioned weapons of bone or flint. He traded for the raw materials to manufacture his weapons and personal ornaments, and his religious beliefs found expression in sculpture and painting. During the Chalcolithic period few improvements seem to have been made in this standard of living, save for the benefits inherent in the experimental use of metal. The most notable development during these two millennia was the growth and wide diffusion of agricultural societies.

THE EARLY BRONZE AGE

The period following the Chalcolithic in Anatolia is now generally referred to as the Bronze Age. In its earlier phases, the predominant metal was in fact pure copper, but the term Copper Age, at one time in vogue, created confusion and has since been discarded. Archaeological convention divides the Bronze Age into three subphases, early, middle, and late. Its beginning, toward the end of the 4th millennium BC, corresponded in Egypt to the Predynastic period and in Mesopotamia to the Protoliterate. It lasted until late in the 2nd millennium. In its earlier stages it is possible to recognize a number of well-defined cultural provinces, distinguished by slight disparities in their archaeological remains: a northwest province, represented by the excavations at Troy; a central Anatolian province, well-documented by the soundings at Alişar, Alaca Hüyük, and elsewhere; an extensive southwestern province, best characterized at Beycesultan; and a Cilician province, represented by Mersin and neighbouring Tarsus. Other, less clearly defined provinces are found in

Subphases of the Bronze Age

the Konya Plain, connected with Cilicia by the Taurus passes, and in the Pontic area. The combined results of excavations in all of these localities have made possible a further chronological division of the Early Bronze Age alone into three successive phases, numbered from the earliest upward.

Early Bronze Age I. In the stratigraphical sequence at Troy, for instance, Early Bronze Age I designates the earliest settlement—a small fortress, hardly more than 100 yards in diameter (Troy I). Supplemented by comparable finds at Poliochni and Thermi on the offshore islands of Lemnos and Lesbos, its contents reveal some characteristics of the new culture. Examples are found of houses built along a megaron plan—a rectangular hall with a central hearth and an open porch at one end. The toughening of copper by the admixture of tin was understood, at least in the Aegean and Mediterranean trading centres. The dead were buried in cemeteries outside the walls, often in large clay vessels. Pottery was invariably monochrome, black or red, and usually burnished to a slight shine.

Early Bronze Age II. The citadel at Troy, rebuilt on a more ambitious scale in the second phase of the Early Bronze Age (Troy II), is characterized by a more advanced metal industry and a more active commercial life. There, hoards of jewelry and other luxury objects were found by Heinrich Schliemann beneath the palace and christened by him "Priam's Treasure." Also purporting to have originated in the Troadic province at this period is the Dorak Treasure, of which only drawings are now available. These suggest an even higher standard of craftsmanship and a wider variety of precious materials. Nor was evidence of luxury restricted to the western peninsula. Contemporary with Troy II were the royal tombs discovered in the Early Bronze Age levels beneath the Hittite city at Alaca Hüyük. There, several generations of a ruling family were buried amid their funerary paraphernalia and private possessions. Men were buried with weapons, women with ornaments and toilet articles as well as domestic vessels and utensils, many of them in precious metals. The tombs themselves were rectangular pits enclosed by rough stone walls and roofed with timber. The tombs allotted sufficient space for wooden furniture, of which only the metal trappings and ornaments have survived. All-bronze furniture has since been found in a comparable group of tombs at Horoztepe, near Tokat to the northeast of Alaca. The combined inventory of artifacts from these various cemeteries and treasures emphasizes the advances that were made in metallurgy and craftsmanship after the Chalcolithic period. Most of the technical processes later used by the Hittites seem already to have been familiar to the metalsmiths of the Early Bronze II period. They included casting in closed molds (the cire perdue, or lost-wax, process), metal inlay, sweating and soldering, hammering and repoussé, granulation, filigree, and even cloisonné. The metals used included copper, bronze, silver, gold, electrum, lead, and finally iron, which then was far more valuable than gold. All these metals are obtainable in Anatolia, but tin had to be imported. Semiprecious stones and other materials used in association with them included rock crystal, carnelian, jasper, nephrite, obsidian, and meerschaum —all native to Anatolia. Ivory, amber, lapis lazuli, and turquoise were imported.

Royal tombs

The constant warfare among the Anatolian states at this time is emphasized by the number and variety of weapons found—daggers, swords, spears, and battle-axes—and also by examples of military architecture. In Troy itself, the city walls, with their great towered gateway constructed of mud brick in a timber framework on a substructure of dressed stone, seem almost disproportionately substantial to the buildings they protect. The main public building—probably a communal assembly hall—is a vast megaron hall with a roof span of more than 30 feet (nine metres). The presumed residential palace, however, is a much less pretentious building, and the whole enclosure is hardly 200 yards in diameter, giving the impression of a military fortress rather than a state capital. There are no obviously religious buildings. Such appear in the corresponding levels at Beycesultan, where they took the form of rectangular shrine chambers, arranged in pairs and perhaps dedicated to male and female deities. The ritual installations and accumulation of votive objects suggest an affinity with the tree, or pillar, cults of Crete.

Military architecture

Early Bronze Age III. The third phase of the Early Bronze Age (the final centuries of the 3rd millennium BC) appears to have been a time of major migrations. About 2300 BC a great wave of Indo-European invaders, speaking a Luwian dialect, seems to have swept over Anatolia from the west, occupying practically the whole southwestern part of the peninsula. Their progress was marked by widespread destruction, and for about a century, at sites where remains of this period have been found, there is every indication of a decline in prosperity. The northern and central provinces, however, remained unaffected. About 2200 BC, when contact with the Luwian-dominated southwest was re-established, profitable trade was resumed, particularly with Cilicia and the North Syrian markets beyond. At this point, at the beginning of the Middle Bronze Age, the attention of archaeologists returns to central Anatolia, particularly to the site of Kültepe (Kanesh), whose history was recorded by literate Assyrian colonists. The Assyrians occupied their own suburb outside Kanesh; but in the main city mound itself, Turkish excavations in the pre-*karum* levels have revealed some characteristics of the indigenous Anatolian population in this final phase of the Early Bronze Age. There is some impressive architecture, including a building thought to be a temple, which is planned around a square megaron hall with a huge circular hearth between four wooden columns, so much resembling the central feature of Mycenaean palaces a thousand years later that it is impossible to think of them as unrelated by tradition. For the rest, strange alabaster idols and fine polychrome pottery from Cappadocia seem purely Anatolian contributions, but other types of pottery and small objects suggest trade with Cilicia and even with Mesopotamia.

Indo-European invasion

THE MIDDLE BRONZE AGE

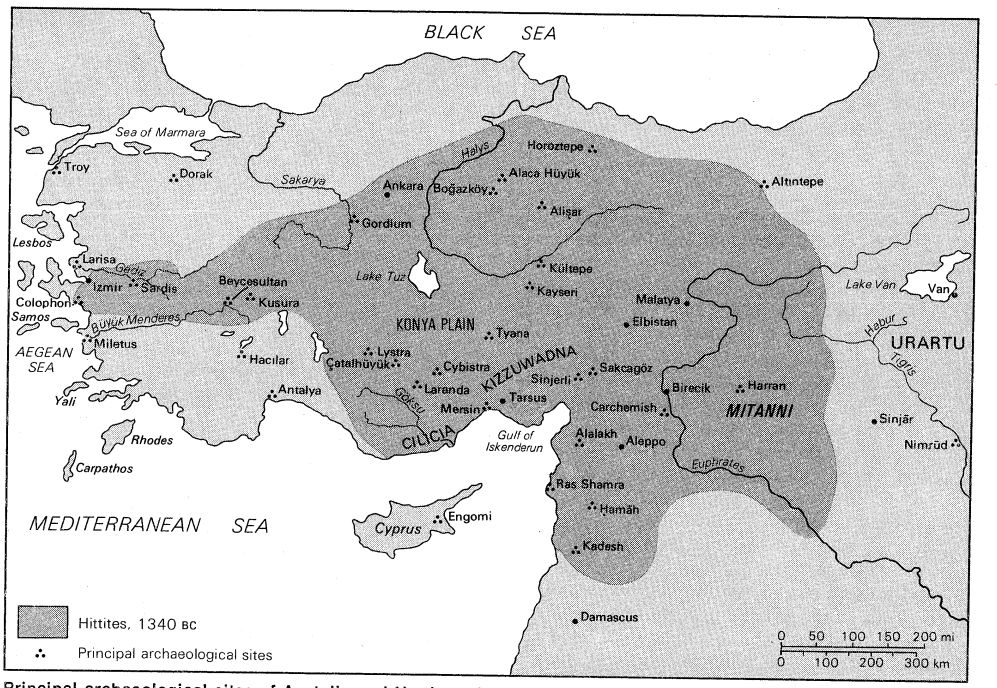
The Middle Bronze Age, beginning in the first decades of the 20th century BC, seems to have inaugurated a period of cultural progress and prosperity in the Anatolian cities. Again, the discoveries at Kültepe are invaluable. The merchants from Ashur occupied a walled suburb that had apparently already existed before their arrival. Some scholars have disagreed as to the status of the colony in relation to the indigenous principality, but most are now satisfied that it was based on mutual agreement rather than vassalage. Certainly the foreigners lived on good terms with their Anatolian neighbours, for there was frequent intermarriage. The *karum* must have resembled a sort of chamber of commerce, with authority to fix prices, settle debts, and arrange transport.

The *karum* at Kültepe

The history of the *karum* falls into two periods, divided by some disaster that resulted in the destruction of the suburb by fire, after which it was rebuilt. Because written records were kept, the approximate dates of these two occupations may be calculated by correlating textual references to the names of Assyrian kings whose dates are already known. The Level II occupation, which was the longer and more productive of the two, must have covered the reigns of Erishum, Sargon I, and Puzur-Ashur (*c.* 1950–1850 BC), while Level Ib was contemporary with that of Shamshi-Adad I (*c.* 1813–1781 BC).

A new chronological system, now in use, divides the Middle Bronze Age into four subphases: the first covering the two pre-Assyrian occupations of the Kültepe suburb (Levels III and IV); the second corresponding to Level II; the third lasting until the destruction of the Level Ib *karum* by a fire that ended the colony; and the fourth phase representing the interval between this event and the foundation of the Old Hittite Kingdom.

The successive occupations of the suburb are paralleled in contemporary building levels in the main city mound, where the palaces of local rulers have occasionally produced vitally important cuneiform texts. In addition to writing on clay, Anatolian scribes in the city also adopted



Principal archaeological sites of Anatolia and Northern Syria to *c.* 1340 BC.
Adapted from S. Lloyd, *Early Highland Peoples of Anatolia* (1967); Thames and Hudson

Anatolian iconography

the use of the cylinder seal, which they decorated with designs of their own. The elaborate repertory of figurative symbolism used for this purpose has been one of the most revealing aspects of the finds at Kültepe, because it emphasizes the existence of an authentic and indigenous Anatolian culture persisting through the vicissitudes of migration and political change. A fully developed Anatolian iconography persisted into later centuries, reappearing almost unchanged in the art of the Hittites.

THE HITTITES

The Hittite occupation of Anatolia. The first suggestion of the Hittites' presence in central Anatolia during the Middle Bronze Age is the occurrence in the Kültepe tablets of Indo-European personal names among pre-Indo-European texts. Pre-Indo-European languages then current in Anatolia included Akkadian (Babylonian), the language of diplomacy in the Near East, and the local dialects of the indigenous central Anatolian language now known as Hattic (Khattian, Khattic, or Hattian). The Indo-European name Kaneshite for the official language, now called Hittite, is a term literally meaning "the language of the town of Nesha" (= Kanesh, Kültepe). Besides Kaneshite, two other Indo-European dialects were found in Anatolia: Luwian (Luvian), spoken by immigrants into southwest Anatolia late in the Early Bronze Age and written with the pictographs commonly called Hittite hieroglyphs; and the more obscure Palaic, spoken in the northern district known in classical times as Paphlagonia.

Anatolian languages

The first knowledge of the Hittites, then, depends upon the appearance of typically Kaneshite names among the predominant Assyrian language of the texts. The problem of the origin of the Hittites has been the subject of some controversy and has not yet been positively solved. On linguistic grounds, some scholars were at first disposed to bring them from the west of the Black Sea, but it has been shown that this theory conflicts with much archaeological evidence. One authority argues for their arrival in Anatolia from the northeast, basing his theory on the burning or desertion during the 20th century BC of a line of settlements representing the approaches to Cappadocia from that direction. The evidence from the Halys cities and Cappadocia, however, does not support this picture of an invading army, destroying settlements in its path and evicting their inhabitants. The impression is rather one of peaceful penetration, leading by degrees to a monopoly of political power. From their first appearance among the indigenous Anatolians, the Hittites seem to have mingled freely, while the more flexible Neshian language gradually replaced Hattic.

More tangible evidence of the political events leading up to the foundation of the Hittite Kingdom is provided by the Hittites' own historical tradition recorded by them in later times. Only a few of the tablets of the Hittite archives found at Boğazköy can be dated earlier than the 17th century BC; nevertheless, certain historical texts of this period have survived in the form of more or less reliable copies made in the 14th or 13th centuries. One of these concerns two half-legendary kings of Kussara (Kushshar) named Pitkhanas and Anittas, whom the Hittites apparently associated with their heredity. The city called Kussara has yet to be identified, but the text gives an impressive list of cities that Pitkhanas had conquered, and among them appears the name of Nesha (Kanesh), which his son, Anittas, subsequently adopted as his capital. Also included in the list is Hattusas (Khattusha), known to be the ancient name of the later Hittite capital at Boğazköy, which Anittas was said to have destroyed. The fact that no direct connection could be inferred between these two kings and the subsequent history of the Hittites has been explained by later archaeological discoveries, which have shown them not only to have been authentic historical persons but native Anatolian (Hattian) rulers, whom the Hittites replaced at the end of the 18th century BC.

Kings of Kussara

The Old Hittite Kingdom. The two main periods of Hittite history are customarily referred to as the Old Kingdom (*c.* 1700–*c.* 1500 BC) and the New Kingdom, or Empire (*c.* 1400–*c.* 1190). The less well-documented interlude of about a hundred years is sometimes referred to as the Middle Kingdom. Among the texts from Boğazköy, preserved or recopied by the imperial archivists, those relating to the Old Kingdom are comparatively few. For many years historians of that period relied for the most part on a single remarkable document: the constitutional decree of Telepinus (Telepinush), one of its last kings. In contrasting the prosperity of the nation under his earliest predecessors with the decadence into which it had fallen at the time of his own accession, Telepinus provides a useful though not always reliable summary of early Hittite history:

Decree of Telepinus

> Formerly Labarnas was Great King; and then his sons, his brothers, his connections by marriage, his blood-relations and his soldiers were united. And the country was small; but wherever he marched to battle, he subdued the countries of

his enemies by night. He destroyed the countries and made them powerless [?] and he made the sea their frontier. And when he returned from battle, his sons went each to every part of the country, to Hupisna, to Tuwanuwa, to Nenassa, to Landa, to Zallara, to Parsuhanda and to Lusna, and governed the country, and the great cities were firmly in his possession [?]. Afterward Hattusilis became King . . . .

One would infer from this that the Hittites regarded their own history as beginning with a king called Labarnas (Labarnash); this inference is confirmed by the use in later times of his name and that of his wife Tawannannas as dynastic titles or throne names of subsequent rulers. A difficulty arises, however, because nothing else is known about this king, nor is it certain that he was the first of his line. The earliest contemporary texts date from the reign of his son Hattusilis (Khattushilish; mentioned by Telepinus), and the most important of them is a bilingual inscription in Hittite and Akkadian found in 1957. In the Akkadian version, his name is given as Labarnas, and it is implied that he is in fact the nephew of Tawannannas. In Hittite he becomes Hattusilis and is given the double title "King of Hattusas" and "Man of Kussara." This circumstance has given rise to the supposition that, whereas the original seat of his dynasty was at Kussara, at some time during his reign he transferred his capital to Hattusas (long ago destroyed by Anittas) and thus adopted the name Hattusilis.

The geographical identity of place-names in Hittite historical texts has always been a subject of controversy, but some of those mentioned in the Telepinus decree are known: Tuwanuwa (classical Tyana, near modern Bor); Landa (Laranda; modern Karaman); Hupisna (classical Cybistra); and Lusna (the Lystra of St. Paul's journeys). These and others are all located in the territory to the south of the Halys called by the Hittites the Lower Land, suggesting the first extension of the Hittite Kingdom from its restricted homeland in the bend of the Halys followed hard upon the establishment of the new capital at Boğazköy. The extent and direction of this expansion may have been unforeseen when the site was chosen. As a mountain stronghold dominating the northeastern corner of the plateau, Boğazköy may at the time have had much to recommend it; but later conquests left it on the periphery of the kingdom, and its security was consequently diminished. This possibility is reflected in the bilingual text, which gives a detailed account of events of six successive years of Hattusilis' reign.

Capital moved to Boğazköy

In the account of the first year's campaign, the obscure place-names give no more than a general impression of a localized operation, perhaps in Cappadocia. In the second year's records, however, the extent of Hittite conquests is more impressive, and there is some justification for Hattusilis' claim to have "made the sea his frontier." In fact, the very first place-name mentioned places Hattusilis beyond the Taurus passes in the plains of northern Syria. Alalkha can hardly be other than Alalakh, near Antioch, the ruins of which were excavated by Sir Leonard Woolley (1937–49). The priority given to this town would suggest an approach to Syria through Cilicia and by the Belen Pass over the Amanus Mountains. Two other cities, Igakalis and Taskhiniya, are unidentified, but Urshu, which Hattusilis destroyed and looted on his return journey, is known to have been on the Euphrates above Carchemish. Curious in this account is the absence of any reference to the important city of Yamhad (modern Aleppo), of which the vassal state, Alalakh, may temporarily have been independent. For the rest of Hattusilis' reign, Aleppo apparently remained the principal power in North Syria, to whose armies and allies his own troops were to find themselves repeatedly opposed.

The third year's record introduces the names of two states later to play an important role in Hittite history. The first of these was Arzawa, a powerful kingdom with extensive territory in the southwest part of the peninsula, against which Hattusilis now organized a campaign. In doing so, he left unprotected his possessions in the south and southeast, and they were promptly annexed by the Hurrians, a people who now enter Anatolian history for the first time. During the preceding century, when the power of Assyria was temporarily in eclipse, the Hurrians had infiltrated into northern Mesopotamia and Syria from the east and soon constituted an important element in the population of both territories. On this occasion, having abandoned his attack on Arzawa, Hattusilis seems to have pressed them back and recovered his losses, but he spent the next two years in re-establishing his frontiers. In the sixth and last year of his recorded activities, he found himself once more opposed to the Hurrian armies in North Syria, this time supported by troops from Aleppo. His feud with Aleppo was never decided in his lifetime, for it is known from other sources that he returned, badly wounded, to his old residence at Kussara, anxious to appoint a successor who might continue the struggle. In this endeavour he was at first singularly unsuccessful, for three of his sons in succession proved unreliable to the point of treason; one of the most remarkable and humanly revealing documents of the period is a long and bitter lament in which he chided them for their infidelity and ingratitude. This text is, incidentally, one of the first examples of the Hittite language written in cuneiform, and it is thought that Babylonian scribes had been imported into the capital for the purpose of devising a formula by which this could be done.

The Hurrians

Hattusilis eventually adopted his grandson Mursilis (Murshilish) as his successor, and he proved a wise choice. His first concern was to "avenge his father's blood" by settling accounts with Aleppo, which he destroyed after conclusively defeating the Hurrian armies. Following this victory, he launched an extraordinary expedition against Babylon and, according to Telepinus, destroyed the city. The fact that his army was able to advance almost 500 miles down the Euphrates and overcome the defenses of the Mesopotamian capital is one that historians have found it difficult to explain. His occupation of the city seems to have been extremely brief, because it was not the Hittites but the Kassites who afterward assumed control of the country and founded a dynasty in Babylonia. The Kassites had penetrated northern Mesopotamia from the east, on the heels of the Hurrians. It is by no means improbable that Mursilis had welcomed them as allies, and at least one authority suspects that the attack on Babylon may have been made possible by their support. Because it must have taken place just before or just after the death of Samsuditana, the last king of the 1st dynasty of Babylon, the event can be dated to 1595 BC. This date also may well have corresponded to the death of Mursilis, for, having returned to his own capital laden with booty, a conspiracy among his relatives resulted in his assassination. The succession of his brother-in-law Hantilis (Khantilish) marked the beginning of the catastrophic period referred to in the edict of Telepinus, during which the Hittite Kingdom came near the verge of extinction.

Campaign against Babylon

A major disaster during this period, which eclipsed other military failures, was the conquest of Cilicia by the Hurrians. This great coastal plain to the south of the Taurus Mountains, hitherto described as the "land of Adaniya" (Adana), then acquired the name Kizzuwadna and became the seat of a Hurrian (Indo-Aryan) dynasty. The cities of North Syria were thus rendered inaccessible to the Hittite armies, except through the Antitaurus passes, and remained so until imperial times. When Telepinus himself approached the task of establishing defensible frontiers, he found it necessary to conclude a treaty with a king of Kizzuwadna named Isputakhsus, and he was also compelled to renounce his claims on the neighbouring country of Arzawa.

Of equal interest in the decree of Telepinus is his program of political reforms. Citing examples of the political evils that had resulted in the past from aristocratic disunity at the death of a monarch, he laid down a precise law of succession, specifying an exact order of precedence to be observed in the selection of a new ruler. He further prescribed that in future

> the nobles must again stand united in loyalty to the throne, and if they are dissatisfied with the conduct of the king or of one of his sons, they must have recourse to legal means of

redress and refrain from taking the law into their own hands by murder. The supreme court for punishment of wrongdoers must be the "pankus" or whole body of citizens.

Hittite general assembly

The meaning of the word *pankus* (*pankush*) has been much discussed, for it has been taken to mean a general assembly in the democratic sense, composed of the fighting men and servants of the king. Because it is known to have been an essentially Indo-European concept and did not survive into imperial times, it has been thought to show that at this period the Indo-European aristocracy had not yet merged, as it did later, with the native Hattian population. There is in fact little other evidence to support this suggestion, and in the inscriptions no specific term or epithet is ever used to distinguish the non-Hittite indigenous population.

The Middle Kingdom. Telepinus is ordinarily regarded as the last king of the Old Kingdom. His death marks the beginning of a more obscure period that lasted until the creation of the Hittite Empire. The Syrian provinces, which Telepinus had been compelled to abandon, fell for the time being into the hands of the Hanigalbat, one of the political units into which the Hurrians had become organized. The Hanigalbat, in turn, surrendered them to Egypt, after the successful eighth campaign of Thutmose III. This war also seems to be the first occasion on which the Hittites found themselves in alliance with Egypt, for it afforded an opportunity for them to attack Aleppo, which they once more managed to capture and destroy. The Hittite indebtedness to Egypt for their help may be inferred from an agreement between the two states, in 1471 BC, by which a Hittite king—presumably Zidantas or Huzziyas—paid tribute to the pharaoh in return for certain frontier adjustments. From 1471 to 1450 BC Syria was more or less dominated by Thutmose III, but this situation did not long survive his death. During these last years of his reign, the national unity of the Hurrians seems to have been revived by the imposition of an alien aristocracy and the foundation of a new Aryan dynasty. The Hittites now found themselves confronted on their southern boundaries by a powerful state, known from this time onward as Mitanni. Early in the reign of Amenhotep III, the Mitannians were able to recover Syria and establish their authority over Kizzuwadna. The profound penetration of Hittite civilization by Hurrian ideas, which became pronounced in later times, can be said to have been initiated at this period. For the moment, the situation was politically disastrous for the Hittite Kingdom, for a firm alliance was concluded between Mitanni and Egypt. This was sealed by a royal marriage between a daughter of the Mitannian king, Artatama, and the young Egyptian king, Thutmose IV. By the time of Tudhaliyas III, the formation of this alliance had encouraged a general revolt in all the Hittite provinces.

Mitannians

The Hittite Empire to c. 1180 BC. The Hittite king Suppiluliumas (Shuppiluliumash, Subbiluliuma) dominates the history of the Near East for the greater part of four decades. He was the son of Tudhaliyas (Tudkhaliash) III, in whose company he had gained military experience before ascending the throne. He spent the first few years of his reign in consolidating the Hittite homeland and in improving the defenses of his capital, Hattusas. It may have been at this time that the greatly extended circuit of city walls was built, enclosing an area of more than 300 acres (120 hectares). He then applied himself to the task of settling accounts with Mitanni, the principal enemy of his immediate predecessors. After an abortive attempt to approach Syria by the conventional route through the Taurus passes and Kizzuwadna, Suppiluliumas next attempted a more carefully prepared attack from the rear by way of Malatya and the Euphrates Valley. Coming from this direction, he met little resistance and was able to enter and sack the Mitannian capital, Wassukkani (Washshuganni; not yet located). West of the Euphrates, most of the North Syrian cities hastened to offer their submission. The King of Kadesh put up some resistance but was defeated; and the Hittite armies penetrated southward almost to Damascus *c.* 1370 BC. The wholesale subjugation of the Mitannian kingdom seems on this occasion to have been a matter of indifference to its Egyptian allies, who, under the apostate pharaoh Akhenaton, had temporarily lost interest in imperial defense. Texts have survived of treaties made after this brilliant expedition that show, for instance, that Nuhassi (central Syria) and Amurru (including most of what is now Lebanon) and such cities as Aleppo and Alalakh then became part of the Hittite dominions. It is not easy to understand why Carchemish, which controlled the Euphrates crossings, was allowed to retain its independence and Wassukkani, somewhere to the east on the headwaters of the Khabur River, to remain untenanted.

Conquest of Syria

Suppiluliumas then returned to his capital, leaving his son Telepinus, "the Priest," to arrange the defense of the Syrian provinces. His task may have been complicated by a new situation that had arisen in the remnants of the Mitannian state. The Mitannian king, Tushratta, was assassinated, and his successor, King Artatama, unwilling to place any further reliance on Egypt, turned to Assyria for an alliance against the Hittites. Meanwhile, in 1340 BC, Suppiluliumas returned to complete his conquest of Syria, capturing Carchemish after an eight-day siege. Telepinus now became king of Aleppo and his brother, Piyasilis (Piyashilish), king of Carchemish. It remained only for Suppiluliumas to obtain control over the old Mitannian capital at Wassukkani; this he did, installing a son of the murdered Tushratta as vassal ruler of a buffer state between himself and the Assyrians.

During this last campaign an incident occurred that illustrates the elevated status then accorded the Hittite King as a result of his conquests. While he was encamped before Carchemish, a messenger arrived from Ankhesenamen (Ankhesenpaaten), the queen of Egypt, with a proposal that he should send one of his sons to become her husband. Suppiluliumas agreed to her request and sent her one of his sons, but, as it proved, he arrived too late; by the time he reached Egypt the Queen had been compelled to marry the ambitious courtier-priest Ay, who thus legitimized his usurpation of the throne. The son was murdered—an outrage that was never avenged, for shortly afterward Suppiluliumas himself died of a pestilence. His eldest son and successor, Arnuwandas (Arnuwandash) III, also died, and the throne descended to the young and inexperienced Mursilis II.

The first Hittite misfortune after the accession of Mursilis II was the loss of the small vassal kingdom based on Wassukkani, the last remnant of the once-powerful Mitannian state. It was invaded and occupied by the Assyrians under Ashur-uballit I, who thus was able to establish a frontier with Syria on the Euphrates. Carchemish and Aleppo, however, remained loyal to the Hittites, enabling Mursilis to face a new threat from the direction of his possessions in southwest Anatolia. Arzawa, with its satellites Mira, Kuwaliya, Hapalla, and the "Land of the River Seha," rose in revolt. A detailed account survives of the two-year campaign in which young Mursilis suppressed this insurrection, killing the Arzawan King and installing Hittite governors as rulers of the several kingdoms. Meanwhile, a threat from the north proved more difficult. The Kaska (Kashku) folk, who inhabited the remote mountain valleys between the Hittite homeland and the Black Sea, seem to have been continually in revolt. Their tribal organization and guerrilla tactics prevented the Hittites from conclusive conquest of the country despite Hittite yearly campaigns. Unrest in Kaska country seems also to have affected the rather nebulous state of Azzi-Hayasa, a client kingdom farther to the east on the upper Lycus River. Suppiluliumas had suffered a good deal of trouble from these people early in his reign, and, in the seventh year of Mursilis' reign they again revolted. The King, who was attending to his religious duties at Kummanni (Comana), entrusted their pacification to one of his generals. While the King was at Kummanni, he was joined by his brother, Shar-Kushukh, king of Carchemish, who was taken ill and died; his death sparked off a revolt in Syria. Fortunately, this was for the moment unsupported by Egypt or Assyria, and the appearance of the King himself at the head of his imperial army proved sufficient to suppress it. Mursilis reigned for

Anatolian revolts

28 years (1334–1306) and bequeathed to his successor, Muwatallis, a substantial empire, securely surrounded by dependent states.

Revival of Egypt

Early in the reign of Muwatallis (Muwatallish), it became clear that Egypt, under its 19th-dynasty kings, was beginning to recover its imperialist ambitions. Seti I led his army into Canaan to restore the system of colonial administration relinquished in the time of Akhenaton and advanced as far as Kadesh on the Orontes. A confrontation between the two powers was avoided until the end of his reign. On the accession of Ramses II in 1304 BC, however, a clash between them became imminent, and Muwatallis enlisted the support of his allies. (The Hittite records at this time are fragmentary, but Egyptian scribes mention for the first time the Dardanians, familiar from Homer's *Iliad*, and the Philistines.) The Hittite and Egyptian armies met at Kadesh, and the battle that followed is one of the first in history of which tactical description has survived. The Hittite specialist O.R. Gurney summarizes the Egyptian text as follows:

> The Hittite army based on Kadesh succeeded in completely concealing its position from the Egyptian scouts; and as the unsuspecting Egyptians advanced in marching order towards the city and started to pitch their camp, a strong detachment of Hittite chariotry passed round unnoticed behind the city, crossed the river Orontes, and fell upon the centre of the Egyptian column with shattering force. The Egyptian army would have been annihilated, had not a detached Egyptian regiment arrived most opportunely from another direction and caught the Hittites unawares as they were pillaging the camp. This lucky chance enabled the Egyptian king to save the remainder of his forces and to represent the battle as a great victory. (From O.R. Gurney, *The Hittites*, Penguin Books, 1952.)

Muwatallis subsequently advanced as far south as Damascus, and the Hittites maintained their ascendancy in Syria. The King then found it necessary to transfer his residence to Dattassa, a city somewhere in the Taurus area, and he assigned the government of his northern provinces to his brother Hattusilis. When Muwatallis died and was succeeded by his son, Urhi-Teshub (Urkhi-Teshub), the boy's uncle became a rival to the throne and, after a seven-year quarrel, was able to displace him forcibly and exiled him to Syria.

The accession of Hattusilis III in 1275 BC inaugurated a period of comparative peace and prosperity. Relations steadily improved between the Hittites and Egypt, perhaps as a result of their mutual interest in protecting themselves against Assyria. In 1269 Hattusilis negotiated a famous treaty with Ramses II, assuring the peace and security of the Levant state. Some years later, a further bond was created by the marriage of his daughter to the pharaoh. This girl's mother was Puduhepa (Pudu-Kheba), the daughter of a Kizzuwadnian priest whom Hattusilis had married 29 years previously. Puduhepa was evidently a woman of strong character who ably supported her husband. Together they reoccupied and rebuilt the old capital city at Hattusas, ordered the recopying of the national archives, and instituted a variety of constitutional reforms. Among the many surviving texts from this reign, one appears to be the King's personal apologia justifying his seizure of the throne and his displacement of Urhi-Teshub, the legitimate heir.

Puduhepa

After the death of Hattusilis, his son, Tudhaliyas (Tudkhaliash) IV, extended his father's system of reforms to include the structure and institutions of the Hittite state religion. In this he was much influenced by his mother, Puduhepa, who, surviving her husband, became coregent with Tudhaliyas. His reign was otherwise comparatively uneventful. In foreign relations the records do introduce two names of states hitherto barely mentioned: Assuwa, later corresponding to the Roman Province of Asia; and Ahhiyawa, a large and formidable country, the identity and geographical location of which have been the subject of a prolonged controversy. Some scholars identify the Ahhiyawans with the Achaeans of Homer, or at least with some subdivision of the Mycenaean world, while others place the country on the Anatolian mainland to the north of Assuwa, making the Ahhiyawans ancestors of the Trojans. There is no doubt, however, that a king of Ahhiyawa at this time interfered in the western Hittite dependencies and that Arnuwandas III, who succeeded Tudhaliyas, consequently found it necessary to reconquer Arzawa.

Phrygian invasion

These final episodes in the saga of Hittite history are difficult to reconstruct; the Phrygian invasion of Asia Minor must already have started, and throughout the Near East a mass movement of peoples had begun that was destined not only to destroy the Hittite Empire but to sweep the Hittites out of their homeland on the Anatolian plateau and into Syria. (S.H.F.L.)

II. Anatolia from c. 1180 BC until the Roman conquest

FROM THE FALL OF THE HITTITE EMPIRE UNTIL THE ACHAEMENIAN PERIOD

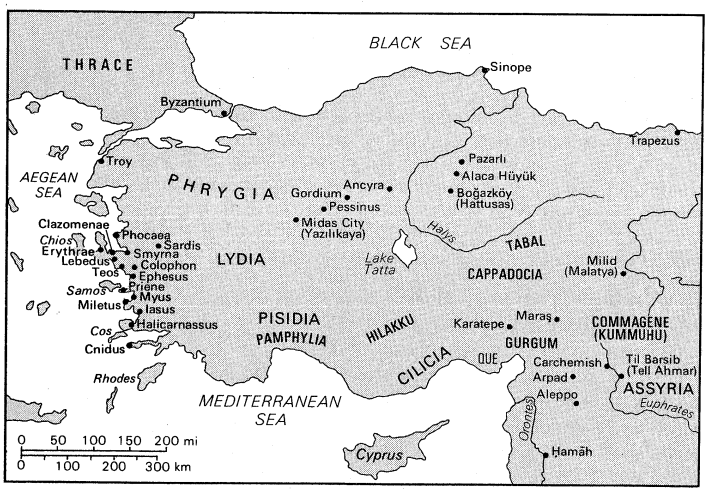
With the end of the Hittite Empire, Anatolia and the whole of the ancient Near East were severely shaken. Migratory groups of the Sea Peoples moving along the south coast of Anatolia and the seashore of Syria and Palestine caused great havoc and upheaval. The Sea Peoples followed the ancient trade route between the Greek Mycenaean world and the coastal cities of Syria, the commercial centres of the Near East. The geographical characteristics of Anatolia facilitate the west–east connection, while the mountain ranges along the northern Black Sea coast and the southern Mediterranean hamper the traffic between north and south.

Anatolia has rightly been characterized as a bridge connecting the Greek world in the West with the great empires of the East. When in the course of a migration new groups passed over this bridge, part of them frequently remained and settled, as had been the case when the Hittites had entered Anatolia. The same happened to the Phrygians, either in connection with or after the fall of the Hittite Empire. The newcomers of the 12th century readily adapted themselves to an existing cultural pattern, and the geography of the country gave rise to the growth of a great number of small local powers and petty chieftains.

Written records are few for the 12th through the 10th centuries, and the picture is not always clear; but the political divisions that emerged in Anatolia after the breakup of the Hittite Empire were roughly the following: A number of Greek city-states were established on the western (Aegean) coast, among them Miletus, Priene, and Ephesus. The southern part of this area became known as Ionia, the northern part as Aeolis. The early history of these cities is known mainly from archaeological finds and from scattered remarks in the writings of later Greek historians. Most of western and central Anatolia was occupied by the Phrygians. In the northeast were the Kaska, a people who had probably participated in the dismemberment of the Hittite Empire. In the southeast were the Luwians, who were related culturally and ethnically to the Hittites. They were organized in a number of small neo-Hittite states (including Carchemish, Malatya, Tabal, and Que) that extended into northern Syria. For the eastern region, archaeological evidence is supplemented by Assyrian texts and by neo-Hittite hieroglyphic inscriptions.

Political divisions after the fall of the Hittite Empire

Phrygia from c. 1180 to 700 BC. The early Phrygians were probably not organized in one strong and centrally governed kingdom. Their origins and the affiliations of their language are still enshrouded in mystery. Greek tradition—and this is still in many cases the best source available—usually dates their migration into Anatolia from Europe around the period of the Trojan War (the early 12th century BC), and the Greeks were convinced that the Phrygians came from Macedonia and Thrace. It is possible, therefore, that the Phrygian language was related to Thracian or Illyrian. From the middle of the 8th century BC, the Phrygians are undoubtedly to be identified with the people called Mushki by the Assyrians, though it seems possible that the Assyrians had earlier used this name as a collective label for their northeastern neighbours. The area occupied by the Phrygians in this early period (12th and 11th centuries BC) is uncertain; many authorities believe they were confined to



Anatolia and northern Syria, *c.* 1180 BC to the 6th century BC.
Adapted from *Westermann Grosser Atlas zur Weltgeschichte;* Georg Westermann Verlag, Braunschweig

the area west of the Halys (Kızıl) River. Parts of the former Hittite capital, Boğazköy, were reoccupied well before 1000 BC. The new settlement was an open, unfortified collection of small, often one-room, houses. The occupants were apparently dissociated from and unaware of the great Hittite past, but it is not certain that they were Phrygians.

Areas of Phrygian settlement

By the 9th and 8th centuries, the Phrygians had formed a centrally organized kingdom in the west with its centres at Gordium and Midas City. Their three main areas of settlement were the hilly country between modern Eskişehir and Afyonkarahisar, the central regions around their capital, Gordium, and the region around Ancyra (modern Ankara), where Phrygian tombs and architectural remains of the 8th to the 6th centuries have been found. Such farther east settlements as Alaca Hüyük, Boğazköy-Hattusas, and Pazarlı temporarily belonged to the Phrygian sphere of influence. Alişar and Çalapverdi were in a kind of no-man's-land between the Phrygians and their Luwian neighbours to the east. The Kaska had by this time probably penetrated into the region between the Halys and the upper Euphrates River. At the time of its zenith in the 8th and 9th centuries, the Phrygian kingdom comprised so large a part of Anatolia that geographically it could well be characterized as the political heir to the Hittite Empire. The invasion of the Cimmerians from beyond the Caucasus at the beginning of the 7th century BC prevented the full realization of this possibility. The American excavations in Gordium itself and in the neighbouring burial tumuli have provided evidence of the great wealth of the Phrygian rulers, reflected in the Greek legends about the Phrygian king Midas. For a later phase of Phrygian culture, the French excavations in Midas City are of prime importance.

Conflicts with Assyria

Phrygia's relations with Assyria are attested to by Assyrian documents. A letter of the late 8th century indicates collaboration between the two powers on an equal basis. Assyro-Phrygian relations, however, were not always friendly; between 715 and 709 BC the Assyrian provincial governor of Que (Cilicia) twice fought the Phrygians before he was finally successful in 709. Before that King Sargon II (721–705 BC) himself had undertaken a campaign against them in 715. Between 718 and 709, a number of East Anatolian and North Syrian Luwian princes sought help from Phrygia in a vain attempt to protect themselves from Assyrian expansionism. The Luwian states, however, were one by one defeated and turned into Assyrian dependencies (see below *The neo-Hittite states* [from *c.* 1180 to 700 BC]). In 709 King Midas himself sent an embassy to Sargon offering submission. During the reign of Sargon's successor, Sennacherib (704–681), the Cimmerians swept through Anatolia, bringing an end to the existence of Phrygia as a major political power. Tradition has it that Midas committed suicide when this happened, and some archaeologists have tried to identify a royal tomb excavated at Gordium as that of the legendary king. Evidences of the Cimmerian destruction of the city are unmistakable. After *c.* 690 the site was abandoned; late in the 7th or early in the 6th century BC it was reinhabited and a new city was built. This late Gordium functioned as the centre of a provincial district, limited probably to the upper valley of the Sangarius River.

Excavations at Gordium have shown that the building and fortification techniques of the Phrygians had reached a high level of perfection; the same is true of their woodcutting, metalwork, and ivory carving. The excellence of Phrygian textiles is known from ancient writings. Caldrons with bullhead attachments found at Gordium show the influence of Urartian craftsmanship, but the differences are enough to indicate an independent local school of bronze working. Other objects reflect the influence of Assyria. Bronze fibulae (clasps), traditionally held to have been a Phrygian invention, have been found in great numbers. According to Greek sources, King Midas was married to a Greek woman from Aeolic Cyme and was the first non-Greek ruler to send offerings to the oracle of Delphi. These traditions indicate that Midas had contacts with Greek coastal cities of western Anatolia. One of the principal ways of identifying an archaeological site is by the type of pottery it contains. Two main types of pottery have been found at sites associated with Phrygia: one polychrome with geometric designs and the other mainly gray or red monochrome. Some archaeologists believe that the polychrome variety, usually called Early Phrygian or Alişar IV in eastern Anatolia where it was first found, is actually Luwian. Whether or not this is true, there were certainly many cultural contacts between the eastern Phrygians and their Luwian neighbours. Geometric patterns typical of Phrygian sculpture appear in Luwian rock reliefs of İvriz and Bor. On the other hand, Luwian influence is clearly present in Phrygian sculptures found at Ankara. There is at least a cultural, if not a political, division in this period between more purely native Phrygia in the west and the eastern Phrygians, with their neo-Hittite affiliations.

Phrygian-Luwian contacts

Around the middle of the 8th century BC the Phrygians adopted an alphabetic script, ultimately derived from the Phoenician alphabet. There is some question as to whether the Phrygians acquired their alphabet in the first instance from a Greek source in the west or south, or whether the Phrygian form of the alphabet was the parent of the Greek. The first supposition seems more likely, since the Greeks probably had more contact with the seagoing Phoenicians than did the inland Phrygians. The oldest Phrygian inscriptions found at Gordium date from the last quarter of the 8th century BC. Another inscription from the same period at Tyana may very well be an authentic document of King Midas, whose name is mentioned on the stone. Toward the end of the 8th century BC, Büyükkale, the citadel of Boğazköy-Hattusas, undoubtedly was a Phrygian settlement. In the early 7th century, perhaps as a result of the Cimmerian invasion, a new system of fortifications was added. Later in that century the settlement extended beyond the citadel to cover most of the area of the former Hittite capital. The Phrygian character of this city is clearly shown by graffiti (wall writing) in Phrygian script and especially by a cult image of Cybele, the main Phrygian goddess, found in a niche at the southeast gate of the citadel. Cybele, a goddess of the "Great Mother of Nature" type, was called Kubaba by the Luwians and, under that name, had played an important role in the religion of Carchemish. Her importance seems to have increased during the Phrygian period, and her statue at Boğazköy seems to show the influence of 6th-century Greek art, as do comparable cult statues from Gordium; from Ayaş, west of Ankara; and from Ankara itself.

The cult of Cybele

The neo-Hittite states (from c. 1180 to 700 BC). The Dark Age after the fall of the Hittite Empire lasted until between 1000 and 900 BC. Carchemish (on the modern border between Turkey and Syria) and Milid (present Malatya, or Arslan Tepe) were the most important Lu-

wian strongholds of this intermediary age, and both were characterized by the same interaction of Luwian and Hurrian influences that had already prevailed during the New Empire period. Reliefs from Malatya, showing the king of that city making offerings to the gods, show a marked similarity to earlier Hittite reliefs at Yazılıkaya and Alaca Hüyük. During the 10th century Aramaean infiltration strengthened and transformed the indigenous Semitic population of Syria; the Aramaeans also penetrated into Luwian areas and sometimes managed to dominate them. Til Barsib (Tell Ahmar) in North Syria was an important Luwian stronghold taken by the Aramaeans in the second half of the 10th century. It became the centre of the Aramaean kingdom Bit-Adini until it was conquered by the Assyrian king Shalmaneser III (858–824). Samal, in the Amanus Mountains (Nur Dağları) of southern Turkey, became Aramaean *c.* 920 BC. Arpad fell shortly after 900 and belonged afterward to the Aramaean state Bit-Agusi. Still later Ḥamāh—the most southern Luwian city—became an important Aramaean power in combination with Aleppo (modern Ḥalab). Aleppo, already a famous capital in the 2nd millennium BC, probably had a numerous Luwian population. The state of Hattina (sometimes called Amqa), which apparently managed to maintain its Luwian character for a very long time, occupied the region at the mouth of the Orontes River, on a site near later Antioch.

The North Syrian states

A great deal more is known about the neo-Hittite states of Syria in the 10th century than about those of inner Anatolia, because so much of the extant source material is Assyrian, and the Assyrian kings had not yet penetrated into Cilicia and Cappadocia.

Til Barsib and Gurgum (capital at modern Maraş) have provided texts from before 900 BC. Most important, however, are the texts from Carchemish, where the subject matter tends to be more diverse than in those of the Hittite imperial age, military exploits being added to the traditional religious subjects. The art of the neo-Hittite states is similar, perhaps under Mesopotamian influence, being more directed toward worldly affairs, frequently depicting hunting scenes and chariot fighting. The principal deities of Carchemish were the Luwian storm god Tarhun (Tarhundas, or Tarhus); Karhuhas, protector of nature's forces; and Kubabas-Cybele, the "queen of Carchemish." Their sacred animals were the bull, the stag, and the lion, respectively. A number of titles used by the kings of Carchemish (*e.g.*, Great King and Hero) are clearly relics of a more glorious Hittite past, but one (*tarwanas*, "judge") is entirely new, possibly reflecting a new political phenomenon. *Tarwanas* has been compared to the Semitic *shofeṭ*, known from the Bible and especially from the Book of Judges, and also to the *tyrannos* of western Anatolia and the Greek world in general. It denotes the judicial capacity of its bearer as distinguished from his military leadership or priestly functions and probably reflects an enlarged concept of kingship resulting from the growing complexity of urban life. Neo-Hittite kings of the 9th century often bore the names of their imperial predecessors; an inscription at Boybeypınarı mentions both a Suppiluliumas and a Hattusilis; at Hattina, kings with the names Labarnas and Suppiluliumas are attested to by Assyrian sources; and during the long reign of a well-documentated dynasty in Gurgum three kings were called Muwatallis.

Assyrian expansion

Tiglath-pileser I of Assyria had invaded Syria *c.* 1200 BC. In the 9th century his successors renewed Assyrian attempts at westward expansion. Ashurnasirpal II (reigned 883–859) received tribute from Carchemish, penetrated into Hattina, reaching the Lebanon and the Mediterranean and returning to Mesopotamia by way of the Amanus Mountains (Samal). Ambassadors from the Luwian regions of Hatti (Carchemish), Hattina, Gurgum, and Milid were among the foreign guests who took part in the celebrations for the inauguration of his new palace in Nimrūd (879 BC). Ashurnasirpal II and his successor, Shalmaneser III (reigned 858–824 BC), both attached great value to the fact that they were able to reach the Mediterranean, but they were unable to permanently subdue the Aramaeans in southern Syria. Included in the Luwian-Aramaean coalition that confronted Shalmaneser III at Qarqār in 853 were forces from the Luwian states of Anatolia, among them Que (the Cilician plain) and Hilakku, the mountainous region to the north of Que. Shalmaneser III made a serious effort to establish Assyrian control over that area; he led four expeditions against Que, one against Tabal (Tubal), and another to Milid (where the tribute of Tabal was brought to him). At this time Tuwatis, the king of Tabal (roughly coinciding with the Hittite Lower Land of the empire period—Lycaonia and Cappadocia south of the Halys), ruled over 20 (according to another text, 24) vassal kings. Apparently, however, the great military efforts of Assyria in this period overtaxed its strength. Toward the end of Shalmaneser's reign a rebellion broke out, and it took more than half a century before the Assyrians were able to renew their western expansion. Hieroglyphic inscriptions from Ḥamāh, the most southerly Luwian stronghold, show how extraordinarily complicated the ethnic situation in this region was. In a text from the middle of the 9th century a king with the Hurrian name Urhilinas—one of the leaders of the coalition against Assyria in 853—records in the Luwian language that he has built a throne and erected a monument for the Semitic goddess Bahalatis. Another contemporary of Shalmaneser III was Halpa-Rudas (Halpa-Rundas) of Hattina. His name has also been found in the hieroglyphic Luwian texts of Tell Tayinat and has helped in the dating of that site. It seems likely that Assyria's contacts with Que, Hilakku, and Tabal, though a threat to their independence, may also have been a strong stimulus to their internal development.

The Karatepe inscription

A great many texts from the various Luwian centres in northern Syria and southeast Anatolia shed light on the history of the 8th century BC. One of the most important of these is a bilingual (Luwian-Hittite and Phoenician) inscription at Karatepe; comparison with the Phoenician version greatly increased scholars' understanding of Hittite hieroglyphics.

The temporary setback in Assyria's westward expansion in the latter part of the 9th century provided a brief period of respite for the neo-Hittite states. This phase ended with the rise of the state of Urartu (centred on Lake Van) in the 8th century. Urartu became a serious threat to Assyria's northern border, and at the same time expanded in a westerly and southwesterly direction, eventually sharing a common border with Phrygia in northern Anatolia and asserting its hegemony over the Luwians. Milid (Malatya) was subdued by the Urartian kings Argishti I (reigned 780–756) and Sarduri II (reigned 755–735); the latter also conquered Kustaspi, king of Kummuhu (Commagene), and forced him to pay tribute around 745. At this time, a king named Asti-Ruwas ruled over Carchemish. A passage from one of his texts seems to indicate that he, too, was dependent on the Urartian king. At Carchemish, the sons of Asti-Ruwas were expelled by a usurper called Araras. In the introduction to one of his texts, Araras emphasizes his diplomatic relations with the state of Mira (on either the south or the west coast of Anatolia), the Mysians (on the northwest coast), the Mushki (Phrygians) and the "Syrians," which may perhaps mean the inhabitants of Cappadocia. In another text he distinguishes between the Luwian, "Syrian," Assyrian, and Aramaean languages. All this points to an active foreign policy in a world that is aware of its fundamental unity as well as of its political and linguistic distinctions. The archaeological evidence —demonstrating the existence of extensive international trade—supports this conclusion.

The fall of Urartu

Under Tiglath-pileser III (reigned 744–727 BC), the Assyrians re-entered the political scene in the west. After Urartu had suffered severe setbacks, first in 743 (in a battle in southern Kummuhu) and afterward in 735, when the Assyrian King penetrated into the heart of Urartu itself, the Luwian and Aramaean kings began to suspect that Urartu was doomed. In 743 Milid, Kummuhu, Arpad, and Gurgum still belonged to the Urartian sphere of influence, but in 740 Tiglath-pileser conquered Arpad, and a large group of princes, among them the

kings of Kummuhu, Que, Carchemish (where a King Pisiris reigned at this time), and Gurgum offered their submission to the Assyrians. King Tutammu of Hattina, who had been strategically safe as long as Arpad had not been conquered, was also defeated and his land turned into an Assyrian province. In 738, Samal, Milid, Kaska, Tabal, and Tuwana (Greek Tyana) came to terms with the Assyrian King. The Assyrian influence had again reached the inner parts of Anatolia. In 732 King Wasu-Srmas (Wasu-Sarrumas) of Tabal was deposed by the Assyrians, and it seems probable that Samal and Que were made into Assyrian provincial territory during the reign of Shalmaneser V (726–722). During the reign of his successor, Sargon II (reigned 721–705), Ḥamāh (720), Carchemish (717), Tabal (713), and Kummuhu (together with Malatya in 708) also ceased to exist as separate states, bringing the era of the independent neo-Hittite states to an end. Shortly afterward the Cimmerians destroyed neighbouring Phrygia.

The Cimmerians, Lydia, and Cilicia, c. 700–547 BC. During the last years of the 8th century BC and the first decades of the 7th century, the Assyrian kings had to fight various wars to maintain their positions in southeast Anatolia. In 705 BC Sargon II himself undertook a campaign in this region, and the unheard-of happened: the Assyrian King was killed in battle. Already in 704 or 703 and later again in 696, Sennacherib (704–681) sent his troops to Que and Hilakku to quell local revolts. On the whole, the Assyrians were not completely successful; while Que (the Cilician plain) remained in their possession, they lost their grip on the more northerly regions of Tabal, Hilakku, and Meliddu.

After the Cimmerians had sacked Gordium, the Phrygian capital (696 or 695), they withdrew to the countryside and confined themselves mostly to a nomadic existence in western Anatolia. So far it has proved impossible to assign any habitation levels or sites in Anatolia to Cimmerian occupation. According to the Greek historian Herodotus, they settled in the area of Sinope on the Black Sea. Herodotus may be right, for this same general area supported the Kaskan nomads of the 2nd millennium BC. Many scholars have concluded from indications in classical authors that a second wave of Cimmerians entered Anatolia from the west and that these western Cimmerians were reinforced by Thracian invaders.

The Lydians

Another new people that appeared in western Anatolia about this time were the Lydians. Their capital and earliest settlement was at Sardis, near present Izmir on the Aegean coast. According to ancient writers, they were the first people to coin money. Their ruling house in the 7th century were the Mermnads, founded by Gyges (reigned *c.* 680–648). The presence of Greek pottery in Sardis' early layers testifies to Lydian contact with the Greeks in this period. The Lydian language belonged to the Anatolian branch of Indo-European and resembled Hittite and Prabaic.

In 679 Esarhaddon of Assyria (reigned 680–669) defeated the Cimmerians under King Teuspa in the region of Hubusna (likely to be a variant of Hupisna-Cybistra), but the area did not thereby become tranquil. In the same year Esarhaddon's troops also fought a war in Hilakku, and a few years later they punished the Anatolian prince of Kundu (Cyinda) and Sissu (Sisium, modern Sis), who had allied himself with Phoenician rebels against Assyrian rule. Again and again the regions to the north of the Cilician plain caused trouble for Assyria. Early in the reign of Ashurbanipal (668–627), however, another Cimmerian invasion threatened the Anatolian states, arousing such alarm that not only Tabal and Hilakku but even Gyges of Lydia sought help from the Assyrians. According to the Assyrian texts, the god Ashur appeared to "Gugu" (Gyges) in a dream, advising him to turn to Ashurbanipal for help. On the same day that Gyges sent his messengers to Ashur, the Cimmerian invaders were indeed repulsed. When Gyges afterward failed to make these temporary relations permanent and instead formed an alliance with the Egyptian king Psamtik, however, Ashurbanipal prayed that "Gyges' body would be thrown down before his enemy," and this indeed happened. During a second attack in 652 Sardis, with the exception of the citadel, was taken by the Cimmerians, and the King was killed. The American excavators of Sardis have found a destruction layer that they connect with this event. Herodotus reports that, like the Phrygian Midas before him, Gyges dedicated offerings to the temple at Delphi, but also that he conducted campaigns against his Greek neighbours at Miletus and Smyrna (Izmir) and that he conquered the Greek city of Colophon. Ardys, his successor on the Lydian throne (651–*c.* 613), again attacked Miletus and took Priene. During his reign Sardis was taken a second time, this time by the Treres, the Thracian tribe mentioned above that operated in close connection with the Cimmerians. According to Assyrian sources, Ardys restored Lydia's diplomatic relations with Assyria.

The defeat of the Cimmerians

The Cimmerian forces were finally beaten by the Assyrians in Cilicia between 637 and 626. At that time the Cimmerian leader was Lygdamis (Tugdamme in the Assyrian sources), who is identified in Greek tradition as the victor over Sardis in 652 and is also said to have attacked Ephesus. The final defeat of Lygdamis is known both from Assyrian sources and from the later Greek geographer Strabo. The Lydian kings Sadyattes (d. *c.* 619) and Alyattes (reigned *c.* 619–*c.* 560) continued their attacks on Greek Miletus. Under Alyattes Lydia reached its commercial and political zenith. He attacked Clazomenae, in 590 took Smyrna, and also subjected many inland regions to Lydian rule. Sappho's poetry, with its references to faraway Sardis, is as eloquent in this respect as the results of modern excavations. The war described by Herodotus between the Lydians and the Medes, expanding out of Iran in the east, probably occurred between 590 and 585. From then on, the Halys River marked the border between the two powers, Lydia on the west and Media (later Persia) on the east.

The growth of an independent Cilicia was one of the most important developments of the last decades of the 7th century BC. It did not include the southern Cilician plain, which came under the control of the Neo-Babylonian Empire after the fall of Assyria in 612. During the conflict between Lydia and the Medes, independent Cilicia and Babylonia acted jointly as mediators, no doubt as two important nonaligned powers of the region. The next and last king of Lydia was Croesus (reigned *c.* 560–546). Famous for his wealth, he ranks with Midas among the Anatolian rulers who made a deep impression on the imagination of the Greeks. Like Midas, Croesus sent offerings to such Greek sanctuaries as Delphi, Miletus, and Ephesus. A number of the relief-decorated pillars of the world-famous Temple of Artemis in Ephesus, one of the Seven Wonders of the World, were presented by him. Stories about his fabulous wealth find some support in the archaeological discovery at Sardis of gold-refining installations from the time of Alyattes and Croesus.

Lydian expansion under Croesus

Croesus completed the work of his predecessors by subduing the Greek cities of Anatolia. He planned to conquer the Greeks of the Aegean islands as well, but the growing threat from the Persians, who by his time had replaced the Medes as the dominant Iranian power, forced him to make an alliance with them instead. According to Herodotus, Croesus ruled all of Anatolia west of the Halys, although the Greek cities probably enjoyed a considerable measure of autonomy. Having secured the support of Egypt, Babylonia, and Sparta (Cilicia remained neutral), Croesus decided to make war against the Persians. Taking the initiative, he crossed the Halys River into Persian territory (547 BC). Both parties fought a battle into the region of Pteria (which may be identical with Boğazköy-Hattusas). Although the battle was indecisive, Croesus decided to return home to his capital, intending to reinforce his troops with allied forces and to renew the war in the following spring. Cyrus the Great, the Persian king, unexpectedly turned after him and took him by surprise. After a short siege, Sardis was taken, and Persian hegemony was established over Anatolia.

Greek colonies on the Anatolian coasts, c. 1180–547 BC. Before the Greek migrations that took place soon after the end of the Bronze Age (*c.* 1200 BC), probably

the only Greek-speaking communities on the west coast of Anatolia were Mycenaean settlements at Iasus and Müskebi on the Halicarnassus peninsula and walled Mycenaean colonies at Miletus and Colophon. The major Greek settlement of Anatolia's west coast belongs to the Dark Age (*c.* 1200–*c.* 1000). In contrast to the colonization of the Mycenaean period, this movement has all the characteristics of a migration. Aeolian territory stretched north of the Hermus (Gediz) River up to Pitane, with Cyme as the most important settlement. According to Herodotus, the Aeolians (coming perhaps from Boeotia and Thessaly or more generally from the eastern part of mainland Greece) once formed a league of 12 cities corresponding to those of Ionia, but their number was reduced to 11 by the Ionian capture of Smyrna. The 12 Ionian cities of historical times were: (the isles of) Chios and Samos, and (from north to south) Phocaea, Clazomenae, Erythrae, Teos, Lebedus, Colophon, Ephesus, Priene, Myus, and Miletus. Among these, Ephesus (as the successor to Apasa, capital of the Luwian Late Bronze state of Arzawa) and Miletus had by far the best claims to historical fame. It is probable that the original number of towns of the 10th and 9th centuries was far larger. There may be a considerable element of truth in the tradition that makes Athens the departure point of the Ionians. Dorian Greeks settled on the Aegean islands of Rhodes and Cos before 900. After this, Cnidus and Halicarnassus were founded on the peninsulas of western Caria. The Dorians formed a league of six, later of five, cities. The Troas (Troy) region was colonized from Mytilene on the island of Lesbos early in the 8th century. By the 8th century, the city league (a group of cities with a common devotion to the same sanctuary) became the normal political institution among the Aeolians, Ionians, and Dorians. A certain measure of urbanization, which in turn was dependent on the development of both local industry and foreign trade, preceded the development of these larger units.

The Ionian cities

Much less is known about the non-Greek populations of the interior. The Mysians, an aboriginal people of the valley of the Caïcus (Bakir) River and the mountains to the north, are mentioned in an 8th-century Carchemish inscription. The Carians, from the hinterland of Miletus and Halicarnassus, enter history as mercenaries in the service of the Egyptian king Psamtik, along with their Ionian neighbours, in the 7th century BC. Of the Lycians, to the east of Caria, nothing definite is known before the 6th century, though archaeological evidence shows that the Greeks had commercial contacts with Lycia as early as *c.* 700. Curiously enough, it was under the aegis of Persian rule that Greek civilization really penetrated into this region. Among the peoples subject to Croesus, Herodotus mentions the Pamphylians, whose country lay in the south, between Lycia and Cilicia. A Neo-Babylonian text of the mid-6th century confirms this, indicating that the Lydian borderline was situated at Sallune (classical Selinus, the most westerly coastal city of Cilicia). There is evidence that the Mycenaean Greeks had contacts with Pamphylia even before 1200, and the name of the Greek hero Mopsus, who figures in the legends surrounding the Trojan War, is associated with the foundation of settlements in both Pamphylia and Cilicia. The appearance of the House of Muksas (Phoenician Mups) in the Karatepe bilingual inscription has suggested that there may be a historical kernel of truth in these traditions, which seem to be a heritage common to both the Greeks and the original Anatolian population. Archaeological finds indicate considerable Greek colonizing activity on the south coast of Anatolia in the 8th century BC and on the north coast in the 7th century. From the middle of the 8th century BC on, Greek merchants were active on the Cilician coast. An important evidence of this can be found in inscriptions of Sargon II (721–705), according to which their activities there began about a generation before his time. Such Greek place-names as Anchiale and Pityoussa occur repeatedly in Assyrian and Neo-Babylonian texts of the 7th and 6th centuries BC relating to the south coast of Anatolia. The North Syrian harbour of al-Mīna' was also of great commercial importance to the Greeks. It is likely that Urartian and other west Asian influences visible in the art of the Greeks and the Italian Etruscans were the result of such commercial contacts. The neo-Hittite state of Hattina on the seacoast of which al-Mīna' was situated probably played an important role in this respect. Of the two main Ionian cities, Ephesus and Miletus, Miletus was very active in colonization. At Ephesus, where the surrounding country produced enough grain and raw materials for the use of the city, there was less pressure for emigration and the founding of colonies. In the early 7th century, the Milesians settled Abydos and Cyzicus in the Propontis (Sea of Marmora) region; Greek tradition suggests that the Lydian king Gyges collaborated in the foundation of Abydos. Dascylium was named after Gyges' father and might be a foundation of the same period. Sinope on the Black Sea coast (founded *c.* 630) was another colony of Miletus. Destruction layers in the excavations of Miletus, Ephesus, and Smyrna, dating from the mid-7th century, suggest that the Ionian cities suffered heavily from the Cimmerian invasion. By the mid-6th century, as has already been mentioned, the cities had all been subjected to Lydian rule.

The Greeks in Cilicia

ANATOLIA IN THE ACHAEMENIAN AND HELLENISTIC PERIODS

Between 546 and 334 BC, Anatolia was dominated politically by the Achaemenian Empire of Persia. Culturally, however, Greek influence continued to be strong and even increased. The coastal regions of Caria and Lycia on the one hand, and Pamphylia and Cilicia on the other, were Hellenized to a considerable degree under the aegis of Persian rule. At the same time, Iranian (Persian) cultural influence penetrated the regions of Armenia, Pontus, Cappadocia, and Commagene. The Persian deities Mithra and Anahita were honoured in Armenia. In Cappadocia, the influence of Iran is clearly visible in the names of the local rulers and in religious practices reported by classical authors. The religious life of eastern Anatolia during this period was characterized by a blend of Iranian, Greek, and indigenous Anatolian elements. The indigenous Luwian groups maintained their linguistic and cultural independence most clearly in such isolated areas as Lycia and western Cilicia, but they are also recognizable in such other southern provinces as Pamphylia, Pisidia, and Lycaonia. The Iranian influence was strong in the northeastern city of Dascylium, originally a Lydian settlement but chosen to be the administrative centre of the satrapy (province) of Hellespontine Phrygia. Aramaic was the official language in the western parts of the Achaemenian Empire, and Aramaic inscriptions in stone and on coins are relatively numerous for Anatolia in the 4th century BC. The Lydian and Phrygian languages continued to be used in their original homelands. The changing nomenclature and boundaries of satrapies and kingdoms throughout this general period present a confusing picture. Under the Persian king Darius the Great (reigned 522–486) Anatolia's original division was as follows: Armina (Armenia), Katpatuka (Cappadocia together with the territories of the Hellespont), Sea Peoples (the south coast together with Cyprus), Sparda (Lydia), and Yauna (the Greek west coast).

The Persian satrapies

The Anatolian Greeks in the Achaemenian period. Of the Anatolian Greek cities, only Miletus had chosen the Persian side in the struggle with Lydia. A number of the others were subjected to Persian rule by force. During the ensuing period, many of these Greek towns maintained a semi-autonomous status, while recognizing Achaemenian overlordship. Outside the cities, occupation forces and military colonies preserved law and order. But in 499, Histiaeus, the Greek ruler of Miletus, led a revolt against Persia. This Ionian revolt was the opening phase of the Greco-Persian Wars. Although the rebels found wide support, not only in the Greek cities of the Propontis region and at the Bosporus but also in Caria, Lycia, and on Cyprus, they lost the decisive sea battle at Lade in 495 BC. In the following year Miletus, the heart of the insurrection, was taken and destroyed. In the last version of the list of satrapies of Darius, Karka (Caria) was added at the end. Apparently it had

been brought under stronger control on account of its support for the Greek cause. During the Greco-Persian Wars, the Anatolian Greeks were compelled to contribute troops and ships to the Persian forces. After the Persian failure to subjugate Greece itself, the Athenians succeeded in pushing back the Persian sphere of influence from many of the coastal districts of Anatolia. A number of cities on the south coast joined the Athenian-dominated Delian League. During the final phases of the Peloponnesian War (431–404), Persia gave support to the Spartan cause. Tissaphernes, satrap in Sardis, concluded a treaty with Sparta in the name of the Persian King (411): the Persians promised the Spartans both financial and naval support; all the Greek cities of Anatolia were to return to the Persian sphere. In spite of the fact that the growing internal weakness of their empire, and perhaps diplomatic reasons as well, restrained the Persians from exercising their full rights under this treaty, Sparta repudiated it and championed the Ionian cause in the early 4th century. After Spartan successes on the continent this war was lost at sea in the Battle of Cnidus (394). Later in the 4th century, however, Persian rule in Anatolia was severely shaken by an insurrection of the Persian satraps of the west (362–359), which subsequently resulted in a considerable measure of local autonomy for the area.

Caria, Lycia, and Cilicia in the Achaemenian period. In the 5th century Caria was ruled by tyrants and princes, some of whom chose the Greek and some of whom chose the Persian side at the time of the Ionian insurrection. Between the middle of the century and the end of the Peloponnesian War (404), Caria belonged to the Delian League. It seems to have been constituted as a separate Persian satrapy. The Carian satrap Mausolus (ruled 377–353) took part in the great insurrection of the western satraps but later changed sides and conquered Phaselis and western Lycia for the Persian King. He likewise profited from the defection of a number of Athens' allies (357), enlarging simultaneously both his own satrapy and the Persian sphere of influence: Rhodes and Cos were added to his possessions. Mausolus made Halicarnassus the metropolis of Caria. The architecture of the city, including the Satrap's tomb, the Mausoleum (another of the Seven Wonders of the World), showed strong Greek influence. The Mausoleum was planned by Mausolus himself but was actually built by his wife and successor, Artemisia (ruled 353–351). Later satraps were Artemisia's brother Idrieus (ruled 351–344), his wife and successor, Ada (ruled 344–341), and Pixodarus, brother of Ada (ruled 341–334).

Mausolus of Halicarnassus

Lycia was conquered by the Persian commander Harpagus after a stubborn resistance by the people of Xanthus. The Lycians had to make a contribution to the expeditionary force led by the Persian king Xerxes in his invasion of Greece (480), but they later sided with Athens. In the latter part of the 5th century, Lycia remained nominally under Persian rule but in practice was almost independent. Greek loanwords in Lycian, the influence of Greece on Lycian art, and the name Pericles, borne by a Lycian king of the 4th century, all attest to a cultural dependence on Greece. The rapid progress of Hellenization in the 4th century is illustrated by bilingual (Greek and Lycian) texts dating from that period. Around 400 BC the Persian grip on the country seems to have been strengthened. Persian rulers are named in inscriptions and on coins (*e.g.*, Artembares, governor of western Lycia). There is evidence that this same Artembares took part in the satrap rebellion. The above-mentioned King Pericles ruled over eastern Lycia between *c.* 380 and 362. Toward the end of his reign Pericles was at war with Mausolus of Caria, who, in all probability, was given western Lycia as a reward for his betrayal of the cause of the satraps. It is uncertain whether any part of Lycia regained its independence before the time of Alexander the Great (334).

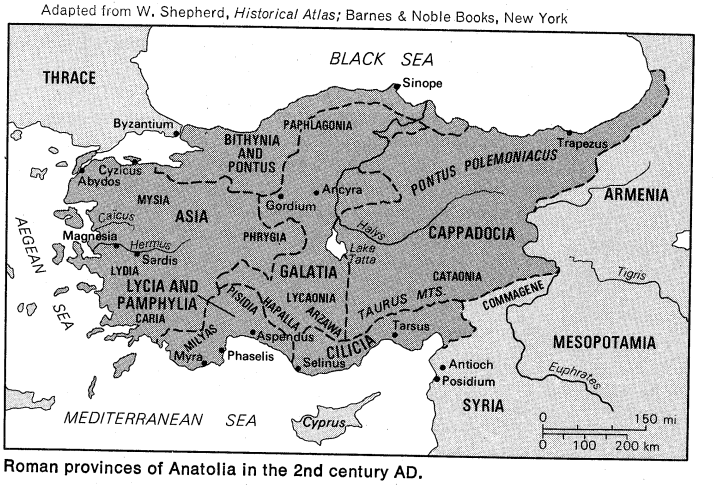
Pamphylia under the Persians

During the 5th century, Pamphylia belonged to the satrapy of the Sea Peoples (and its successors), but its cities were allowed to issue their own coinage. After the Greek victory over the Persians at the Battle of the Eurymedon (fought in Pamphylia *c.* 469), Aspendus and one or two other cities of the south coast were incorporated for a time into the Delian League. In 449, by the terms of the peace concluding the Greco-Persian Wars, the Persians recovered control of Pamphylia; but they seem to have respected its autonomy. Inscriptions from the Pamphylian city of Side (modern Selimiye, Turkey) in a local Sidetan script (and language) together with the legends on Sidetan coins prove the existence in this city of a strong indigenous population group between the 5th and 3rd centuries BC. The history of Cilicia under the Persians can be divided into two distinct periods: the period from 547 to 401, when it was a kingdom recognizing Persian overlordship, and the period between 401 and 334, when it was under the rule of a Persian satrap. During the first period the land was governed by an indigenous dynasty of kings, all of whom bore the name Syennesis. In the second period, the Persians probably actually controlled no more than a narrow coastal strip of western Cilicia, where there was a numerous Greek population.

Anatolia in the Hellenistic Age (334–c. 30 BC). In 334–333 BC the Macedonians under Alexander the Great conquered Anatolia from the Persians and proceeded to destroy the Achaemenian Empire. Alexander's empire was short-lived; quarrels among his successors brought about its fragmentation before 300, and by 275 three dynasties, descended from three of his commanders, had been established in various parts of the territory he conquered: the Seleucids were based in Syria, the Ptolemies in Egypt, and the Antigonids in Macedonia. Anatolia itself was divided: Lycia and Caria were governed by Ptolemaic Egypt, the Seleucids governed most of the other parts of the peninsula. Pamphylia changed hands frequently, but Cilicia, Hellespontine Phrygia, Phrygia, Lydia, southern Cappadocia, and Cataonia were Seleucid satrapies. In the early 3rd century, the states of northern Anatolia (led by Heraclea, Byzantium, Pontus, and Bithynia) formed a league against the Seleucid king Antiochus I. In 278 three Celtic tribes that had migrated across Europe to the Dardanelles were taken as allies by Nicomedes I of Bithynia. The Celts invaded and ravaged Anatolia, until they were defeated by Antiochus in 275. Thereafter they were settled in northern Phrygia by Nicomedes and Mithradates, where they served as a buffer against the Seleucids. The district they occupied was thereafter called Galatia (from Galli, the Latin word for Celts).

The Celtic invasion

Adapted from W. Shepherd, *Historical Atlas;* Barnes & Noble Books, New York



Roman provinces of Anatolia in the 2nd century AD.

In the middle of the 3rd century, Cappadocia became an independent kingdom, and the rulers of Pergamum on the Aegean coast began to enlarge their territory. The Cappadocian leader Ariaramnes carved out a kingdom for himself by incorporating into his own possessions the territory of other local dynasts. Pergamum, originally a mountain fortress, became in the long run an important continental power through the careful manoeuvring of its rulers, Philetaerus and later his nephew Eumenes

(ruled 263–241). Attalus I (ruled 241–197) took advantage of the growing weakness of the Seleucid kingdom to further expand his influence. He broke the power of the Galatians in two battles before 230, adopted the title of king, and for several years (228–223) ruled over the entire Seleucid territory north of the Taurus Mountains.

Antiochus III (reigned 223–187) temporarily restored Seleucid power in Anatolia. By 220, Attalus I was again restricted to practically the original borders of his kingdom. Disturbed by the renewed expansionism of the Seleucids, Egypt, Rhodes, and Pergamum appealed to Rome for help in 200, claiming that Antiochus had formed a pact with Rome's neighbour, Philip V of Macedonia. In 197 Antiochus conquered the entire coast of Anatolia from Cilicia to the Hellespont, while also attacking Pergamum in the interior. In 196 he crossed the Dardanelles and brought the conflict over to Europe. After some hesitation the Romans intervened against him in the Syrian War (192–189). After two defeats, first at Thermopylae and afterward in Magnesia (not far from Sardis), Antiochus was forced to accept the peace of Apamea, which definitely made Rome the predominant power in the Hellenistic East (188). Rome reorganized the Anatolian states: Lycia and Caria were allotted to Rhodes (when this period of Rhodian domination ended in 167, Lycia became a Roman protectorate); Antiochus III was forced to surrender all Seleucid possessions in Anatolia except the Cilician plain; the principal Anatolian powers were then Rhodes, Pergamum, Bithynia, Pontus, and Cappadocia.

Roman hegemony

A new and final stage of Roman involvement was reached when Attalus III (138–133), the last in the line of the Attalids, bequeathed the kingdom of Pergamum to Rome. All of western Anatolia was then reorganized as the Roman province of Asia. The remainder of the peninsula came under Roman rule in the 1st century BC and remained part of the Roman, and later the Byzantine, Empire, until the 11th century AD, when it was partially occupied by the Seljuq Turks. All of Anatolia was taken over by the Ottoman Turks in the 14th century. (For the later history of Anatolia, see ROME, HISTORY OF; BYZANTINE EMPIRE; SELJUQS; and OTTOMAN EMPIRE AND TURKEY, HISTORY OF THE.) (Ph.H.J.H.t.C.)

BIBLIOGRAPHY

Anatolia to c. 1180 BC: The main body of up-to-date information on this subject with full bibliographies relevant to the various periods is contained in the following fascicles of the new *Cambridge Ancient History:* J. MELLAART, *Anatolia: before c. 4000 B.C. and c. 2300–1750 B.C.* (1964), and *Anatolia: c. 4000–2300 B.C.* (1962); HILDEGARD LEWY, *Anatolia in the Old Assyrian Period* (1965); O.R. GURNEY, *Anatolia: c. 1750–1600 B.C.* (1962), and *Anatolia: c. 1600–1380 B.C.* (1966); A. GOETZE, *Anatolia: From Shuppiluliumash to the Egyptian War of Muwatallish* (1965). The following two sources are earlier and semipopular summaries: O.R. GURNEY, *The Hittites*, rev. ed. (1961), dealing with Hittite history; and SETON LLOYD, *Early Anatolia* (1956), archaeology from the prehistoric to the Phrygian periods. More recent popular summaries, fully illustrated and based on more recent archaeological evidence are: J. MELLAART, *Earliest Civilizations of the Near East* (1965); and SETON LLOYD, *Early Highland Peoples of Anatolia* (1967). A very fully illustrated account of excavations at a prehistoric site of paramount importance to world prehistory is J. MELLAART, *Catal Hüyük: A Neolithic Town in Anatolia* (1967). A recent and temporarily definitive treatise on the subject is JOHN GARSTANG and O.R. GURNEY, *The Geography of the Hittite Empire* (1959). A standard work until the postwar period but now quite outdated is JOHN GARSTANG, *The Hittite Empire* (1929). An article on a vital and controversial aspect of religious iconography relating to the intrusion of Hurrian ideas into Hittite culture is E. LAROCHE, "Le Panthéon de Yazilikaya," *Journal of Cuneiform Studies*, 6:115–123 (1952). An article pioneering some new ideas about the arrival of the Hittites in Anatolia is J. MELLAART, "The End of the Early Bronze Age in Anatolia and the Aegean," *American Journal of Archaeology*, 62:9–33 (1958).

Anatolia from c. 1180 BC to the Roman conquest: (Phrygia from c. 1180 to 700 BC): A brilliant survey of archaeological material on the Cimmerians and Phrygians may be found in K. BITTEL, "Kimmerier, Phryger und Skythen in Kleinasien," ch. 3 of *Kleinasiatische Studien* (1942). The results of the new American excavations at Gordium have appeared in the Preliminary Reports by R.S. YOUNG in the *American Journal of Archaeology* since 1955. M.J. MELLINK edited the acta of an important archaeological conference on the nomadic impact in both Anatolia and Iran during the Iron Age under the title *Dark Ages and Nomads c. 1000 B.C.* (1964). R.D. BARNETT recently dealt with this general period in fasc. 56 of the new *Cambridge Ancient History—Phrygia and the Peoples of Anatolia in the Iron Age* (1967). Although in fact devoted to Hattusas, the Hittite capital, K. BITTEL, *Hattusha, the Capital of the Hittites* (1970), also contains much information on the Phrygian parts of Anatolia, especially in ch. 6. *(The Neo-Hittite states of Anatolia from c. 1180 to 700 BC)*: B. LANDSBERGER, *Sam'al: Studien zur Entdeckung der Ruinenstätte Karatepe* (1948), is a brilliant study of southeast Anatolia during the beginning of the 1st millennium BC, based not only on all of the Assyrian and northwest-Semitic texts that are relevant to this region but also on the results of the excavations. P.H.J. HOUWINK TEN CATE, "Kleinasien zwischen Hethitern und Persern," ch. 2, in E. CASSIN, J. BOTTERO, and J. VERCOUTTER (eds.), *Die altorientalische Reiche*, vol. 3 (1967), is a general account of Anatolian history between c. 1180 and 547 BC. *(The Cimmerians, Lydia, and Cilicia, c. 700–547 BC)*: G. RADET gave an exhaustive collection of all the data on Lydian history found in classical authors in *La Lydie et le monde grec au temps des mermnades, 687–546* (1893). The archaeological results of the new American excavations at Sardis are regularly presented by G.M.A. HANFMANN in the *Bulletin of the American Schools of Oriental Research* since 1959. *(The Greek colonies, Mysia, Lycia, Pamphylia, etc. from c. 1180 to 547 BC)*: The history of these regions in the first half of the 1st millennium BC can only be written by archaeologists on the basis of archaeological evidence. H. METZGER, *Anatolie*, vol. 2 (1968; Eng. trans., 1969), gives an account of these archaeological data with a marked emphasis on the indigenous population. On Greek colonization and commercial adventures in the East, see J.M. COOK, *The Greeks in Ionia and the East* (1962); and J. BOARDMAN, *The Greek Overseas* (1964). *(The Anatolian Greeks in the Achaemenian period)*: A recent treatment is H. BENGTSON, *Griechen und Perser* (1965; Eng. trans., *The Greeks and the Persians,* 1969). *(Caria, Lycia, and Cilicia in the Achaemenian period)*: G. BOCKISCH, "Die Karer und ihre Dynasten," *Klio* 51:117–175 (1969), is based on full use of all the Greek evidence, both from classical authors and from inscriptions, but the study disregards the few available data from Persian sources. O. TREUBER, *Geschichte der Lykier* (1887), is the old, but by no means antiquated, collection of all the evidence on Lycian history before it became possible to incorporate evidence from indigenous sources such as inscriptions and coins. P.H.J. HOUWINK TEN CATE, *The Luwian Population Groups of Lycia and Cilicia Aspera during the Hellenistic Period* (1961), contains an introductory historical chapter, based on both the Greek and the indigenous evidence. *(Anatolia between 334 BC and the Roman period)*: W.W. TARN, *Hellenistic Civilisation* (1947), gives a general survey of the Hellenistic period with full emphasis on historical evidence. A trustworthy study of the intricate geographical problems connected with the Hellenistic history of the Anatolian states is E. MEYER, *Die Grenzen der Hellenistischen Staaten in Kleinasien* (1925). A storehouse of information, also for the preceding periods, is D. MAGIE, *Roman Rule in Asia Minor*, 2 vol. (1950).

(S.H.F.L./Ph.H.J.H.t.C.)

Anatolian Cultures

Anatolia, as defined in this article, is modern Turkey without Thrace, the European portion of the country. It consists largely of a high central plateau bordered on the north, south, and east by ranges of coastal mountains that vary considerably in size and in density. The great diversity in the character of the terrain, the position of the country between two climatic zones—the Mediterranean and the Asian—and the proximity of the sea on three sides combine to produce extremities of climate and vegetation.

The Turks moved into Anatolia in the 11th century and established various sultanates, the most renowned being the Seljuq dynasty of the 11th and 12th centuries. The Ottoman dynasty was founded in northeast Anatolia in the mid-13th century and gradually conquered Anatolia and Thrace, until, in 1453, the Ottomans overthrew the Byzantine Empire and made Constantinople (Istanbul) the capital of an empire that eventually stretched from

the gates of Vienna to North Africa, Baghdad, and the Crimea. Military defeats and internal restructuring slowly weakened the empire until final dismemberment after World War I; Anatolia, the heart of the Ottoman Empire, became the Republic of Turkey in 1923, under its great leader, Mustafa Kemal Atatürk (see SELJUQS; OTTOMAN EMPIRE AND TURKEY, HISTORY OF).

Population of Turkey

The population in 1970, according to Turkish census estimates, was 35,700,000 (including Thrace), with an annual growth rate of about 2.6 percent. Densities are very low in the east, southeast, and south, fairly low on the central plateau, and highest in the west and along the Black Sea coast. The proportion of people living in villages is about 60 percent; moreover, a large proportion of the new urban residents still have village roots, and many of the smaller towns contain large farming populations. The current nomadic population is unknown, but it is reckoned to be only tens of thousands.

Turkey does not have a class of great landlords, though large estates exist in the west, south, and east; but, although workers and villagers are often strikingly dignified, the gap between them and the Westernized city dweller is marked, and the society is strongly ranked.

Aside from the Turkish majority, there are a number of minority ethnic and linguistic groups. The 1965 census gave figures for people speaking a minority language either as a mother tongue or as a second language. It is probable that many persons giving a minority language as their second language would claim identity as a member of that minority. Thus, to derive a rough notion of the size of minorities in Turkey, one could almost add the figures in the two columns in the accompanying Table.

Minority Populations by Mother Tongue and Second Language Best Spoken

	mother tongue	second language
Kurdish	2,219,500	1,752,900
Arabic	365,300	357,100
Zaza	150,100	112,100
Circassian	58,300	55,000
Greek	48,100	82,144
Georgian	34,300	49,000
Armenian	33,100	23,282
Laz	26,000	59,100
Pomak	23,100	37,000

Source: Republic of Turkey, *Census of Population*, 1965.

The same census (of 1965) indicated that less than one percent of the population was non-Muslim (constituting various Christian sects and Jews).

TRADITIONAL CULTURE PATTERNS

Human settlement in Anatolia has been traced back well before 6000 BC. Much archaeology is in progress on the Hittite and other ancient empires, and on Greek, Roman, Byzantine, Seljuq, Crusader, and Ottoman remains. Indeed, Anatolia is one of the richest archaeological areas in the world. Both genetically and culturally the resident peoples undoubtedly have influenced the waves of immigrants throughout history.

General characteristics of the Anatolian peoples

This complexity makes general statements somewhat hazardous, but the national characteristics and culture of the present peoples of Anatolia may be said to have two main strands. First, the Turks—and many of the minority ethnic groups—looking back to a nomadic past, show a pride in manliness and toughness. Even educated Turks are mindful of the people who came out of Central Asia, conquered the Arab and Byzantine empires of the Near and Middle East, and set themselves up as rulers—such great names as Timur (Tamerlane) and Genghis Khan and the Seljuq and Ottoman dynasties. The other strand is Islām; the Asian invaders were Muslims before they reached Anatolia in the 11th century.

At the personal level, Turks are proud, dignified, and quarrelsome; the villagers turn easily to violence, because to fail to do so is shameful. One result is a high rate of homicide—killing for honour in village feuds. At the national level, most men willingly accept conscription into the army and are proud of their military prowess. Turks are also intensely hospitable. Paradoxically, this hospitality is consistent with their pride, for generosity to the stranger is everywhere a highly acceptable means of establishing warm relationships while preserving social distance and defining the guest as an outsider.

The emphasis on manliness and toughness and the tradition of Islām combine in traditional values concerning women and families. Traditionally men and women have lived in separate social worlds; men are unequivocally deemed superior. It is the man's responsibility to see to the honour of his womenfolk, and any sexual misdemeanours by them reflect on him and his sons, unless he can redeem them by violence. Traditionally, moreover, Islām has had political consequences. Government has been seen as something religiously guaranteed; even the secularist reformer Atatürk was known to the devout majority as the *gazi*, the conqueror of the infidels.

Material culture. *Among the villagers.* As their main crop, most Turkish villages grow cereals—mostly wheat and barley but also, in some areas, maize, rye, oats, and rice. In the mountains and coastal areas, however, there are many other crops: potatoes, cotton, tobacco, sugar beets, nuts, citrus fruit, grapes, figs, and tea. Dairy farming hardly exists as a separate activity, because cattle are bred for work as well as for meat and milk. Sheep and goats, found everywhere, are an important part of the rural economy, producing milk as well as meat and wool.

Agricultural and pastoral seasons

Each season has its tasks and activities. Except in the south and west, winter is a period of frost, snow, and social activities. Animals are often kept indoors and fed mainly on chopped straw. With the spring thaw, ploughing and sowing are soon under way. After a month or so of less urgent work, the hay harvest is followed immediately by the main grain harvest, a period of intense activity lasting some six to eight weeks; everyone works, some people 16 to 20 hours a day. The crops are reaped, carted to the threshing floors, threshed, winnowed, and finally carted away to market or to storage.

The annual cycle of animal husbandry is even more variable. Sheep and goats that have been away for the winter return in the spring and may be sent away again to upland pastures—*yayla*—at the height of summer. How many, how far, and for how long depends on the particular village. In mountainous areas, many villagers are settled nomads, still owning upland pastures away from their main village. Cattle, water buffalo, horses, and donkeys are pastured on the village lands when they are not working.

Tools and housing

The tools used for these tasks resemble those found in other parts of the Middle East and are clearly very ancient in form. The light wooden plow, or *saban*, is drawn easily by one pair of oxen or buffalo. It has an iron-tipped share but no moldboard, so that it does not turn a furrow. Sowing traditionally is done by hand, reaping with a sickle or scythe. The crops are carried to the village for storage on four-wheeled horse-drawn carts or on the traditional two-wheeled ox cart, the *kağnı*. Threshing involves driving an ox-drawn sledge about five feet long over the crops, round and round, day after day. Flint teeth on the underside of the sledge break the grain from the ears and chop the straw into chaff. This mixture is winnowed by hand with wooden forks and put into woollen sacks.

The typical Turkish village house is a rectangular flat-roofed building, the colour of the local unbaked brick or stone from which it is made, one or two stories high. The poorest contain a single room to house family, livestock, and possessions; the very prosperous encompass joint households containing up to twenty people or so, living in large compounds off which lead many living rooms, stores, stables, and barns. The vast majority of homes fall between these extremes. The household head and his wife live and sleep in the main room of the household, which usually has a mud floor. This room contains the bedding, stacked away in the daytime, the cooking stove (in some areas a beehive-shaped oven built into the floor),

shelves with cooking pots and implements, a gaily painted chest containing the woman's trousseau and personal possessions, and a chest for flour and grain. People sit on rugs or mats spread on the floor. Many homes also have at least one room fitted as a guest room, normally for the use of men only. It has built-in divans running along the walls and very often a stone or wooden floor. Tables and chairs, once rare, are now becoming more common. This description cannot possibly take account of the very great local variations. The most marked contrast occurs in the forests of the northern mountain ranges, where the village homes are made of timber with red roofs. Brick, cement, and, very recently, cinder block are also now becoming common, wherever people can afford them.

Village dress

In recent years Turkish men have adopted increasingly the styles and sombre colours of European male dress. Fezzes and turbans were abolished by law in 1925, and most peasants now wear cloth caps. The famous Turkish baggy trousers, exceedingly full in the seat, are still quite common, but the traditional cummerbund and colourful shifts or waistcoats are rare. Shepherds, whose work may involve withstanding intense cold, wear a special large cloak made of felted wool and a hood with attached scarf that winds around the head and protects the ears. The village women still largely preserve traditional costume. They wear some locally customary combination of baggy trousers, skirts, and aprons. In many areas it is still possible to identify a woman's town or village and her marital status by her dress; village women in Turkey have never worn the veil but cover their heads and mouths with a large scarf.

Most village areas contain weavers, masons, carpenters, and smiths and tinsmiths. Some villagers go to town for craft services, and a number of craftsmen travel around the villages—particularly specialists, such as sieve makers or sawyers.

Among the townspeople. It is impossible to summarize in a few words the "material culture" of the towns and cities that were lately the central part of a great empire and that are now profoundly influenced by European fashions and technology. Most towns, large and small, nevertheless, still contain markets where simple lockup shops stand side by side in rows. Usually these are arranged by craft or wares—coppersmiths, jewellers, cobblers, tailors, motor mechanics, and so on. Retailers also are grouped by commodity—cloth, shoes, hats, dry groceries, hardware, carpets.

Traditional Anatolian town houses, still the normal thing in many Anatolian towns, were built in stone or wood, usually of two stories, with wooden floors and sometimes beautifully carved ceilings. The upper story often protrudes cantilever fashion into the street. With whitewashed walls and red tiled roofs, these small towns sometimes present a much more civilized appearance than most villages, though just as often the distinction between large villages and small towns is barely visible.

Although most townsmen wear Western dress, many townswomen still wear the *çarşaf*, traditionally consisting of two large sheets, that effectively conceal everything but the eyes; the veil is now exceedingly rare.

Among the nomads. Traditionally, much of Anatolia, more especially the east, was peopled by nomads or transhumants, who migrated seasonally between upland and plain. Most of their descendants, herding sheep and goats, move from place to place in tents, but some possess houses in two or even three villages at different altitudes, moving from one to another at the appropriate seasons.

Central importance of Islām

Religion and beliefs. Islām lay at the centre of the legal, political, and educational structure of the Ottoman Empire, formally defining religious minorities—and, by implication, the unorthodox Muslims—as second-class citizens but ignoring ethnic differences within religious communities. At this imperial level Islām, especially Orthodox Sunnī Islām, was a highly formalistic and demanding religion.

For the orthodox—and even the less orthodox—Islām imposes many duties. Men and women must keep themselves in a state of ritual purity, pray five times a day, fast for a month every year, and strive, if possible, to visit Mecca at least once in a lifetime. Islām provides basic ideas about the nature of morality, sin, and punishment, about relations between men and women, about cleanliness and impurity, and about political submission (see ISLAM and ISLAM, HISTORY OF).

For centuries there has also existed within Islām a tradition of mysticism and holy men and teachers who founded orders explicitly distinct from, even opposed to, orthodox Islām. These orders were often known as dervishes (see ISLAMIC MYSTICISM). Some, for example, the Mawlawīyah (Mevlevî) and Bektāshīyah (Bektaşî) orders, were officially accepted and recognized in the Ottoman Empire; others existed at a popular level. Many of them incorporated ideas of love, passion, and salvation, as well as teachings inconsistent with or opposed to formal orthodoxy. These orders were in varying degrees secret societies, so that members often practiced the outward formalities of orthodoxy and kept their more heterodox beliefs and customs to themselves. The orders were sometimes sects or movements or institutions with urban branches, buildings, and monasteries that people might join; others were rural or tribal communities or parts of such communities.

Traditional beliefs

At the village level, a range of beliefs concerning magic and superhuman powers exists: the evil eye, the ancient belief that certain persons, consciously or unconsciously, can harm or kill with a glance; ritual healing and fortune-telling; and various personal and agricultural rites that are local versions of Islāmic belief and practice. It is, for example, still common to make offerings at the shrines and tombs of local holy men. People, houses, carts, and lorries carry amulets and religious texts to ward off evil, and people suspect each other of various forms of malicious magic, especially of inducing impotence on wedding nights. Pieces of paper with Qur'ānic verses written on them are thought especially potent in the performance of magic.

Social structure. Traditionally, until World War I and after, Anatolian society could be roughly divided into four divisions. At the top were the rich or the rulers, mostly people who lived in or had close connections with the main cities, Istanbul and Izmir, and many of whom were increasingly Westernizers and reformers. On the second level were the local townsmen, people who on the whole stayed put in the towns—landowners, merchants, craftsmen, and also farmers—all of whom lived within and accepted the main political and legal system.

Third were the villagers, who formed the vast majority, about 80 percent, of the total population. Most villages were small, self-contained settlements, relatively isolated and economically self-sufficient and, to a large extent, autonomous with regard to their own internal political and social affairs. Although some villages or some lands around villages were owned by absentee landlords, most Anatolian villagers owned their own land and produced for their own needs. Generally, their own contacts with the central authorities involved paying taxes to the tax collectors or rents to the absentee landlords (who in turn paid the taxes); the government provided only minimal services in return. The villages that produced export products, such as cotton and tobacco, and thus had a more involved economic relation with the cities, were relatively few. In all cases, villagers were primarily concerned with power and prestige in their immediate social environment. The village looked inward.

It is also important to realize that rights to land often depended (and still today depend) more on the acceptance and recognition of kin and neighbours than on the legal rules of the state; moreover, in any case, landholding was not always very secure and depended on force or political influence. Quarrels over land were normal and permanent features of such a society, often involving violence.

The fourth division of Anatolian society consisted of the pastoralists in the mountainous areas, who were organized tribally, sometimes with powerful chiefs, and who held rights to the territories that they pastured.

These tribes on the whole enjoyed even more political autonomy and were even more isolated from urban society than were the villagers. The large tribal organizations found in other areas of the Middle East, however, apparently did not exist in Anatolia.

This fourfold division is arbitrary. A detailed historical look at one area at one point in time would reveal that there are all kinds of complexities and that, within each of the divisions, there are great differences of power and wealth.

Division of labour between men and women

Social roles and kinship. Anatolian men in the villages are responsible for the major agricultural work, for looking after the oxen, and for making all contacts outside the home, both official and economic, including shopping. Women—widows, for example—might do men's work, but men never undertake women's tasks. One consequence of this is that men are more dependent on women than are women on men; and a bereaved widower who has no other adult women in his household may well remarry within a few days or weeks.

Women in Anatolia, as elsewhere, are concerned with the care of children and their houses and with the preparation and cooking of food. They are also responsible for the milking, for caring for the chickens, for making cakes of winter fuel from dung and straw, for weeding vegetable plots near the village, and for reaping barley and other short-stemmed crops.

The structure of social relationships is in innumerable ways profoundly affected by the sharp social segregation of men and women. Although women can and do exercise power and influence, not least by gossip and the control of reputations, they are seen and accepted by both sexes as inferior at every level from religious and metaphysical reality to the details of the daily chores.

This segregation is related to attitudes toward sex and sexuality, which people treat as ritually impure and somewhat shameful. Women are measured by rigid standards of purity; sex is a forbidden topic between close kin; and a young couple is forbidden to show any interest in each other if anyone else, even a member of the household, is present. A man leaving on a trip does not say good-bye to his wife publicly, nor does he greet her publicly on his return.

The typical household

Ideally, a traditional household consisted of a man, his wife, his adult sons and their wives, and his young children and grandchildren. On the death of the household head, this large household broke up into as many first-generation households as there were sons, each beginning the process again. The former high death rate among adult men, the lack of living sons, and, very rarely, quarrels between generations made these large households a minority of all households at any one time. Thus, although probably most villagers live some part of their lives in such a household, most village households at any given time contain only parents and children, with perhaps other random relatives. The average size of a household is probably between five and six persons.

In most areas, household heads are grouped in patrilineal lineages or clans—that is, a group of men descended through males only from a common ancestor, usually a great grandfather, but perhaps an ancestor further back still. Such lineages are concerned primarily with mutual support and defense within the village, and the members often have adjacent houses and lands. In some areas in the past, lineages had wider political importance; and they were more important among tribal peoples.

Marriage

Marriage is permitted between close kin under the incest rules of Islām. A man has a right to his brother's daughter as a bride for his own son; and this right has often been exercised. Men also sometimes marry their sons to more distant women of their lineages or to women related through marriage. Many marriages, however, involve completely unrelated people. Whatever the precise arrangement, for societies of this kind marriage is of central importance. From the day of their own marriage every couple starts to consider and prepare for the marriage of their children. Marriage is thought of less in terms of personal choice and compatibility and more in terms of political and economic relationships.

Traditional village weddings involve elaborate ceremonies and last several days. Large transfers of wealth are often involved. Regional variations are considerable, but commonly a man must still make a marriage payment to the father of his son's bride and must also pay for the wedding, the total cost amounting to as much as or more than one year's total income for an average household, without counting the need to provide a new room or house.

Kinship carries strong obligations of mutual support and interest. People look to their kin for day-to-day sociability, for hospitality in other villages, for help in trouble, for cooperation in weddings and funerals, and for aid nowadays in urban migration, in finding jobs, or in getting official favours. Kinship and marriage ties have had important political and economic implications, both at higher levels of power in the towns and in links between towns and villages.

Roles of youth. Birth is surrounded by ceremonies and visits among the women. Very young children in crowded male-dominated households are not allowed to cry, and motherhood in the first months is more or less a full-time job, day and night. Breast-feeding continues normally to the age of two or even later; weaning is sudden.

Small girls are expected to take charge of younger siblings from a very early age and to cooperate in women's work. After ten or eleven, they are progressively more secluded. Marriage was once thought possible at about twelve or thirteen years of age and normal by about sixteen. No girl without severe handicaps remains single.

Boys have far more freedom in growing up, and from about the age of eight they are expected to pasture animals and take on other farming jobs. By twelve they have learned to plow. Boys are circumcised—often in groups—at any age from about six to twelve. The actual operation is performed by recognized professionals, publicly and ceremonially. The fortitude shown by the boys reflects on the family reputation.

Deference to seniority

Seniority is the basic criterion for relationships of superiority and deference. Children are expected to be obedient and submissive, and on all public occasions sons are to defer to fathers, daughters to mothers and mothers-in-law, and younger siblings to elder siblings. A son, for example, should not smoke or even offer an opinion in his father's presence. The common word for elder brother in Turkish, *ağabey*, is a combination of two words for respected superiors, and the word for elder sister, *abla*, also carries strong overtones of respect. It is true that the villages have no formal age grades and that circumcision, though absolutely necessary for manhood, does not imply any specific admission to a grade. But the traditional deference patterns exclude personal intimacy between fathers and sons and encourage sociability among equals. Thus, informally, men do divide themselves into groups by age.

Marriage also is a "rite of passing," since it changes many existing rights and duties, and for each partner marks the passage from dependent youth to full adult status.

CONTEMPORARY CHANGES

Major trends. Change in Anatolian society is now rapid, complex, and extremely uneven. It is even difficult to know when to use the past tense and when to use the present. These changes, like those in many other developing countries, include rapid growth in population, communication, production, size of cities, and administration and education. The government is helping to decrease regional isolation by greatly improving village roads and trunk routes between cities. Radio services and the national and local press have also undergone vast expansion.

Changes in employment

Up to about 1950 incomes were low and relatively unchanging for most of the population, but since then the economy has grown and prospered so much that the average real per-capita income has risen by about 75 percent. The benefits of this economic growth are, of course, unevenly distributed, but in general it allows greatly increased consumption, especially of factory-made goods.

Another effect is a vast increase in jobs available in towns and cities. This has attracted migrant labour in the form of men who work in urban centres but still keep their families in the village tilling the land. It has also meant that many village households have uprooted themselves and moved to towns and cities, vastly increasing the urban populations. In 1950 there were five towns of over 100,000 persons, numbering together about 1,700,000; by 1970, 20 towns of over 100,000 numbered together 6,700,000. Ankara itself rose from 289,000 in 1950 to 1,200,000 in 1970.

At the same time, through political and administrative pressure and increasing efficiency, secularization and modernization have increasingly pervaded the rural areas and small towns. State schools have increased in numbers in the countryside, introducing more national or cosmopolitan ideas. Nearly all children attended primary schools by 1972. Bureaucracy has introduced registration of births, deaths, and marriages and more complex systems of credit and law. Peasants now use banks for their savings. Land disputes are now often settled by official and legal means rather than by local social pressures. Legal divorce has tended to replace socially recognized separation. The list of changes is endless.

Widening horizons of the villagers

It is possible to think of these complex changes under three heads. First, the range of social relationships is widely extended. Village men need no longer deal with the world outside through their village leaders; they can deal themselves with merchants, salesmen, banks, government offices, employers, and fellow workers in the cities. Women, who twenty years ago hardly left the village except to visit neighbouring villages, now regularly visit their close relatives who are living in cities.

Second, the range of information reaching rural areas is vastly increased. Most children are literate, most homes have radios, most men go to the cities, and most people know a friend or relative who has worked in western Europe—mainly in West Germany. Although much of their information may be partial or misleading, their increased knowledge of the world is enormous; and as their incomes rise, their range of choices with regard to goods, occupations, and even customs correspondingly increases enormously.

Third, many of the goals and standards that guide life also change. Most fundamentally, people are less committed to the village and to success in village eyes, so that conformity to customs and traditions is less consistent. By and large, the village people increasingly are facing outward, toward the city and the nation.

Changes in cultural traits. *Arts and crafts.* The new national elite in Turkey is European oriented. The policies of Atatürk, especially the change from Arabic to Latin script, consciously cut educated Turks off from their immediate Ottoman cultural heritage. At the same time, however, came an emphasis on Turkish culture, especially folk culture. Thus, Turkey has now national writers, composers, musicians, artists, and architects who seek to weave Turkish themes into European cultural forms. With this goes a revived interest in the specifically Turkish heritage—literature, dancing, music, and architecture. Turkey has excellent museums.

Traditional handicrafts, on the other hand, suffer, as everywhere, from the competition of factories and from the growth of explicit commercialism and tourism. Village girls, for example, who used to weave *kilims* and carpets of their own design for their trousseaus, now make carpets for the market, following fixed instructions given to them by local merchants.

Religion and beliefs. Changes in religion in Turkey are controversial. The state is constitutionally secular, but it still controls the religious establishment through the Presidency of Religious Affairs, attached to the prime minister's office. During the early republican period up to 1950 no religious teaching was permitted, but now modern religious schools and theology faculties have been created, and religious lessons are allowed in state schools. Many courses and groups outside the state system have been set up to teach children religion, and the number of new mosques is very large. Thus the deep attachment of the majority to Islām has been demonstrated. The Shīʿī Muslim minority are said to support firm official secularism, largely because they have tended to benefit from the weakening of orthodox institutions; they were never part of the national system, and thus were less directly affected by official secularism. In the long term, at least among the younger educated generation, a decline in the range of activities influenced by Islām would seem probable, along with a decline in the numbers and devoutness of Muslims. But except for a small elite, the Turkish people remain deeply committed to their identity as Muslims and to an Islāmic world view.

New elements of independence among youth

Kinship and marriage. Changes in kinship, family, and marriage seem to stem from economic and demographic changes. Young men can now easily establish economic independence and the very large number of children who now survive creates new problems in large joint households. Far more men with living fathers are heads of independent households, though the proportion of village three-generation households has also increased. Young men still expect traditional marriages arranged by their elders; and a respectable bride spends, if possible, at least the first year of her married life in the house of her husband's father.

According to the ideal, one married son should in any case remain with the parents for good. Within the household there appears to be a decline in paternal control and authority, but the young are still deferential, and the basic attitudes toward the roles of the sexes and the emphasis on feminine modesty have changed little. Universal formal education and the possibilities of upward social mobility or migration for work have given young people a different view of the world from that of their fathers, but this generation gap is new, and significant changes in customary behaviour are slow in developing.

The new household

One radical change is partly concealed by the fact that most village households retain land. Traditionally, almost all households were economic units for production as well as consumption. In spite of segregation of the sexes, men and women lived and worked in daily contiguity and complementary dependence within a household. What is emerging is a new type of household—one in which the breadwinner's occupation has nothing to do with the household and in which the women are left with purely housekeeping tasks. In the villages, many men are likely to be away for months at city jobs, managing their own domesticity; some agricultural tasks may be performed by machine-owning neighbours and not the household women and children. In the towns, men are away during working hours, and a wife is confined to domestic tasks in the home and immediate neighbourhood. If she does earn, her work may be quite unrelated to her home.

In some cases, the ties of kinship between households may be weakened, both by permanent migration and by the increase in other types of social relationships; but the new situation also provides new opportunities to exploit kinship ties—especially in finding jobs, in providing short-term hospitality to migrants, and in solving urban housing problems. Similarly, while in some cases kinship loyalties may decline sufficiently to reduce feuding, homicide resulting from feuds is by no means a thing of the past.

Impact of change. These changes taken together involve a large cultural input of ideas, symbols, techniques, and social arrangements, borrowed from more industrialized societies, especially Europe and the United States. But it is naive to contrast "modern" and "traditional" as though the one is simply replacing the other. The processes of change are far more complex and subtle. The changing pattern of social relationships undoubtedly is diminishing local social controls and spreading information, technology, ideas, and consumer goods, all of which necessarily increases knowledge and dramatically alters some specific goals and standards. Yet the more fundamental assumptions and values of Anatolian culture persist through this multiplexity of changes.

BIBLIOGRAPHY. No good general books in English cover the material of this article. The only comprehensive work is

in French and quite out of date: VITAL CUINET, *La Turquie d'Asie*, 4 vol. (1890–95). Otherwise, one must depend on a vast range of specific studies, travel books, histories, and so forth. Much good recent work is in Turkish, French, or German—especially, for example, on nomads. One excellent work that illustrates the contemporary complexity is XAVIER DE PLANHOL, *De la plaine pamphylienne aux lacs pisidiens, nomadisme et vie paysanne* (1958), a human geography of the region of Anatalya. SETON LLOYD, *Early Anatolia* (1956), summarizes pre-classical archaeology up to that date. FREDERICK W. HASLUCK, *Christianity and Islam Under the Sultans*, 2 vol. (1929), emphasizes popular cults and the transition to Islām. MARY GOUGH, *The Plain and Rough Places* (U.S. title, *Travel into Yesterday*; 1954), is a readable personal account of modern Turkey. For Istanbul traditions, see HALIDE EDIB ADIVAR, *Memoirs of Halidé Edib* (1926); for a recent study of Turkish villagers, see PAUL STIRLING, *Turkish Village* (1965). On recent changes, see EDWIN COHEN, *Turkish Economic, Social and Political Change* (1970), a general optimistic appraisal; and JAN HINDERINK and M.B. KIRAY, *Social Stratification As an Obstacle to Development* (1970), a sociological account of drastic changes in four villages near Adana. A useful introduction to modern Turkey is GEOFFREY L. LEWIS, *Turkey*, 3rd ed. (1965).

Anatolian Languages

The term Anatolian languages in its most comprehensive use includes both the Indo-European and non-Indo-European languages spoken in Anatolia (Asia Minor) before the Greco-Roman period. The Anatolian languages are known only from texts of the 2nd and 1st millennia BC; the earliest evidence is that of the so-called Cappadocian tablets (19th–18th centuries BC). The term Asianic is sometimes used as an alternate designation for the Anatolian languages, but, since the discovery in 1915 that Hittite, the main Anatolian language, is an Indo-European language, there has been a tendency to use Asianic in a more restricted sense for the non-Indo-European languages that existed in Anatolia before the entry of the Indo-Europeans. These are called substratum languages.

Hattic (or Hattian), also misleadingly called Proto-Hittite, is the best known substratum language. It is completely unrelated to Hittite and its sister languages as well as to Hurrian, a language also spoken in Anatolia.

The Anatolian group of Indo-European languages consists of Hittite, Palaic, Luwian, Hieroglyphic Luwian, Lydian, and Lycian. Hittite, Palaic, and Luwian are known from 2nd-millennium cuneiform texts found in the excavations in Boğazköy-Hattusas since 1905; Hieroglyphic Luwian is found on scattered inscriptions and seals from Anatolia (mainly the southern area) and northern Syria dating mainly from later times (*i.e.*, between *c.* 1200 and 700 BC, although there are earlier examples from the empire period, *c.* 1400–*c.* 1190 BC). Lydian and Lycian are known from texts in alphabetic script from *c.* 600 to 200 BC. It seems fairly reasonable to add the Carian language of southwest Anatolia to this list as well as other less well documented languages like Sidetic. More to the east, in the Caucasus region centring around Lake Van, Hurrian of the 3rd and 2nd millennia BC was replaced in the 1st millennium BC by the related Urartian language. Both of these languages are definitely non-Indo-European.

Historical background of ancient Anatolia. It is customarily assumed that the Indo-Europeans entered Anatolia around or shortly after 2000 BC, although there are no specific archaeological data that might enable scholars to specify the period of entry or the route the invaders followed. On the basis of the agricultural terminology used in Hittite, it has been suggested that the entry into Anatolia was not a warlike invasion of predominantly male groups. If such had been the case, the influence of substratum languages would have been likely, but, on the contrary, the word stems used are definitely Indo-European. The differences in the terminology used in other Indo-European subgroups indicate that the "Anatolians" seceded from the parent group at an early date, before the common agricultural nomenclature came into being. On the other hand, Hittite shares the Indo-European notion of the hereafter, pictured as a pastureland with grazing cattle "for which the dead king sets out."

Indo-European entry into Anatolia

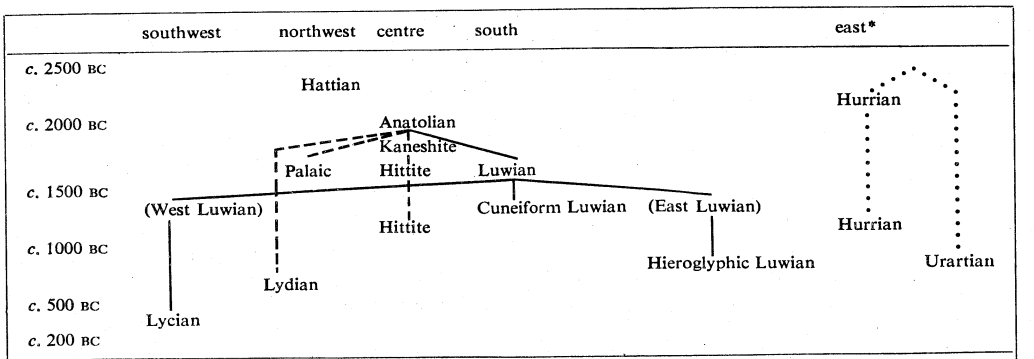
In recent times, there has been a tendency among linguists to postulate an eastern route of entry into Anatolia by way of the Caucasus, because certain grammatical features—*e.g.*, the loss of the feminine gender—might be explained as having been caused by prolonged contacts with Caucasian languages. It is likely that the Indo-European forebears of the later speakers of Hittite, Palaic, Luwian, and Lydian entered Anatolia together, following a common route, because the Anatolian languages share a considerable number of losses as well as innovations that presuppose a long common past.

In the central parts of Anatolia, within the bend of the Halys River (modern Turkish, Kızıl Irmak), and in the northern regions, Hittite and Palaic were profoundly influenced by Hattic as a substratum language. The Hattian culture also changed the political and religious concepts of the newcomers, and a clear cultural dependency of the Indo-Europeans on the older Hattian population is evident. Some scholars have stressed the likelihood that farther to the south the Luwians might have been conversant with a different substratum. In view of the absence of textual evidence, and because knowledge of the Luwian vocabulary is rather restricted, it is perhaps not surprising that this possible substratum element escapes definition. (For the history of Anatolia in the 2nd and 1st millennia BC, see ANATOLIA, ANCIENT.)

The most important invaders of Anatolia in the "Dark Age" (after 1190 BC) were the Phrygians. Their language is definitely Indo-European, but it bears no relationship to the Anatolian subgroup. Rather, it seems akin to Thracian, Illyrian, or possibly Greek. Greek, in the second half of the 1st millennium BC, and, later, Latin, from the 2nd century onward, entered central Anatolia as languages of a ruling caste. Much earlier—beginning in Mycenaean times—the west coast had attracted Greek settlers. In the first half of the 1st millennium, the southern and northern shores also attracted Greek-speaking peoples. To the east in the Caucasus region, other Indo-Europeans, the Armenian-speaking invaders, penetrated into the former Urartian territory well before the beginning of the Persian period, probably in the 7th and 6th centuries BC. During Persian times, a Persian ruling caste entered eastern and also northeastern Anatolia and was still clearly recognizable in the Hellenistic and Roman periods (*e.g.*, in Bithynia, Pontus, Cappadocia, and Commagene). Late data on names and scattered remarks made by Patristic authors (Fathers of the Church) indicate that until late Roman and perhaps even Byzantine times, some Anatolian dialects remained in use in certain isolated parts of the interior.

Early research on Anatolian languages

Classification of the languages. Research on the Anatolian languages began in 1821 with the Lycian language and passed an initially fruitful phase in the 1880s with work on Hieroglyphic Hittite (nowadays referred to as Hieroglyphic Luwian). In 1902, the Norwegian Assyriologist Jørgen Alexander Knudtzon's study on the Arzawa letters was published; these were two letters exchanged between a king of Arzawa and Pharaoh Amenhotep III that had been found in the Amarna archive. They were written in the Hittite language in cuneiform writing. In 1915, research reached a climax with the interpretation of Cuneiform Hittite by the Czech Orientalist Bedřich Hrozný. In all four of these highlights, the discovery that the texts in question were Indo-European was either clearly expressed or more discreetly implied. This conclusion was based on both the nominal (noun) declension and the verbal conjugation: the languages had a nominative ending in *-s*, the accusative in *-n*, verbal endings like *-ti* and *-nti* for the 3rd person singular and plural of the present tense, and an imperative form like *estu* "let it be." These features were deemed to be sufficient proof of their Indo-European origin. Study of the Anatolian subgroup of Indo-European thus began with Lycian, the last Anatolian offshoot in the temporal sequence, then passed the intermediary stage of Hieroglyphic Luwian, and reached the 2nd-millennium Hittite language in 20th-century research. For the relationship between members of the Anatolian subgroup, see the accompanying diagram.



*Hurrian and Urartian are not related to the Indo-European group of Anatolian languages.

The non-Indo-European Hurrian and Urartian languages are related to one another, but modern research indicates that Urartian should not be considered as a direct continuation of Hurrian.

HISTORY AND DEVELOPMENT OF THE ANATOLIAN LANGUAGES

Languages using cuneiform writing and Anatolian hieroglyphs. *Hattic.* The Hattic language appears as *hattili* in Hittite cuneiform texts. Called Proto-Hittite by some, it was the language of the linguistic substratum inside the Halys River bend and in more northerly regions. Apparently the Indo-European newcomers of Hittite stock were named with the same designation as their predecessors. All the Hattic material preserved by Hittite scribes belongs to the religious sphere of life: rituals (*e.g.*, connected with the erection of a new building), incantations, antiphons, litanies, and myths. Among the Hattic interpolations in Hittite texts, there are some to which a Hittite translation has been added. It is impossible to ascertain the length of time that the Hattians had been present in Anatolia before the Indo-Europeans entered the country, but it seems certain that during the Hittite New Empire (*c.* 1400–*c.* 1190 BC) Hattic was a dead language.

Records of Hattic

Hattic studies began in 1922 with the work of the German Assyriologist Emil Forrer. In 1935, Hans G. Güterbock, a German-born Orientalist, published a large group of texts containing Hattic material and in so doing completed the publication of the Hattic texts stemming from the Winckler excavations (1905–12). Important studies on the subject have continued to appear since then.

Hittite. The Hittite language is known from the approximately 25,000 tablets or fragments of tablets preserved in the archives of Boğazköy-Hattusas, excavated by German archaeologists beginning in 1905. In Hittite cuneiform texts, the language is referred to as *nesili* (*nasili*) "language of Nesa," or *nesumnili* "language of the Neshite." Earlier Hittite linguistic material may be found in the indigenous proper names and a few loanwords from the local dialect that are recorded in the Cappadocian tablets (the commercial correspondence in Assyrian of Assyrian colonists living in Anatolia, especially in the emporium at Kültepe, near modern Kayseri, between *c.* 1900 and 1720 BC). The data from Kültepe are sometimes referred to as "Kaneshite" (from Kanesh, the old name of Kültepe); this is obviously the modern equivalent of the word *kanisumnili* "language of the Kaneshite" found in a Hittite text. It is possible, or even likely, that Kanesh and Nesa do, in fact, refer to the same entity.

Hittite tablets from places outside of the Hittite capital are rare; only stray examples have been found—*e.g.*, in Tarsus, Alalakh, Ugarit, and Amarna. These findings attest to the growth of a great Hittite empire, especially between *c.* 1400 and *c.* 1190 BC. Old Hittite, the written embodiment of the earliest Indo-European language that has been discovered so far, is known from some tablets preserved in an "old ductus" type of handwriting that was typical of copies from the Old Kingdom period (*c.* 1700–1500 BC). The intermediary "Dark Age" between *c.* 1500 and *c.* 1400 BC is sometimes referred to as the period of the so-called Middle Hittite language. Most of the Old and Middle Hittite texts, however, are preserved in copies from the later empire period.

Archives of Boğazköy-Hattusas

The archives of Boğazköy-Hattusas have been found in various places in the citadel, in the Great Temple complex, and in the "House on the Slope." Although the majority of the texts are concerned with religious subjects (oracle texts, hymns, prayers, myths, rituals, and festival texts), these archives also contain material of historical, political, administrative, literary, and legal character. The cuneiform adopted by the Hittite scribes is a variant of a writing system of Mesopotamian origin that closely resembles the ductus and shapes prevalent in tablets of the 17th century BC (layer VII) from Alalakh (modern Atsana in southeastern Turkey). It is possible that the cuneiform script might have been introduced as a result of the Hittites inducing Syrian scribes to transfer their activities to the Hittite capital during the early part of the Old Kingdom, shortly after 1650 BC. It has also been posited, with good reason, that the newly acquired script was first used to write Akkadian and was only later employed for Hittite as well. In addition to the genres enumerated above, the "scholarly literature" deserves to be mentioned. This consists of the material considered by the scribes to be essential for their training; it includes word lists, omens, and ritual prescriptions, all reflecting an encyclopaedic approach aimed at complete coverage of the subjects concerned. The Sumerian texts found in these archives belong to this class of literature. For treaties and correspondence with foreign powers, Akkadian was used as the diplomatic language of that period. Therefore, both Sumerian and Akkadian formed part of the curriculum of the qualified scribes, these languages belonging to the "eight languages" found in the Hittite archives.

In actual fact, the first decipherer of Hittite was the Norwegian scholar J.A. Knudtzon, who pointed out in 1902 that the language of the so-called Arzawa letters (*i.e.*, Hittite)—found in the Amarna archive—had an apparent affinity with Indo-European. Because the cuneiform script had already been deciphered, Knudtzon, and Bedřich Hrozný after him, were able to "read" their texts. Thus their discovery consisted more in the interpretation than in the actual decipherment of the written material. The first series of German excavations, lasting from 1905 to 1912, produced about 10,000 tablets. It was work on this corpus that familiarized Hrozný with the contents of these tablets and led him to his epoch-making discovery that Hittite was indeed Indo-European (1915).

Palaic. Palaic, which appears as *Palaumnili* "language of the Palaite" in Hittite cuneiform texts, was the language of the region of Pala (probably Blaëne in the Greek period), in northwest Anatolia. During the Old Hittite kingdom, Pala, Luwiya, and Hattusas formed the three major provinces of the Anatolian part of the Hittite territory. From the intermediary "Dark Age" onward, Kaska nomads made their influence felt in northern Anatolia, and this resulted in a decline of importance for this region.

Linguistic affiliations of Palaic

The Indo-European character of Palaic was first advocated by Emil Forrer (1922). Part of the text material is preserved on tablets in "old ductus." The knowledge of the limited vocabulary leaves much to be desired, but parallels—especially in the inflection of the noun, the forms of the demonstrative, relative, and enclitic pronouns, and the verbal endings—vouch for a close relationship to Hittite and Luwian.

Luwian. Luwian (or Luvian), the language of Anatolia's southern coast, is known from texts stemming from three major periods: (1) the Hittite New Empire (*c.* 1400–*c.* 1190 BC); (2) the period of the Neo-Hittite states (*c.* 1190–*c.* 700 BC); (3) the period of the Lycian monumental inscriptions (*c.* 400–200 BC). In addition to the various time periods, there is also a variation in writing system—Mesopotamian cuneiform, Anatolian hieroglyphs, and an alphabet derived from a Greek source—and dialectal differentiation. There are indications that as early as the 15th and 14th centuries BC, there was a West Luwian dialect (the precursor of alphabetic Lycian) and an East Luwian dialect (the forerunner of the later Hieroglyphic Luwian of the Neo-Hittite states). Both of these differed from the Luwian found in the archives of Boğazköy-Hattusas, which was possibly a central dialect.

As in the case of Palaic, the pioneering work on Luwian written in cuneiform was done by Emil Forrer (1922). Following this work, new text materials were published in 1953, closely followed by both grammatical and vocabulary studies as well as a standard dictionary of Cuneiform Luwian (1959).

The Anatolian hieroglyphic system has a long history, with its logographic beginnings dating back to early Hittite stamp seals of the 18th and 17th centuries BC; the youngest texts seem to date from the last quarter of the 8th century BC. The geographical range of the inscriptions is great, stretching from Sipylus and Karabel in the extreme west to Alaca Hüyük and Boğazköy-Hattusas in the north, Malatya, Samsat, and Tell Ahmar (Til Barsib) in the east, and Hama and ar-Rastān in the south. During the "Dark Age" of the 16th and 15th centuries BC, the early writing grew into a fully developed writing system with logograms (word-signs), syllabic values, and auxiliary signs. During the New Empire, the script was already in use for a multitude of purposes (rock inscriptions, seals, and wooden tablets for everyday use in the temple and the army). Whether an example of the empire period such as the Aleppo inscription already reflects the Luwian language is a moot question but seems likely. It is certain that the later inscriptions of the Neo-Hittite states were in Luwian.

Work on Hieroglyphic Luwian

The first attempts to decipher Hieroglyphic Luwian, made by the British archaeologist Archibald H. Sayce, were fortunate in some fundamental details, but it was not until the 1930s that systematic and mutually stimulating research by scholars of several countries led to the establishment of a number of syllabic values for the characters as well as to a correct analysis of the sentence structure of the inscriptions. In his publication of the (bilingual) Hittite royal seals (in 1940, 1942), Hans G. Güterbock bridged the gap between the inscriptions of the empire period and the late Neo-Hittite states; the seals found in the French excavations at Ugarit (in northern Syria) served a similar purpose. The most important recent finding was the discovery in 1947 by Helmuth T. Bossert, a German archaeologist, of the Karatepe bilingual inscriptions, written in Phoenician and Hieroglyphic Luwian.

On many points the Luwian vocabulary is still an enigma. The unity between the various Luwian dialects and the close relationship of Luwian to the other members of the Anatolian subgroup, however, is secured by several linguistic parallels, especially in the singular inflection of the noun, the forms of certain pronouns, the verbal endings, and a number of lexical (vocabulary) correspondences.

Hurrian. In earlier stages of research, the terms Mitanni language and Subarian were used as designations for Hurrian. In Hittite cuneiform texts, *hurlili* "language of the Hurrian" is used. In the last centuries of the 3rd millennium BC, Hurrians were already present in the Mardin region, which, from a geographical point of view, belongs to the North Mesopotamian plain. In Mesopotamian texts (from the time of the Akkad dynasty) some Hurrian personal names and glosses have been found. The customary assumption is that this non-Semitic and also non-Indo-European ethnic group had come from the Armenian mountains. During the beginning of the 2nd millennium BC, the Hurrians apparently spread over larger parts of southeast Anatolia and northern Mesopotamia. Still later, during the intermediary "Dark Age," they are supposed to have infiltrated into Cilicia and the adjacent Taurus and Antitaurus regions (Kizzuwadna in 2nd millennium texts). Before the middle of the 2nd millennium BC, an Indo-Aryan ruling caste wielded some type of authority over parts of Hurrian territory. Some names and words in ancient Near Eastern texts bear witness to their presence. Among these words are a group of technical terms related to the training of horses that found its way into Hittite treatises on that subject; they are most important from a historical point of view. After Sumerian, Akkadian, Hattic, Palaic, and Luwian, Hurrian and these Indo-Aryan glosses constitute the sixth and seventh additional languages of the Hittite archives.

Sources of Hurrian texts

Hurrian texts have been found in Urkish (Mardin region, *c.* 2300 BC), Mari (on the middle Euphrates, 18th century BC), Amarna (Egypt, *c.* 1400 BC), Boğazköy-Hattusas (Empire period), and Ugarit (on the coastline of northern Syria, 14th century). Amarna yielded the most important Hurrian document, a political letter sent to Pharaoh Amenhotep III. From Mari came a small number of religious texts; from Boğazköy-Hattusas, literary and religious texts; and from Ugarit, vocabularies belonging to the more "scholarly literature" described above and Hurrian religious texts in Ugaritic alphabetic script. Hurrian personal names, found in texts from many sites (Boğazköy-Hattusas, Alalakh, Ugarit, and especially Nuzu), constitute a second linguistic source of major importance.

The research on Hurrian started in the 1890s with simultaneous contributions by several scholars. Subsequently, Bedřich Hrozný (1920) and Emil Forrer (1919, 1922) discovered the presence of Hurrian material in the Boğazköy-Hattusas archives.

Urartian. The terms Chaldean and Vannic have also been used as designations for Urartian during earlier stages of research. Urartian is not a late dialect of Hurrian but a separate language, although both stem from a common parent. During the 9th through 6th centuries BC, Urartian was used in northeastern Anatolia as the official language of the state of Urartu, which centred around the district of Lake Van but also extended over the Transcaucasian regions of modern Russia and into northwestern Iran and at times even into parts of North Syria. The Urartian texts are written in a variant of the Neo-Assyrian script and consist mostly of monumental inscriptions (annals, votive inscriptions related to building and irrigation activities), some small inscriptions on helmets and shields dedicated in the temple, and a few economic cuneiform tablets. Two bilingual inscriptions in Urartian and Assyrian that apparently correspond very closely provided the key to the understanding of the language; the stylistic resemblances to Assyrian texts of the same period guided the further interpretation.

Archibald H. Sayce was the first scholar to devote his attention to Urartian in the 1880s and 1890s and continued his activities until 1932. More important were the philological contributions of the German historian Carl F. Lehmann-Haupt between 1892 and 1935. The first reliable description of Urartian grammar was published by the German Orientalist Johannes Friedrich (1933).

Next to the Urartian texts in cuneiform writing, there also existed an indigenous hieroglyphic script that is still undeciphered and is too meagrely represented to warrant a serious attempt.

Languages using a derivative of the Phoenician or Greek alphabet. *Phrygian.* The Phrygian inscriptions and graffiti may be separated into two groups, the Old Phrygian

Old Phrygian and New Phrygian

texts in a typical Phrygian alphabet dating from *c.* 730–450 BC, and the New Phrygian inscriptions (sepulchral texts in the Greek alphabet) stemming from the 1st and 2nd centuries AD. The Old Phrygian texts may be divided into a central group (Midas City and the central area) and an eastern group (found in Gordium), with offshoots in a still more eastern direction marking the utmost Phrygian expansion (inscriptions in or around Hüyük near Alaca, in Boğazköy-Hattusas, and in Tyana). An important recent finding—and the longest Old Phrygian text to date—is the rock inscription near the village of Germanos (modern Soğuk Çam) in Bithynia (found in 1966). The total number of Old Phrygian texts now stands at about 80; more than 50 of these are from Gordium and represent about one-quarter of the available material. There is a consensus of opinion on the Indo-European character of the Phrygian language; most scholars think that Phrygian is somehow connected to the Greek branch of Indo-European languages, although, at an earlier stage, some scholars considered the possibility of a connection with the Anatolian branch of Indo-European, and others proposed a relationship with Thracian and Illyrian.

In a publication of new material from Gordium, the U.S. archaeologist Rodney S. Young cautiously suggested that the Old Phrygian alphabet may be dependent on a prototype in use on the North Syrian or Cilician coasts. The old idea that the Phrygian alphabet was dependent on a Greek one (and not vice versa) need not be abandoned in that case. Historically, such a derivation would present no problems, because the presence of Greek settlements in these areas during the second half of the 8th century BC is amply attested to by both Assyrian annals and late Greek historical sources as well as by archaeological findings. Internal evidence from the Phrygian alphabet, presented by the French linguist Michel Lejeune, serves as proof for some researchers that that alphabet derived from the Greek one.

Lydian. Of the more than 70 Lydian texts (*e.g.*, sepulchral inscriptions, votive texts, many graffiti), more than half have been found by United States excavators at the Lydian capital, Sardis. Two small Greek–Lydian bilingual texts were far less helpful than the famous Aramaic–Lydian text. A few texts (about ten) may go back to the 6th or 5th centuries BC, but many more stem from the 4th century. The Lydian alphabet was derived from an East Greek prototype; the superfluous signs in the Greek alphabet were used for specific Lydian sounds, and additional signs were either borrowed from other "Anatolian alphabets" or freely created.

Research on Lydian

Important results concerning Lydian were reached using a strictly combinatory method; *i.e.*, passages were compared that expressed similar contents in a slightly different manner in order to obtain a better understanding of the language's structure. This stage of the research culminated in a conclusive article by the Italian Piero Meriggi on the Indo-European character of Lydian (1936). Subsequently, other scholars published evaluations of the Lydian data, a dictionary, and a grammar book. The study of Lydian is hampered by many lexicological uncertainties, but there is at least a growing consensus on matters of grammar leading to the common notion that Lydian belongs to the Anatolian subgroup of Indo-European. The final obstacle to this classification as Anatolian was removed in 1959 by the Italian Onofrio Carruba, who proved that Lydian, like the other members of the Anatolian branch, does not possess a separate feminine gender. Lydian shares common features with Hittite, Palaic, and Luwian and should therefore be acknowledged, it seems, as a fourth independent member of the Anatolian subgroup.

Carian. A great number of the more than 100 Carian inscriptions are graffiti found in Egypt that were left behind by Carian mercenaries in the services of Egyptian pharaohs of the Saitic period (664–525 BC). In recent years, more monumental inscriptions have been found in Caria itself, while Carian clay tablets have also been discovered. In the mid-20th century, several scholars concluded that Carian writing consists of a purely alphabetic script and is not a mixed system of both single letters and syllabic signs as was formerly thought. It is a likely but still unproven assumption that Carian may also be classified in the Anatolian subgroup of Indo-European.

Lycian. More than 150 Lycian monumental inscriptions have been found so far, which, with very few exceptions, are sepulchral in character. They are written in an indigenous Lycian alphabet that is based not on an East Greek prototype (as its Lydian replica) but on a West Greek one. Although the Lycian coin legends are still usually dated from the period between 500 and about 360 BC, the tradition of the Lycian monumental inscriptions is now thought to have continued for a longer period, into the 3rd century BC. During the beginning of the research in the first half of the 19th century, extensive use was made of a good bilingual text that offers a faithful Greek translation. In the first phase of research, which ended about 1880, Lycian was investigated by an etymological method by which it was linked up either with Greek or the Iranian languages. A more reliable combinatory method was later introduced, but the most fecund phase in the study of Lycian occurred at the end of the 19th century, when the Scandinavian school of scholars cooperated closely in the publication of several important studies. In 1945, Holger Pedersen published a synthesis of all data that seemed to indicate a relationship of Lycian with Hittite; thus Pedersen proved conclusively that Lycian belongs to the Anatolian branch of Indo-European languages. This conclusion was slightly modified when the British scholar Franz J. Tritsch (in 1950), and, later, the French scholar Emmanuel Laroche showed that Lycian should be more specifically compared to Luwian.

Relationship of Lycian with Anatolian

Sidetic. The historical detail preserved by the Greek historian Arrian that the city of Side on the Pamphylian coast possessed a particular, indigenous language has been strikingly confirmed by legends on Sidetan coins of the 5th(?) through the 3rd(?) centuries and by five inscriptions from the 3rd and 2nd centuries BC (two of which are bilingual). There is a curious likelihood that this alphabet was directly derived from a Semitic writing system rather than from a Greek prototype, but Greek influence was not absent, as is clearly evidenced in the Greek bilingual texts and by a loanword from Greek. The first reliable study of Sidetic was made by Helmuth T. Bossert in 1950. In the case of Sidetic, even the value of a group of signs is still undecided, and research has not yet reached a stage in which a fruitful analysis of the texts and a classification of the language are within sight.

LINGUISTIC CHARACTERISTICS OF THE ANATOLIAN LANGUAGES

Non-Indo-European languages. The non-Indo-European Hattic is an agglutinative language; that is, it combines several elements of meaning into a single word. In the conjugation of verbs, it uses prefixes that are attached to the word stems, which are mostly monosyllabic or bisyllabic. Hattic nouns consist of a free number of syllables and have both prefixes and suffixes. There are, however, no formal distinctive features to distinguish nouns and verbs.

Both the Hurrian and the Urartian languages differentiate between stems and suffixes, but there is again no sharp distinction between noun and verb. Many suffixes may be added onto one another in a row, but within the often prolonged suffix series a detailed order is rigidly observed. Among the suffixes added to the noun, several subgroups are distinguished; one group might be compared to the case endings of the Indo-European languages. One of the most characteristic phenomena of this group is the distinction between a subject case (the "nominative") and an "agentive" case. The agentive marks the actor or subject of a transitive verb when the object is expressed by its counterpart, the "nominative." The subject case is characterized by a lack of ending on the stem; it marks the subject in nominal sentences (sentences without verbs) and occurs with intransitive verbs and as the object of transitive verbs.

Subject and agentive cases in Hurrian and Urartian

Phrygian. The New Phrygian texts especially favour the attribution of the Phrygian language to Indo-Euro-

pean. They contain such data as *ios* as relative pronoun (Indo-European **io-s*, Greek *hos*), a demonstrative pronoun that is either comparable to Indo-European **ki-/ko-* or to **so* (an asterisk indicates a hypothetical reconstructed form), and the form *ad-daket* "he adds" related to Latin *addit* and to Greek *é-thē-ka*.

The Anatolian subgroup of Indo-European. *Grammatical characteristics.* Characteristic of the Anatolian languages is the absence of the dual number ("you and I") and the lack of feminine gender in the declension of nominals (nouns, pronouns, and adjectives). There is a division between an animate (common) gender and an inanimate (neuter) gender. In Hittite, a neuter may not be the subject of a transitive action verb; in that case, an *-ant* suffix is added before the neuter nominative ending in *-s*. This *-s* ending persists in the whole subgroup. The case system of Old Hittite is still fairly complicated, but in the subgroup as a whole there is a clear tendency toward a greater simplicity. The presence in Hittite of an archaic irregular class of nouns is a striking characteristic; *e.g.*, there are alternate *r* and *n* stems as in *utter/uttanas* "word, affair" and *water/witenas* "water."

The Anatolian inflection of pronouns conforms to the traditional Indo-European pattern by being different from that of the nouns, but, at the same time, it shows some striking peculiarities. Typical Anatolian pronouns are: Hieroglyphic Luwian *amu*, equivalent to Lycian *amu*, *emu*, *ẽmu* "I, me" (compare Hittite nominative *uk*, accusative *ammuk*); and Hittite nominative *zik*, accusative *tuk* "you," as compared to *ti*, *tu* in Palaic. Some of the languages have enclitic pronouns; *i.e.*, pronouns pronounced as being part of the preceding word. A demonstrative pronoun *aba-* ("that," but in some member languages also "this") is present in Hittite, Palaic, Cuneiform and Hieroglyphic Luwian, Lycian (*ebe-*), and Lydian (*ebad* "here, there"), and an interrogative or relative pronoun *kui-* (compare Latin *quis*) is common to Hittite, Palaic, and Cuneiform Luwian. The corresponding terms for *kui-* in Hieroglyphic Luwian, Lycian, and Lydian also seem to be phonetic variants of the same original pronoun.

Anatolian verbs

The Anatolian verbal system is simple: it has two moods (indicative and imperative) and two tenses (present and preterite). There are some traces—either to be classified as debris or as the nucleus for a future development—of an aorist *-s* fixed to the stem; *e.g.*, *kaness-* "to acknowledge" (compare Greek *gi-gnō-sk-ō*); *kalless-* "to call" (compare Greek *kaleō*, aorist *é-kale-s-a*). (The aorist is a verb form denoting action without reference to its duration or completion.) A mediopassive "voice" is present in Hittite (*es-a-ri* "he is seated"; *ki-tta-ri* "he is lying"), Palaic, Luwian, and perhaps in Lydian. (The mediopassive expresses a type of reflexive meaning ["He washes himself"] or passive meaning ["He is being washed"].)

Reduplication (repetition) of the verbal stem occurs in the entire Anatolian subgroup. It adds an iterative or intensive nuance to the meaning, but it does not function in a system of tenses as in Greek. Very typical of the Anatolian subgroup are verbal suffixes like the causative *-nu-* (compare Hittite *war-* "to burn," *warnu-* "to kindle," *harg-* "to perish," *harganu-* "to ruin, to destroy"). In principle, these formations can be built on any verbal stem whose meaning permits such an addition. It should be stressed that in Hittite a normal expression for a "state" consists of a nominal sentence (that is, a sentence without a verb but with a noun, an adverbial expression, or a participle as predicate); sometimes, however, the verb *es-* "to be" is used as the carrier of modal or temporal nuances. The total absence of the Indo-European perfect (describing a "state" resulting from a recently concluded action) becomes very clear by the usage of the adverb *nawi* "not yet," which occurs with a present tense in Hittite (but which would employ a perfect tense, such as "has been," in English and other Indo-European tongues).

Very characteristic of the Anatolian subgroup is a strong preference for the linking together of particles and enclitic pronouns to form "chains" that are placed at the beginning of the sentence or clause. The first component of such a "chain" usually is a stressed part of the sentence or otherwise a sentence connective (like *nu* in Hittite, *a* in Luwian).

Phonological characteristics. In the Anatolian vowel system, *a*, *e*, *i*, and *u* are present, but *o* is curiously absent. In Lycian, the Greek value omicron has been used for the Lycian *u*, but in Lydian the existing *o* seems to be a secondary development. A main dialectal criterion is the treatment of Anatolian *e*: in Old Hittite, there still was a differentiation between *e* and *i*, but in later Hittite, an *-e* at the end of a word changed to *-i*. In Luwian, *e* tended to appear as *a*. Vowel gradation (*i.e.*, a change of vowel) that reflects meaning change plays a role in Hittite (*e-es-zi* "he is" versus *a-sa-an-zi* "they are") but was impossible in Luwian because of the sound change. Both Lycian and Lydian possess separate signs for nasalized vowels (*ã* and *ẽ*).

Laryngeal theory and Anatolian data

Advocates of the so-called laryngeal theory (first proposed by the Swiss linguist Ferdinand de Saussure in 1879) have found their postulate partly confirmed by Anatolian data. This theory maintains that the different forms of certain words in the various Indo-European subgroups can be satisfactorily explained only by assuming that all the known Indo-European languages have lost certain guttural sounds (laryngeals) that were originally present in the parent speech. In 1927, both the Polish linguist Jerzy Kuryłowicz and the French scholar Albert L.M. Cuny announced their discovery that in Hittite an *h* sound was preserved in positions in which a laryngeal would have formerly been (compare Hittite *hant-* "front" to English *anti-*; Hittite *pahhur* "fire" but English *pyre*). But the Anatolian evidence for the laryngeal theory is certainly not without problems, and the adherents of the theory consider that other laryngeals disappeared in Hittite as well.

Lexical data. Some examples of correspondences in vocabulary are given in the Table. It has often been remarked—and not without reason—that although the grammar of the Anatolian languages would be recognizably Indo-European, the vocabulary would be less so. This is usually attributed to the deeply penetrating influences exercised by strange surroundings, not only while the "Anatolians" were "en route" but also after their arrival in Anatolia.

The relationship with the other subgroups. The relationship of the Anatolian branch to the rest of Indo-European has often been defined in the United States on

Anatolian Lexical Correspondences

	Hittite	Palaic	Cuneiform Luwian	Hieroglyphic Luwian	Lycian	Lydian
to make, to do	iya-		aya-	aia-	a-/e-	i- (?)
in	anda		anda	anta	ñta/e	(-)ẽn-/(-)ẽt-
to be	es-/as-	es-/as-	as-	as-	es-	e- (?)
house	pir-/parn-		parn-	parn-	*prñn-	bira-
to give	pai-/piya-	pi(ya)-	piya-	pi-	pije-	bi-
up	ser/sara		sarri	logogram	hri	
high, superior	sarazzi-			*sarli- (?)	hrzzi-	serli-
god	siu-/siwi-/siwa-/siun-/siuni-/siuna-		massani-	masana/i-	mahani-	civ-
divine	siunali-		massanalli-			civvali-

*Indicates that the following form is an unattested, reconstructed form.

Indo-Hittite hypothesis

the basis of the "Indo-Hittite hypothesis." That is, Hittite or Anatolian on the one hand and Proto-Indo-European on the other were both supposed to descend from a common parent. This hypothesis attributes too much weight to the Anatolian evidence. It was demonstrated as early as 1938 that the Anatolian branch should be placed on a par with the rest of the Indo-European subgroups and not as a coequal with Indo-European itself. Nowadays, the Indo-Hittite hypothesis is very rarely defended. Another extreme position states that the Hittite-Luwian-speaking group (another designation for the Anatolian subgroup) left the Indo-European parent group comparatively late, after the Greek and Armenian divisions had done so and approximately at the same time as Indo-Iranian. If this theory were true, there would be no need to use the Anatolian data for a thorough revision of the reconstructed Proto-Indo-European language, because these data would be less relevant, at least not more so than Indo-Iranian and Greek, on which the old reconstruction was based. A third opinion—prevalent in the French school of Indo-European studies—holds that the Hittite or, preferably, the "Common Anatolian" data are of special importance, because the Anatolian languages are particularly archaic. According to this theory, similarities in morphology (word elements) between the Celtic, Italic, and Hittite–Luwian groups and Tocharian (an Indo-European language of central Asia) seem to imply that the dialects from which these groups evolved were in peripheral positions in the Indo-European language area and were probably the first to move away from the main group.

BIBLIOGRAPHY. The best general handbook on Anatolian history and civilization, in German, is A. GOETZE, *Kleinasien*, 2nd ed. (1957), with succinct chapters on the Anatolian languages (pp. 45–63). An excellent, more popular account of Anatolia with marked emphasis on the Hittite period is O.R. GURNEY, *The Hittites*, rev. ed. (1961). This book deals with languages and races in ch. 6. The general handbook on the Anatolian languages, both Indo-European and non-Indo-European, written by scholars, is A. KAMMENHUBER *et al.*, *Altkleinasiatische Sprachen* (1969), although in some chapters the book is perhaps too detailed for the general reader. J. FRIEDRICH (ed.), *Kleinasiatische Sprachdenkmäler* (1932), is still a very valuable collection of texts, chosen to illustrate the various languages of Anatolia. Although written from the viewpoint of the Indo-Hittite hypothesis, E.H. STURTEVANT and E.A. HAHN, *A Comparative Grammar of the Hittite Language*, rev. ed. (1951), was a major achievement. In many respects it is still the best grammar of Hittite. H. PEDERSEN, *Hittitisch und die anderen Indoeuropäischen Sprachen* (1938), was directed against the Indo-Hittite hypothesis and should be used with Sturtevant and Hahn's book. J. PUHVEL's contribution "Dialectal Aspects of the Anatolian Branch of Indo-European," in H. BIRNBAUM and J. PUHVEL (eds.), *Ancient Indo-European Dialects*, pp. 235–247 (1966), gives a good survey of the main characteristics of the Anatolian subgroup. The best recent introduction to the "Indo-European problem" with a notable emphasis on Anatolian matters is R.A. CROSSLAND, *Immigrants from the North*, fasc. 60 of the *Cambridge Ancient History*, rev. ed. (1967). The best introduction to the Anatolian languages and to Hittite literature may be found in two highly authoritative and very well written scientific contributions by H.G. GUTERBOCK: "Toward a Definition of the Term Hittite," *Oriens*, 10:233–239 (1957) and "A View of Hittite Literature," *Journal of the American Oriental Society*, 84:107–115 (1964). The history of the various decipherments has been dealt with by J. FRIEDRICH, *Entzifferung verschollener Schriften und Sprachen*, 2nd ed. (1966).

(Ph.H.J.H.t.C.)

Ancestor Worship

The term ancestor worship describes, in a broad and loose sense, a variety of religious beliefs and practices concerned with the spirits of dead persons regarded as relatives, some of whom may be mythical. Although far from universal, ancestor worship exists or formerly existed in societies at every level of cultural development.

Nature and significance. The core of ancestor worship is the belief in the continuing existence of the dead and in a close relation between the living and the dead, who continue to influence the affairs of the living. Beliefs in a surviving element of the human person (*e.g.*, the soul) and in an afterlife have been held in all kinds of societies. Attitudes toward the spirits of the dead vary from love, respect, and trust, mingled with special feelings of reverence, to outright fear; the attitudes are sometimes ambivalent. The spirits of the dead are often thought to help the living, but they often are thought to do harm if they are not propitiated. All societies give ritual attention to death or to the souls of the dead, but not all of these practices may appropriately be called ancestor worship. If death itself, rather than the ancestral relationship, is the focus of attention, the name death cults is more appropriate. The deification of dead heroes is similarly poorly distinguished from ancestor worship. Death cults, the worship of dead heroes who may or may not be regarded as ancestors, and clearly distinguishable rites of ancestor worship may all exist in the same society.

Ancestors venerated by elaborate rites are those persons who in their lifetimes held positions of importance, such as heads of families, lineages, clans, tribes, kingdoms, and other social groups. Depending on the manner in which kin are organized into social groups, ancestral spirits that are worshipped may be limited to one sex, or may include both sexes. Among primitive societies that trace descent only through male lines, for example, the titular positions of prestige are held by males, and only male ancestors are significant.

Ancestral spirits that are worshipped also vary in nearness or remoteness in time from the living. In some societies only the spirits of the recently deceased are given attention; in others, all ancestors, near and remote in time, are included. In still other societies, one ancestor, real, honorary, or mythical, may be the focus of attention, and he is often regarded as a hero (see also HERO WORSHIP).

Importance of kinship

Basic patterns and functions. The presence or absence of ancestor worship relates in a general way to the importance of kinship in the societies concerned. In the primitive world society is ordered and life made possible through bonds of kinship, though intergenerational continuity of kinship extending to deceased forebears may not be regarded as important. Among societies of higher levels of cultural complexity, the importance of kinship and the size of actively functioning kin groups decreases. Where continuity of kinship and inheritance of property are very important, elders are characteristically regarded with respect, and the persistence of bonds of affinity with ancestors is favoured, as was the case in traditional China and Japan. Societies in which the only important kin group is the nuclear family composed of parents and immature children, and where economic support as well as emotional well-being among adults do not depend upon kinship, are ill-suited for the development or maintenance of ancestor worship. Illustrative examples are modern China and Japan, where sociocultural changes, brought about by adaptation to Western civilization and modern technology, have included a great decline in the importance of kinship and the size of kin groups, and where traditional practices of ancestor worship have correspondingly declined.

Ancestor worship includes all of the attitudes and acts usually associated with the worship of nonancestral gods and spirits. According to some scholars and theorists, ancestral spirits are anthropocentric conceptions similar to other supernatural beings; that is, the spirits have the qualities of personality and the capabilities of man, to which supernatural potency is added. The spirits see, hear, feel, understand, and communicate with the living; they make moral judgments; they are wishful, willful, joyful, angry, stern, permissive, kind, cruel, and sometimes capricious; and they have all the other emotions and traits of human beings. All of the behaviour and practices that are customary with regard to other kinds of supernatural beings are found in rites of ancestral worship—veneration and propitiation in the forms of prayers, offerings, sacrifices, the maintenance of moral standards, and festivals of honour that may include pageantry, music, dance, and other forms of art. Where ancestral spirits directly control the affairs of the living, their continued favour is sought by established periodic rites, and their special aid

may be requested at times of crisis. Perhaps the only truly distinctive ritual acts of ancestor worship are commemorative ceremonies, held annually or at other fixed intervals, and tendance of graves, monuments, or other symbols commemorating them (see also DEATH RITES AND CUSTOMS).

Motives for acts of piety toward ancestors are diverse, and they differ from devotional acts toward other gods or spirits principally in reflecting the idea that the spirits continue in some measure to be kin and are active participants in the life of the community. Rituals directed toward ancestral spirits maintain communion with them in ways that reflect human regard for the deceased elders and desires to aid them in their spiritual existence. These rites and devotional acts also seek to gain spiritual and practical benefits for the living. The powers the ancestral spirits are believed to possess vary greatly from society to society, as do the powers of other supernatural beings. Their powers may be weak or strong, generalized or specific. In some societies, their supernaturalistic roles include that of being intermediaries between living relatives and the gods. Where neglected ancestral spirits are thought to be harmful to the living, the goals of ritual observances may include or emphasize the desire for protection from them. Whether ancestral spirits are themselves gods with powers or are intermediaries, communion with them is a form of transcendence of ordinary states of existence, which may be a conscious or unconscious goal of the acts of devotion.

Ancestor worship in religions of the world. Until the 19th and 20th centuries ancestor worship in various forms and of varying importance in total religious complexes was widely but irregularly distributed throughout the world. In most societies, however, it was only one element of a complex of supernaturalism, and seldom a dominant feature. The spread of European culture weakened, displaced, or otherwise put an end to ancestor worship in most primitive societies, and technological, social, and ideological changes discouraged its continuation in culturally advanced societies.

In primitive societies. Among primitive societies, well-developed ancestor cults are limited principally to peoples of sub-Saharan Africa, Melanesia, and some tribal groups of India and adjacent parts of Asia. The greatest development was in Africa, where ancestral spirits are commonly an important part of the roster of supernatural beings. In the aboriginal kingdoms and near-kingdoms of sub-Saharan Africa, the spirits of kings and paramount chiefs were often regarded as generalized ancestors and were venerated by all members of society. Spirits of the heads of clans, often a mythical couple, were also worshipped, as were the spirits of founders of lineages and of deceased heads of individual families. Ancestral spirits of kings and high chiefs often were believed to have power over matters of concern to the entire society, such as rain and the growth of crops and cattle, whereas spirits of heads of families, lineages, and clans influenced matters of immediate concern to the particular social groups. Acts of piety were numerous and included sacrifice, prayers, and hospitable celebrations that honoured the spirits by story-telling and other forms of entertainment. The spirits were generally regarded as very helpful to their living descendants and were propitiated in established cyclic ceremonies as well as at times of crisis when help was needed.

Ancestral spirits in Africa and Melanesia

In Melanesia the spirits of the dead generally were held to be important, and in some societies were the focus of much attention. An outstanding example is the ancestor cult of the Manus of the Bismarck Archipelago, where Sir Ghost, the spirit of the living male head of the household, was the tutelary god of the family and supervised the behaviour of its members. Only spirits of the newly dead were worshipped, and when the head of a household died, the old tutelary god was discarded. The skull of the deceased household head was placed above the entrance in the dwelling, where it watched the conduct of all within, giving rewards and punishments in accordance with their deeds, and protected the family from the malign influences of the guardian spirits of other families.

Among the Trobriand Islanders near New Guinea, the dead had two spirits, one of which was a harmless ghost that vanished a few days after death. The other spirit, *baloma*, had an eternal existence as an ancestral spirit abiding in another world. Death rites were clearly distinguishable from ceremonies directed to the *baloma*. Communion was maintained with these ancestral spirits, who returned to their villages when annual feasts were held by their living relatives, and also appeared in dreams and trances. Seers with special powers of communication with the supernatural world brought news of the spirits by visiting them in person in the land of the dead. The *baloma* were propitiated but were never feared. They were invoked and believed to provide supernatural aid in numerous acts of magic, especially in connection with the raising of the crops upon which the livelihood of the people depended.

Elsewhere in primitive societies, ancestral spirits sometimes were important, but nowhere were they the lone or primary supernatural beings. In aboriginal Polynesia, where people of high social status were regarded as descendants of the gods, the spirits of kings and high chiefs had power to help men, but they were not the objects of worship to any great extent. Indian tribes of North and South America seldom gave much ritual attention to ancestors. Among tribes of the present-day United States, for example, the greatest ritual elaboration was among the Pueblo tribes of the Southwest, whose complex ceremonial calendar included impressive rites honouring generalized ancestral spirits. These ancestors were impersonated in ritual and appealed to for aid in providing rain, bountiful crops, and general well-being.

In Eastern societies. Among the civilizations of Asia, the classic examples of ancestor worship have been China and Japan. In both societies, however, reverence for, rather than worship of, ancestors is a more nearly accurate description of the beliefs and practices. In China the ancestor cult is extremely ancient and emphasized continuity of familial lines. Reverence for elders was an act of filial piety strongly supported by the teachings of the sage Confucius of the 6th–5th centuries BC. The family was viewed as a closely united group of living and dead relatives rather than a group of individuals. Unity of the larger kin group was also stressed through devotional acts at clan temples that honoured all ancestral spirits. Rites of reverence were held in the home, at temples, and in graveyards. Ancestral shrines containing tablets bearing the names of recent ancestors and especially notable forebears were maintained in the homes, and rites were observed before them. Temple rites were also observed; funerals and commemorative ceremonies were elaborate, and custom also called for pilgrimages to graves. Motives for performing rites involved concern for the welfare of the ancestors, who were thought to require solicitous care and, since the kin group was an indivisible unit with common goals and fortunes, a means of continuing to receive the aid and cooperation of the deceased relatives.

Ancestor worship in China and Japan

The early background of Japanese ancestral veneration is obscure, but most of the historically known practices are adaptations of Chinese customs. Some sort of ancestor cult may have existed in the native Shintō religion before the diffusion of Buddhism from China in the 6th century. When Buddhism came to Japan it was a comprehensive religious system. With the passage of time and in coexistence with the Shintō religion, Japanese Buddhism began to emphasize death rites and commemorative ceremonies, and Shintō became more concerned with matters of daily life. Confucianism was never an organized religion in Japan, but quasi-religious Confucian ideals of filial piety were very important and were sometimes incorporated in the teachings of Japanese Buddhist sects, thereby reinforcing ancestral veneration. Japanese rites, like those of China, consisted of elaborate funerals and many commemorative rites at the home, the temple, and the grave. A great annual ceremony honours all spirits of the dead, who return to their homes at that time. Until recent years Shintō rites of passage at death also were conducted in the home.

The state of ancestor worship in modern China is un-

clear, but it may be disappearing. In modern Japan ancestors have declined in importance, and Buddhist ritual tends to emphasize funerals, giving less attention than formerly to later commemorative ceremonies.

Ancestor worship in India

In India the vast, locally variable, and unorganized complex of theology and rites of the Hindu religion conspicuously includes ceremonies honouring ancestors, but the cult of ancestors is nevertheless a relatively small part of the full religious system. Characteristically, funeral rites are very elaborate and have many motifs of supernaturalism, among which is attention to ancestral spirits. The practices relating to ancestors reflect ideas concerning reincarnation and the system of castes, which are in turn closely intertwined. One of the manifest goals of the funeral rites is to guide the spirits of the deceased during the perilous time between death and rebirth. The idea that one's moral behaviour in life determines his fate in the next life is also connected with caste affiliation. Sinful behaviour will bring rebirth in a low caste. Among some castes and in some regions of India, annual rites are performed in honour of the spirits of the ancestors of the male heads of household, who are believed to give spiritual aid in promoting the growth of crops and in other important matters of everyday life.

In ancient Near Eastern and European societies. Ancestor worship in various forms existed among the ancient civilization of the Mediterranean, where cults of the dead sometimes also existed, and among later European peoples. Ancient Egyptian religion featured a cult of the dead but gave little attention to ancestral spirits except to those of royalty, which were venerated by the people and especially honoured in rites observed by their royal descendants. In ancient Babylonia a cult of the dead also existed, and among the members of the ruling class ancestral spirits were honoured by festive rites and sometimes deified. Beliefs and practices of late Zoroastrianism (a religion founded by the 6th–7th century BC Iranian prophet Zoroaster) included rites for the spirits of the dead, who were believed to have power in the affairs of the living.

Ancestor worship in Greece, Rome, and other ancient European societies

In ancient Greece ancestor worship overlapped with hero worship. Some ritual attention was given to spirits of household heads and political leaders, and the spirits of men whose deeds were heroic were sometimes elevated to immortality and made the objects of rites of reverence. In ancient Rome ancestor worship was a familial cult activity. Ancestral spirits were believed to have influence on mortal life and to return to visit their relatives, when rites were held in their honour.

Among various northern and eastern peoples of Europe, ancestral spirits also held some importance. Ancient Celts, Teutons, Vikings, and Slavic groups conducted rites of propitiation and sacrifice.

Theories and interpretations. The 19th-century sociologist Herbert Spencer regarded fear and consequent propitiation of the souls of ancestors as the earliest form of religion, an interpretation that later scholars set aside as unverifiable. Reflecting the decline of ancestor worship among societies of the world, modern scholarship has seldom given much concern to this subject in isolation but has instead followed the trend of the social sciences in considering ancestor cults in relation to other elements of religious complexes, the social order, and the whole of culture. Early writings often expressed the idea that all people fear death, but this idea is questioned by Hindu thought, in which extinction rather than eternal life is the ultimate goal. It is generally held that all peoples have some beliefs of an afterlife. Such beliefs are presumed by anthropologists to be very ancient, as seems evident from Paleolithic burials dating from perhaps as much as 60,000 years ago containing stone tools, shells, and other objects, and the abundance of later prehistoric burials with grave goods.

Modern interpretations of ancestor worship

No recent study of wide comparative scope attempts to interpret the significance of ancestor worship, and modern interpretations of these practices view them as having essentially the same social and psychological value as other beliefs and practices of supernaturalism. Through their symbolic representations of kinship and of the social hierarchy of kin groups, the beliefs and acts of ancestor worship may be seen as establishing and reinforcing ideas of social roles and identities, thereby contributing to psychological well-being and social harmony. Joint rites promote social solidarity, and characteristically display and thereby reinforce the social order. Where ancestral spirits have power in mortal affairs, they are psychologically and socially significant in ways similar to those of gods and other supernatural beings. In the many matters of life over which man lacks secular control, for example, the intervention of ancestral spirits alleviates anxiety. But, quite like various other beliefs of supernaturalism, ideas about ancestors may also be seen sometimes to instill as well as to allay anxiety. In this connection ancestor cults may have an important moral significance by serving as sanctions of social conformity. In many societies, improper behaviour is usually thought to reflect unfavourably upon both living and dead relatives, whether or not ancestor cults exist. Scholars often interpret the primary significance of ancestor cults in some societies as a sanctioning force. Ancestral spirits are viewed as approving or disapproving the behaviour of their descendants, in a generalized way, without exercising specific sanctions, or they may be regarded as vigilant protectors of morality. Among the Manus of Melanesia, the ancestral tutelary spirit was believed to punish all moral defections by removing the "soul stuff" from the wrongdoer, thereby causing illness. When the offense was serious, death followed unless penitential acts were performed. Where, as among the Manus, ancestral wrath might affect any member of the group for an offense committed by another member, the sanctioning force is viewed as powerful, operating so that all living members monitor the behaviour of each other.

Special value is attributed to the rites of ancestral reverence in promoting familial solidarity, and, to the extent that such rites are emphasized, in promoting the unity of larger kin groups and entire societies. Information is most abundant on traditional practices of familial ancestor worship in China and Japan. In these societies in former times the individual was submerged in the family, and rites of ancestral reverence may thus be viewed as both reflections and reinforcements of the social order. As acts of supernaturalism, they place a special stamp of approval upon familial roles, unity, and continuity.

BIBLIOGRAPHY. J.T. ADDISON, *Chinese Ancestor Worship* (1925), descriptive information on traditional practices in China; E. BENDANN, *Death Customs* (1930, reprinted 1969), descriptive material useful in distinguishing between death cults and ancestor worship; R.F. FORTUNE, *Manus Religion* (1935, reprinted 1965), a detailed account, including information on the role of ancestor worship as a sanction for ethical behaviour; J.G. FRAZER, *The Belief in Immortality and Worship of the Dead*, 3 vol. (1913–24, reprinted 1968), a classic work presenting extensive information on its subject among societies of the world; J.R. GOODY, *Death, Property and the Ancestors* (1962), a detailed account of mortuary customs of the Lodagaa of West Africa; B. MALINOWSKI, "Baloma, the Spirits of the Dead in the Trobriand Islands," in *Magic, Science and Religion and Other Essays* (1948).

(E.N.)

Ancient Epigraphic Remains

In the context of this article ancient epigraphic remains are the traces of ancient languages that can be recovered from inscriptions on ancient ruins and durable objects such as coins. The article is confined to languages of Europe and the Mediterranean (including Asia Minor and the Near East) that are not fully covered elsewhere in the language articles. Related articles are ETRUSCAN LANGUAGE; ITALIC LANGUAGES; GREEK LANGUAGE; ANATOLIAN LANGUAGES; INDO-EUROPEAN LANGUAGES; HAMITO-SEMITIC LANGUAGES; WRITING, FORMS OF; ALPHABETS; HIEROGLYPHIC WRITING.

Epigraphic remains in general. The epigraphic remains of languages with literary traditions are important, because the exact form in which the inscriptions were written or engraved is preserved; in other words, epigraphic remains are not subject to the changes that take place in the course of centuries of copying and recopying literary texts. Furthermore, some inscriptions contain the

language of the lower classes, which normally is not preserved otherwise. Even the inscriptions do not always represent the actual state of the language in a particular period, however, for official inscriptions—treaties, laws, public announcements—tend to preserve obsolete inflections, forms, syntax, and style; and the vulgar inscriptions—*e.g.*, graffiti—often borrow from the language of the upper classes.

Epigraphic remains of a language without literary tradition represent the only primary source of knowledge concerning the language. If the script is already known, the problem is one of interpreting the text. Etruscan, for example, is in an alphabet derived from the Greek and can be read. If both the script and the language are unknown, some key for the decipherment has to be found. The text may be bilingual, and this may provide a clue. If it is not, the first step is usually to equate a special group of signs to the name of a person or a place known from other sources. If the guess is correct, the approximate value of a small number of signs is gained. The circumstances in which the text is found may suggest the meaning of single words, groups of words, and whole phrases. The recognition of relationships with other languages can lead to further progress in decipherment. This makes it possible to take elements, especially inflectional and structural elements, of the related languages and to compare them with categories still unexplained in the language being studied. A decipherment will prove successful only if a certain number of texts is preserved, because it is necessary to check the assumed meanings of words and the assumed functions of endings and suffixes in different contexts. Otherwise the arbitrariness and the precariousness of the assumptions remain too great, so that further statements on the language are not justified.

Methods of decipherment

Asia Minor and the Near East. *Phrygian.* Old Phrygian is known from a small number of rock inscriptions in a script derived from the Greek alphabet. The inscriptions date from the 6th and 5th centuries BC. Late Phrygian is found in approximately 100 rock inscriptions in the Greek alphabet from Roman times (2nd to 4th century AD). The language is Indo-European but has no special connection with the Hittite-Luwian group of languages, also spoken in Asia Minor.

Carian. Only a few inscriptions, of about the 4th century BC, were found in Caria, in southwest Asia Minor. The majority of the Carian texts consist of graffiti carved by Carian soldiers on various objects in Egypt; these inscriptions date from the 8th or 7th to the 5th century BC. They are written in a syllabic script that has not yet been completely deciphered. The language probably belongs to the Hittite-Luwian group.

The language of Side in Pamphylia. The language of Side in Pamphylia (on the southern coast of Asia Minor), often called Sidetic, is attested on a few short inscriptions on stone, including two bilingual inscriptions from the 3rd century BC, and on coins from the 5th and 4th centuries BC. The alphabet may be derived from the Greek; the language seems to be Indo-European.

Pisidian. The Pisidian language, which was spoken in Pisidia, directly north of Pamphylia, is represented by a dozen inscriptions in Greek characters from Roman times. From its features it could belong to the Hittite-Luwian group.

Crete. Most of the objects on which inscriptions occur are of clay. The syllabic writing system of Crete, which shows a more or less continuous development from the hieroglyphic to the linear pattern, is referred to as the Minoan or the Aegean writing system.

Hieroglyphs. Hieroglyphs, the oldest form of the Cretan scripts, date from 1900–1600 BC and have not yet been deciphered. The inscriptions are impressions of seals on clay tablets. The later form, found on clay amulets and other clay objects, is protolinear; that is, the signs are simplified, and there is a tendency toward a linear script.

Linear A. The older of two linear scripts, Linear A is a further simplification of the protolinear script and is found in more than 300 inscriptions (150 clay tablets with Linear A script are from Ayía Triáda [Hagia Triada], near the south central coast of the island), written between 1600 and 1450 BC.

Linear B. Clay tablets containing inscriptions in Linear B script have been found on Crete and on the Greek mainland. The inscriptions represent the earliest surviving examples of the Greek language.

Other writing systems. Other writing systems associated with Crete are attested by only a single document or a limited number of documents. The so-called disk of Phaistos, for example, has on both of its sides a series of a sort of hieroglyph arranged in a spiral line.

Other sites of epigraphic remains. *Cyprus.* The Cypriot inscriptions are written in two scripts, the Cypro-Minoan script, derived from the Minoan writing system of Crete and not yet deciphered, and the Cypriote syllabic script. The Cypro-Minoan script is found on clay balls and tablets, on copper bars, and on other objects dating from the period 1500–1050 BC. The Cypriote syllabic script, found on stones, vessels, and tools and dating from the 7th to the 3rd century BC, was used for two languages, the so-called Eteocypriot language and the Greek dialect of Cyprus. Greek invaders adopted this script from the older, non-Indo-European inhabitants.

Cypro-Minoan and Cypriote syllabic scripts

Balkan peninsula. Only a few ancient documents from the Balkan peninsula have survived. The inscriptions on two rings, from Ezerovo and Kolmen in Bulgaria, are the only directly preserved specimens of the Thracian language. The inscriptions are in the Greek alphabet and date from the 5th or 4th century BC. The Dacian language was used in an inscription on a clay vessel found near Dialul Grîdiştrei, in Romania. The inscription, of unknown date, is written in the Roman alphabet.

Ancient Italy. The vast majority of inscriptions found in southern Italy are in Greek, Etruscan, Latin, and other languages dealt with in articles cited in the opening paragraph of this article.

The Messapic language, an Indo-European tongue closely related to Illyrian (traces of which are found on the east side of the Adriatic), was used in stone inscriptions found in Apulia and Calabria provinces on the southeast coast of Italy. The inscriptions, in an alphabet derived from the Greek alphabet of Tarentum, date from the 6th to the 1st century BC.

Among the wide variety of inscriptions discovered in Sicily, that on a Centuripe vase of the pre-Roman period is the best known. The language of the inscription, called Sicel, is apparently Indo-European.

In ancient Liguria, in northwest Italy, a number of pre-Roman stone inscriptions are in a language of Indo-European origin called Ligurian.

In ancient Raetia, northeast of the Italian peninsula, inscriptions on various objects are also from pre-Roman times. The language of the inscriptions, Raetian, is definitely non-Indo-European.

France and northern Europe. In the territory of ancient Gaul, now occupied by France, about 60 stone inscriptions in the Celtic language known as Gaulish were found. A few inscriptions in this language were also discovered in northern Italy, dating from the 3rd century BC to the 3rd century AD. For the older inscriptions, the Greek alphabet was used; for the more recent, the Roman characters of the imperial period. About 60 graffiti of about 50–60 AD from La Graufenseque, with lists and accounts of potters, also cast light on the Celtic language of ancient Gaul. Celtic was also used in more than 300 inscriptions in the so-called ogham alphabet. Most of the ogham inscriptions were found in southern Ireland, but a few texts of particular importance are situated in Wales and in adjacent territories. The inscriptions date from the 4th to the 6th century AD and thus are older than the oldest literary tradition for Celtic.

Iberian peninsula. The Iberian peninsula is the territory now occupied by Spain and Portugal. More than 500 inscriptions on stone and also on such objects as vases and lead tablets have been found. Dating from the 3rd to the 1st century BC, most of the inscriptions are in either of two languages: the non-Indo-European Iberian language of the coast and the Pyrenees or a Celtic language of the inland (Celto-Iberian). The characters in the script

Iberian and Celto-Iberian inscriptions

used—called the Iberic script—contain Greek and Phoenician elements, with two classes of signs: (1) signs standing for whole syllables and (2) signs denoting a single vowel or consonant.

BIBLIOGRAPHY. ERNST DOBLHOFER, *Zeichen und Wunder: Die Entzifferung verschollener Schriften und Sprachen* (1957; Eng. trans., *Voices in Stone: The Decipherment of Ancient Scripts and Writings*, 1961); and JOHN CHADWICK, *The Decipherment of Linear B* (1958), deal with methods and techniques to be adopted in decipherment. VLADIMIR I. GEORGIEV, *Introduzione alla storia delle lingue indeuropee*, 2nd ed. (1966), is very important for the areas of Crete, Cyprus, Asia Minor, and the Balkan Peninsula because of its rich bibliography; the interpretation of the documents, however, has to be read with caution because there is a tendency to find Indo-European idioms everywhere. The various languages, groups of languages, or groups of epigraphic remains are treated in the following books (containing texts in transcription and translation with commentary and a sketch of grammar, if not otherwise stated): GUNTER NEUMANN, "Lykisch," and ALFRED HEUBECK, "Lydisch," in *Handbuch der Orientalistik*, vol. 2, *Altkleinasiatische Sprachen* (1969); P.H.J. HOUWINK TEN CATE, *The Luwian Population Groups of Lycia and Cilicia Aspera During the Hellenistic Period* (1961); ROBERTO GUSMANI, *Lydisches Wörterbuch* (1964); OTTO HASS, *Die phrygischen Sprachdenkmäler* (1966); R.D. BARNETT, *Phrygia and the Peoples of Anatolia in the Iron Age* (1967), gives an excellent survey, but includes no texts and no linguistic interpretation; STERLING DOW and JOHN CHADWICK, *The Linear Scripts and the Tablets As Historical Documents* (1971), also no texts and no linguistic interpretation; JOSEPH WIESNER, *Die Thraker* (1963), deals with the history of the peoples of the Balkan Peninsula—has no texts and no linguistic interpretation; FRITZ VON LOCHNER-HUTTENBACH, "Illyrier und Illyrisch," in *Das Altertum*, 16:216–228 (1970), gives a survey (documents in the Illyrian language do not exist, the only inscription hitherto attributed to Illyrian is actually Greek); R.S. CONWAY, S.E. JOHNSON, and JOSHUA WHATMOUGH (eds.), *The Prae-Italic Dialects of Italy*, 3 vol. (1934); VITTORE PISANI, *Le lingue dell'Italia antica oltre il latino*, 2nd ed. rev. (1964); OTTO HAAS, *Messapische Studien* (1962); ULRICH SCHMOLL, *Die vorgriechischen Sprachen Siziliens* (1958); GEORGES DOTTIN, *La Langue gauloise* (1918); ANTONIO TOVAR, *The Ancient Languages of Spain and Portugal* (1961).

(He.Mi.)

Andaman Sea

The Andaman Sea is part of the northeastern Indian Ocean and covers an area of 308,000 square miles (798,000 square kilometres). Through such ports as Bassein, Moulmein, Tavoy, and Mergui—as well as Rangoon itself—it forms Burma's most important sea link with other nations; it also forms part of a shipping route between India and China, via the Strait of Malacca. It is bounded to the north by the Irrawaddy Delta of Burma; to the east by peninsular Burma, Thailand, and Malaysia; to the west by the Andaman and Nicobar Islands, which are under Indian administration; and to the south by the island of Sumatra (part of Indonesia) and by the Strait of Malacca. The sea takes its name from the Andaman Islands. It is 750 miles long, from north to south, and 400 miles wide. Less than 5 percent of the sea is deeper than 10,000 feet, but in a system of submarine valleys east of the Andaman–Nicobar Ridge, depths exceed 14,500 feet.

Trading vessels have plied the Andaman Sea since ancient times. It formed part of the early coastal trade route between India and China and from the 8th century onward formed a link in a thriving trade between India and Sri Lanka (Ceylon) to the west and the Burmese ports of Thaton, Martaban, and Tavoy to the east.

The sea's northern and eastern third is less than 600 feet deep. The Irrawaddy River is continually forming a delta shelf of silty clay, incised by a shallow channel system that merges into the submarine Martaban Canyon to the south. During the Tertiary Period (70,000,000 to 10,000,000 years ago), a continental shelf was formed adjacent to the Malay Peninsula. Subsequent faulting and folding molded parts of this shelf into two submarine terraces and a shelf basin north of Sumatra. A discontinuous veneer of shelly quartz sands covers the continental margin, while the shelf basin has accumulated over 3,000 feet of silty clays. The continental slope is 500 miles long with a relief variation of 6,000 feet.

Geology. The Andaman–Nicobar Ridge has a core of serpentinite (primarily hydrous magnesium silicate) and associated rocks, overlain by at least 10,000 feet of folded and faulted sandstone and shale. The western and central half of the sea, between the ridge and the Malay Continental Slope, is 3,000 to 10,000 feet deep, and consists of a broad central trough, two seamounts, a system of smaller seamounts and rift valleys, and the volcanic Narcondam and Barren islands. This basin's basaltic rock is partially and thinly covered by Irrawaddy clays. This silt blanket is thin because most Tertiary sediment was probably shed into an adjacent trough that was then molded into the Andaman–Nicobar Ridge. The basin was produced approximately 10,000,000 years ago when the Andaman–Nicobar Ridge was rifted from the shelf in a movement that continues today.

Climate and water characteristics. Southeast Asia's monsoonal regime governs the sea's climate and waters. In the winter, light northeasterly winds cause south to southeasterly Andaman currents to flow into the Bay of Bengal. Humidity is low, the sea receives little rainfall or runoff, and surface salinities are high. In May southwest monsoon winds cause westerly currents. Huge volumes of runoff water from Burma flow into the sea during the summer monsoon, forming a marked pattern of low surface salinity in its northern third.

Andaman water—categorized as Equatorial India Water—is stratified in three layers. A surface layer 200 to 300 feet thick, with temperatures between 75° and 84° F (24°–29° C) and seasonally variable salinities, is separated by a strong thermocline (layer of water in which temperatures drop more than 1° C per metre of depth) from a 1,500-foot-thick intermediate layer, with a maximum salinity of 35 parts of salt per 1,000 parts of seawater. Down to 4,000 feet the temperature potential is 40.5° F (4.7° C), and salinity is 34.86 parts of salt per 1,000 parts of seawater. Deeper waters, replenished by waters from the Bay of Bengal, maintain these temperatures and salinity levels down to the bottom.

Economic resources. The plankton biomass in the surface waters is relatively low. Bottom fauna is extremely poor, while coral reefs occur only along the Andaman and Nicobar Islands. The Malay Shelf waters, however, favour molluscan growth, and there are about 250 edible species of fish—20 of which are common—in the intensively fished coastal waters.

Mineral resources are limited. Southern extensions of the Burmese oil fields may underlie the Irrawaddy Shelf, and petroleum may also occur in the shelf-basin off Sumatra. Submerged extensions of the tin and ilmenite deposits of the Malayan coastal plains are intensively sought after, while tin dredging close to the shore has begun off the coast of Thailand.

The two largest ports are Pinang (Malaysia) to the southeast and Rangoon (in Burma) to the north, and most maritime traffic is along the coast of the Malay Peninsula. Westbound shipping enters the sea through the Strait of Malacca and exits through the Great Passage to Colombo or the Ten Degree Channel to Madras.

BIBLIOGRAPHY. On the physiography, submarine morphology, geology, and geologic history, see K.S. RODOLFO, "Bathymetry and Marine Geology of the Andaman Basin, and Tectonic Implications for Southeast Asia," *Bull. Geol. Soc. Am.*, 80:1203–1230 (1969); on bottom deposits, "Sediments of the Andaman Basin, Northeastern Indian Ocean," *Mar. Geol.*, 7:371–402 (1969).

(K.S.R.)

Andean Civilization, History of

The term Andean civilization refers to the aboriginal American Indian cultures that developed and flourished in western South America from several thousand years BC to their subjection by the Spanish conquerors in the 16th century AD. In its basic economic elements, it persists today in the large Indian populations of Peru, Bolivia, Ecuador, and Colombia. Though termed Andean, it also includes, especially in its early phases, the peoples and cultures of the western coast. The article is divided into the following sections:



Principal sites associated with Andean civilization.

I. Pre-Inca civilization
- The Late Preceramic
- The Initial Period and the Early Horizon
  - The Initial Period
  - The Early Horizon
- The Early Intermediate Period
  - The south coast
  - The north coast
  - The north highlands
  - The south highlands
- The Middle Horizon
- The Late Intermediate Period and the Chimú Empire
  - The Late Intermediate Period
  - The Chimú Empire

II. Inca civilization
- The ecological background of Inca expansion
- The origins and expansion of the Inca state
  - Settlement in the Cuzco Valley
  - The beginnings of external expansion
  - Pachacuti Inca Yupanqui
  - Topa Inca Yupanqui
  - Huayna Capac
  - Civil war on the eve of the Spanish Conquest
  - The Spanish Conquest
- Inca culture at the time of the conquest
  - Social and political structure
  - Administration of the empire
  - The prevention of rebellion
  - Inca technology

I. Pre-Inca civilization

The Indian peoples of the Andes never developed a written language. The Spaniards recorded oral traditions of the Incas, but for pre-Incan peoples the record is archaeological, and archaeologists, by studying the sequence and distribution of artifacts, have mapped the progress of Andean civilization from the 1st millennium BC to the complex empires preceding Spanish Conquest.

Peru must have been occupied by closely related peoples throughout the succession, but for much of it different areas developed with a considerable degree of independence. During three periods, called the Early, Middle, and Late Horizons, wide areas were united, and this is demonstrated by the existence of widespread art styles. In the case of the Middle (Huari) and Late (Inca) Horizons this is attributable to military conquest.

THE LATE PRECERAMIC

The Late Preceramic in Peru is considered to begin about 3500 BC, at which time there was considerable diversity in the occupation of the coast. In the centre and north lived people who cultivated beans, squash, cotton, and chili peppers and exploited the sea, catching fish with cotton nets and shell or composite hooks, collecting shellfish, and hunting sea mammals. One group at Chilca, south of Lima, built conical huts of cane thatched with sedge. The dead were buried wrapped in twined-sedge mats and guanaco skins. Some people camped in winter on the *lomas*, patches of vegetation outside the valleys that were watered at that season by fogs. In summer, when the *lomas* dried up, they built camps along the shore. The *lomas* provided wild seeds, tubers, and large snails; deer, camelids (probably guanaco), owls, and foxes were hunted. The *lomas* had long been shrinking, and the winter camps were abandoned (*c.* 2500 BC) in favour of

permanent fishing villages. Nowhere are the deposits thick enough to show stratification, but they have been arranged in chronological order by comparing the implement types and noting their distribution within the shrinking patches of vegetation. Some small patches still survive.

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

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

Little is known of the highland Late Preceramic. The caves at Lauricocha at about 13,000 feet in the central Andes, which had been occupied by deer and camelid hunters since nearly 8000 BC, were still used, at least as summer camps, by hunters who employed small leaf-shaped points. Gourds, squash, cotton, and lúcuma, with seed plants such as quinoa and amaranth, were cultivated in the Ayacucho Basin before 3000 BC; maize (corn) and beans came within the next millennium. There are also ground stone implements for milling seeds. It is claimed that llamas and guinea pigs were domesticated long before.

After about 2500 BC came a great increase in the speed of development, which is best known on the coast. Population increased, and stable settlements were established in many places. By 2000 BC there were perhaps 100 villages on the coast with populations of 50 to 500 people, with a few of up to 1,000, indicating a total population of about 50,000. This was a far cry from the thinly scattered bands and occasional villages of, say, 1,000 years before. There are considerable variations from place to place, but most sites show a predominance of seafood, namely fish, shellfish, sea lions, and sea birds.

Population growth

On the north central coast, the stretch between the Casma and Huarmey rivers was heavily populated. One site, at Culebras, is a large village on a terraced hillside, with semi-subterranean houses whose underground parts are lined with basalt blocks and whose upper parts are built of lighter materials such as adobe blocks. They originally had hard clay floors, and some had a guinea-pig hutch consisting of a stone-lined tunnel connecting two rooms at floor level. The guinea pig, normally vegetarian, appears to have been taught to feed on small fish. A site at Huarmey provides the earliest known instance of maize on the coast, and it also occurs in the top Preceramic levels at Culebras. Burials at Culebras were tightly flexed and wrapped in twined mats and cotton cloth and accompanied by gourd vessels and beads and pendants of stone, shell, or bone. The skulls of these people were deformed by having been bound to cradleboards in infancy. There was a cemetery, but many burials were under house floors. No ceremonial buildings are yet known in this area.

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

Ceremonial architecture

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

Apart from one site, Kotosh, near Huánuco, little is known of the highland final Preceramic. Here there are structures of undressed stone chosen to present flat wall surfaces, set in mud, covering an area of at least 600 by 300 feet, in some parts of which is a succession of buildings piled up to a considerable height. Among these are two superimposed temples, the lower being a rectangular structure on a stepped platform about 26 feet high. The floor was surrounded by a broad, low bench, and each outside wall has two or three rectangular niches. The walls and floor were covered with two coats of mud plaster, and beneath the central niche at one end was a pair of crossed forearms modelled in the same material. This temple was later buried in boulders surrounded by a retaining wall and covered by a new floor on which a second temple was built, of which little remains. The burial of the first temple to act as a raised foundation for the second recalls the construction at Río Seco.

THE INITIAL PERIOD AND THE EARLY HORIZON

The Initial Period. The next period, called the Initial Period, begins with the introduction of pottery. The pottery does not appear at one time at all the known sites, but the earliest examples have radioactive carbon dates of about 1800 BC, although the U.S. anthropologist John H. Rowe, in the most recent attempt to construct a consistent dating system, considers 2100 BC more reasonable. These examples occur on the central coast between Las Haldas, in the Casma-Huarmey region, and Lima; at Kotosh, in the central highlands; and at Yarinacocha, in the eastern lowlands. They are considerably later than the earliest pottery at Puerto Hormiga in Colombia (before 3000 BC) and Valdivia in Ecuador (*c.* 2700 BC). The period ends with the spread of the Chavín cult about 1000 BC (Rowe estimates 1400 BC).

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

Technological advances

New ceremonial centres show considerable diversity. Examples include La Florida, a huge pyramid in Lima that formed the nucleus of a yet unmapped building complex. The Tank site at Ancon consists of a series of stone-faced platforms on a hill. Las Haldas has a platform and three plazas; two smaller similar sites are also known. The old centres at El Paraíso and Río Seco had been abandoned, but, in the highlands, Kotosh continued to be occupied. Any constructions at Yarinacocha in a wet, stoneless area would have been of wood or other perishable materials and would have left no traces.

The variety of the pottery suggests that it was derived from several different sources. In the Lima area, which is better known than most, the earliest examples are neckless jars and incurved bowls with thickened rims and rounded bottoms, very uneven in shape. Later, some

types are pebble polished, and the jars become thinner. Other later types include bottles with straight spouts, which may have simple incised or applied decoration, and open bowls. Finally, as the period draws to a close, there are tan-coloured decorated wares, with punctate or red-painted areas outlined by incised lines, as well as orange ware with black stripes. A type found on the south coast is a small, double-spouted bottle with simple negative-painted decoration, the first appearance of a form long-lived in that area and of a decorative technique that later spread widely over the country.

In the highlands, the earliest pottery at Kotosh consists predominantly of simple bowls with somewhat constricted mouths, and bowls with gently rounded bases meeting the vertical to outsloping concave walls at a sharp angle. There are rare double-spout-and-bridge bottles, closed vessels with two tubular spouts connected by a solid bridge. The ware is mainly dark gray, black, or dark brown to sombre red, and it may have a red slip. The decoration, which is either applied to a broad zone covering most of the walls, or, on the neckless jars, forms a ring round the mouth, consists of linear incision, hatching, stamped circles, punctation, or excision. Postfired painting in red, yellow, and white frequently covers excised, hatched, and stamped areas. In spite of being across the eastern watershed, Kotosh has little in common with Yarinacocha save some similar decorations and the double-spout-and-bridge bottle, which is rare at both sites.

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

The Early Horizon. The Early Horizon is introduced by the appearance and rapid spread of the Chavín art style, ending the regional isolation of the Initial Period. It is the first of the three horizons that unite wide areas of the Central Andes at different times. It covers most of the north and centre of Peru, and its influence affected a good deal of the south coast, excluding only the southern highlands and the extreme south of the coast. The art style, which is regarded as the expression of a cult, is expressed in painted textiles (of which few have survived), in pottery, and chiefly in stone carvings. Archaeologists at one time generally agreed that the chief object of worship was a cat, probably the jaguar, but this has been questioned, although many natural bird, animal, and human forms have feline mouths and other attributes. Feline representations are widespread, whereas some unquestioned deities are confined to the immediate neighbourhood of Chavín.

Beginnings of unified art style

Most Chavín temples seem to have been ceremonial centres without people living around them, although that at Chavín itself seems to have been accompanied by a considerable town. The remainder appear to have been focuses for scattered settlements. The most elaborate temple known is that at Chavín, which lies at a height of 10,530 feet on a tributary of the Marañón, east of the Callejón de Huaylas. The temple consists of a group of stone platforms formed of rubble faced by walls of coursed masonry in which two thin courses alternate with one thick one. They are honeycombed with galleries running parallel to the walls at different levels and ventilated by shafts. The oldest part of the temple is a U-shaped structure, with the open top of the U facing east; the rectangular central arm contains a cruciform gallery, at the crossing of which stands a remarkable prismatic shaft of white granite, some 15 feet high, carved in low relief to represent a standing human figure with snakes typifying the hair and a pair of great fangs in the upper jaw. This figure, which has variously been called El Lanzón, the Great Image, and the Smiling God, has in its setting "an awe-inspiring quality which can be felt even by a present-day unbeliever" (Rowe), and it is thought to have been the chief object of worship in the original temple. The southern arm of the temple was subsequently twice widened by rectangular additions, into which some of the original galleries were prolonged. After the second addition, the two were joined by a free-standing facade having a central portal with a lintel supported on two cylindrical columns. The lintel bears 14 eagles in low relief, supplied with feline jaws with prominent fangs behind the beak, and each column is entirely covered by a mythical bird bristling with feline fangs and faces. These have been interpreted as attendants of the god worshipped in this part of the temple, who had perhaps superseded the Smiling God and could have been the god shown on the famous Raimondi Stone, now in Lima. This shows the Staff God, a standing semihuman figure having claws, a feline face with crossed fangs, and a staff in each hand. Above his head, occupying two-thirds of the stone, is a towering, pillar-like structure fringed with snakes and emerging from a double-fanged face, which Rowe interprets as a symbolic treatment of his hair as a tongue coming out of a mouth. Unlike the Smiling God, this figure occurs in areas as far from Chavín as the north and south coasts. Except for the columns, which are of black slate, the stones of the facade are light in colour on the south side and dark on the north. East of the facade is a small sunken court of the same period, which contained a number of slabs with carvings in low relief, and to the east of this is a much larger court surrounded by platforms. Within this court is a square, slightly sunken area, in which was found the Tello obelisk, a rectangular pillar carved in low relief to represent a caiman and covered with Chavín symbolic carvings, such as bands of teeth and animal heads. This is considered to be an object of worship like the Smiling God and Staff God. Carvings found on and around the temple include a cornice of projecting slabs, on the underside of which are carved jaguars, eagles, and snakes, and a number of tenoned heads of men and the Smiling God; they are thought to be decorations or the attendants of gods rather than objects of worship.

The temple at Chavín

Another Chavinoid highland temple is at Kuntar Wasi, or La Copa, in the upper part of the Jequetepeque Basin. It is a triple-terraced platform on the top of a hill, supporting a much damaged temple, with stone carvings of Chavín character—including three-dimensional statues, not found at Chavín, of men with feline features.

On the coast, the temples are built mostly of adobe. In the Nepeña Valley, two temples—Cerro Blanco and Punkurí—differ so much that they must also differ in age, but it is not known which is the earlier. Cerro Blanco is a massive platform of conical adobes and stones, supporting rooms with walls bearing Chavín decoration, including eyes and feline fangs, modelled in mud plaster in low relief and painted red and greenish yellow. Punkurí has a low, terraced platform with a wide stairway on which stands a feline head and paws, modelled from stone and mud, and painted. By the paws was buried a woman, believed to have been sacrificed. At Moxeke and Pallca in the Casma Valley to the south, there are terraced, stone-faced pyramids with stone stairways. The first has niches containing clay-plastered reliefs of mud, stone, and conical adobes, showing felines, snakes, and human beings of Chavinoid character painted in polychrome. Also in Casma is a temple at Cerro Sechín, consisting of a series of superimposed platforms with a central stair, on either side of which, at the bottom level, stands a row of irregularly shaped flat stones with incised designs showing standing men carrying clubs, severed heads, and other designs. Lacking Chavín characteristics, these have been interpreted variously as ancestral Chavín or derived from it, the latter being the

more plausible. There is a Chavín ceremonial centre at Garagay in the Lima Valley but none to the south.

Chavín pottery

Chavín pottery is best known from the decorated types found in the galleries in the temple at Chavín and in graves on the northern coast, where it is called Cupisnique. Until the end of the period, the ware is monochrome—dull red, brown, or gray—and is hard and stonelike. Vessels are massive and heavy, especially in the early part of the period. The main forms are open bowls with vertical or slightly expanding sides and flat or gently rounded bases, flasks, and stirrup-spouted bottles. The surface may be modelled in relief or decorated by incision, stamping, brushing, rouletting, or dentate rocker-stamping, all of which may be applied to particular zones in contrast with other smooth ones. Some bowls have deeply incised designs on both the inside and outside faces. Many of the forms and decorative features, apart from specifically Chavinoid designs (particularly feline fangs), were already present at Kotosh in the previous phase. There are considerable time changes in Chavín pottery; for example, the earliest stirrup spouts are relatively small, very thick and heavy, and the spout has a thick flange. As time goes on, the stirrups become lighter and the spouts longer; the flange is reduced and finally disappears. The necks of the flasks undergo similar changes. The decoration on some of these is extremely striking; one has incised flower designs, and another has a roughened surface in which there are a number of concave circular depressions with a notably high polish. The Cupisnique stirrup-spouted vessels, some of which are modelled in the form of human beings, animals, or fruits, are the beginning of a north-coast tradition of naturalistic modelling, which persists throughout its history. Toward the end of the period, a bichrome (dark red on cream) pottery comes into use.

There is a considerable area on the south coast with its focus in the Ica Valley, where strong influences from Chavín are found in the Paracas pottery style, and two painted textiles in pure Chavín style have survived from the same valley. Paracas pottery is very different from that of Chavín, but various motifs have enabled the two to be correlated closely. Paracas begins at practically the same time as Chavín, say about 1000 BC (Rowe says 1400), and lasts throughout its span and beyond it, perhaps to about 200 BC. The most characteristic form of Paracas pottery is a closed globular vessel with a somewhat flattened base, which has two narrow spouts connected by a flat bridge, or more frequently one spout is replaced by a human or bird head. Simple round-based bowls are very common. The ware is most commonly black or very dark brownish, and much of the surface is covered with decoration outlined by incision and painted in polychrome with hard, shiny, resinous colours after firing. A panel bearing a feline face on one end of a spout-and-bridge vessel is one of the most frequent forms of decoration. Paracas is also distinguished for its gorgeous embroidered textiles, generally found in the mummy bundles of the important dead. Embroidery had a popularity at this time that it afterward lost, but a surprisingly wide range of weaving techniques were also used in various parts of the coast.

Metalworking began in the Early Horizon. In the north, ornaments such as cuffs, crowns, and ear disks were made of thin-hammered sheet gold, with Chavinoid repoussé (relief) designs, and these are found particularly in the Lambayeque Valley. Welding and soldering were employed in the making of some of these. Objects of silver are occasionally found. In the south, copper ornaments and hooks for spear throwers are found in Paracas contexts, but information about their manufacture is contradictory, since one authority says that cold hammering and annealing were used, and another says casting. In both areas metal begins fairly late in the period.

A cemetery at Tembladera, in the Pacasmayo Valley, has yielded resin-painted polychrome pottery with specific resemblances to that of the end of Paracas far to the south. This pottery must mark a southern intrusion into the north, but its relation to the later stages of Chavín in that area is not known.

THE EARLY INTERMEDIATE PERIOD

The Early Horizon was succeeded by the Early Intermediate Period, formerly called Florescent or Classic, since "art and technology reached their peak" (Lanning) in the areas where it was best known, namely Moche in the north and Nazca (Nasca) on the south coast. The full range of pre-Columbian plants was cultivated, the most important addition to the list at this time being the potato, doubtless introduced from the highlands. Weaving with the belt loom continued to flourish, and llama and alpaca wool, as well as cotton, were everywhere used in a wide variety of processes, though they had been introduced earlier in the south. In metallurgy, all the pre-Columbian processes and alloys were in use, except bronze. Gold, silver, copper, and their alloys were used in making vessels and ornaments, such as crowns, collars, ear and nose ornaments, and false faces for mummy bundles. In the northern Moche culture, useful objects, such as mace-heads, lance points, and blades for digging sticks, were made of copper. Gold was obtained from placer deposits; that is, from superficial deposits of sand and gravel. There is no evidence about the sources of silver and copper, though copper, at least, was smelted from ores in later times. As in the past, hammering, embossing, and annealing were employed in fashioning metals. Metals were also cast in molds and by the cire-perdue (lost-wax) method, whereby a refractory mold is built up around a wax pattern and then heated to drain off the wax. Gold–copper alloys were gilded by leaching out the copper at the surface and burnishing; metals were sometimes welded together by heating and hammering.

Metallurgy

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

Nazca pottery

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

The north coast. On the north coast, the cultural peak already mentioned was reached in the Moche culture, but this appears to have been preceded by two ill-dated phases. The first, called Salinar, is best represented in the Chicama Valley, though it exists also in Moche. Like Cupisnique, it has elaborately modelled vessels, and the stirrup spout continues. There are also jars with a narrow neck and strap handle, as well as globular jars with a spout connected by a bridge to a modelled figure, which may hold a whistle. Human beings, birds, and animals are shown, but the modelling is less natural either than Cupisnique before or Moche afterward, and the

human figures are lively caricatures, sometimes involved in sexual activities. The pottery is oxidized, and its red surface is often decorated with simple white painting.

The second, called Gallinazo, developed chiefly in the Virú and Santa valleys and on a small scale in Moche and Chicama. The salient characteristic of the decorated pottery is negative painting, found on many forms including spout-and-bridge and double vessels, and simple stirrup-spouted jars. Many vessels have figure modelling less naturalistic than that of Moche.

The Moche culture

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

There were no towns in the northern valleys, and the general pattern of settlement seems to have been dispersed communities, built in places where they would not use valuable agricultural land and situated in such a way as to suggest dependence on one of the ceremonial centres.

The relationship between the Moche and Gallinazo peoples is not certain. Both peoples had pyramids, and the large Gallinazo rooms, or courts with decorated walls, may correspond to the Moche "palaces." Among dwelling types, irregular conglomerations of adobe rooms are found in both. The distribution of pottery phases indicates that Moche I and late Gallinazo were contemporary and that Gallinazo was displaced from the Virú Valley in Moche II, presumably by a conquest that extended to its southern limit in Nepeña in Moche III. The distribution of Moche V suggests a northward withdrawal to Chicama, perhaps under pressure from southern invaders (see below *Middle Horizon*). Since the chief difference between Gallinazo and Moche seems to be in their decorated pottery, they may have been closely related peoples.

Recuay pottery

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

The Cajamarca Basin is the site of a pottery style that is entirely independent of known outside influences and that spans at least the Early Intermediate Period and the Middle Horizon. It has lightly painted running-scroll designs, which vaguely recall writing (whence the name cursive), as well as small animals and faces, in brownish black or red on a cream background, mostly on open bowls with ring bases. It was traded widely in the north, and south as far as Huari, during the Middle Horizon. On the central coast, the Early Intermediate Period is not so well known as it is in the north and south, and its pottery styles are less spectacular. But a detailed study has enabled the old succession of White-on-Red (corresponding to Salinar and Gallinazo in the north) followed by Interlocking (corresponding to Moche) to be subdivided. Using the new divisions, it has been determined that several new ceremonial centres with pyramids were established for a brief time toward the end of the period, while the older ones were abandoned. The new centres indicate that the region had considerable religious prestige, and by the end of the period their abandonment and the growth of Pachacamac show that this prestige was becoming focussed there.

Pucará and Tiahuanaco

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

Tiahuanaco, near the Bolivian side of the lake, is a famous ceremonial centre, whose stone ruins are surrounded by a hummocky area littered with potsherds. The area must have been occupied by adobe buildings, the remains of a considerable town. The occupation of the ceremonial centre is believed to have begun very early in the period, and some incised and painted pottery, decorated with step designs and animals in red, dark gray, and white on a light yellow to chestnut ground, has been found in the deeper levels. This pottery could be related to the incised and painted ware of Pucará. The ceremonial buildings, whose exact age is uncertain, include a large stepped pyramid or platform called Akapana, with foundations of buildings on the top; a semi-subterranean temple with stone heads tenoned into the walls; and a low rectangular platform called Calasasaya, enclosed by a retaining wall of upright stones alternating with smaller rectangular blocks. In one corner of the platform stands a great monolithic doorway, not in its original position, cut from a large block of lava. At the top of the doorway is carved a central low-relief figure attended by three rows of smaller winged figures that appear to run inward toward him. The central figure, carrying staves that may represent a spear-thrower and darts, has been likened to the Chavín Staff God and for convenience may be called the Doorway God. Versions of the Doorway God and his attendants are found almost everywhere within the range of Tiahuanaco influence in the subsequent Middle Horizon. Another feature of the site is a number of large and finely finished stone blocks with niches, doorways, and recessed geometrical decorations. Tiahuanaco masonry is sometimes held together by accurately cut notches,

sometimes by copper clamps set in either straight or T-shaped grooves. Several massive monolithic statues have been found in and around Tiahuanaco, the largest being 24 feet (seven metres) high. They resemble pillars bearing relief designs and some carry beakers.

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

THE MIDDLE HORIZON

The Middle Horizon began about AD 600 (Rowe says 540) and is marked by the spread of influences from the south highlands that united wide areas. One centre was Tiahuanaco, which affected the highland area from the Titicaca Basin southward to the Chilean coast; the other was Huari, near Ayacucho, which was itself influenced by Tiahuanaco and finally affected most of Peru. The Tiahuanaco sphere is chiefly known from finds of pottery and wood carving in pure Tiahuanaco style. The purity of style suggests a rapid spread and points to the existence of a short-lived empire, but the evidence is quite insufficient. On the other hand, the Huari sphere has with reason been called an empire. It began with interactions between the coastal valleys of Ica and Nazca, and Huari, where Nazca influence was strong. At about the same time, Tiahuanaco motifs, including the Doorway God and his attendants, began to appear on pottery at and near Huari and would appear to represent religious influences. Large numbers of massive ceremonial urns, which carried these motifs, were sacrificed by being broken at shrines at Huari, a neighbouring site, and Pacheco in the Nazca Valley. Local pottery styles also carried some Tiahuanaco details, apart from the religious figures, and the widespread distribution of these in the central and south coast and in the highlands north to the Callejón de Huaylas indicates that the Huari expansion had begun. After a period of consolidation, the expansion was intensified, eventually reaching Cajamarca and the Chicama Valley in the north and the Ocoña Valley far down on the south coast, by about 800. The growth of the empire and its nature is shown by a number of features. There was the spread of Huari pottery styles and local copies of them, some bearing the Doorway God and other religious figures but many with neutral or secular motifs such as bands of chevrons. These are generally polychrome wares, and figures appearing on them—mythological, human, or animal—may have the eyes divided vertically into black and white halves, as at Tiahuanaco. A result of the increasing dominance of Huari styles was the obliteration of the old pottery styles over the whole coast from Nazca to Moche. The southern burial custom of huddled, cloth-wrapped mummies spread northward along the coast, displacing the older fashion of extended burial. The presence of great groups of storage buildings at Piquillacta in the Cuzco Valley and at Viracochapampa, near Huamachuco far to the north, suggests military activity like that of the Incas later. On the coast, some great cities in the north—of which Chan Chan, near Trujillo, is the best known—originated at this time, apparently under southern influence, and the rectangular Great Enclosure Compounds in the Virú Valley may be an expression of the same phenomenon. All these changes, taken together, point strongly to military conquest.

The rise of the Huari Empire

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

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

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

In the Marañón drainage near Cajamarca, is the great fortified hilltop of Marca Huamachuco, containing buildings of irregular stones whose walls survive to a considerable height. They are grouped about a circular citadel and are surrounded by a double wall. This was probably the enemy who provoked the campaign for whose prosecution Viracochapampa was built. A walled town at Kuélape, about 150 km north and somewhat similar, may have been built by the same people.

When the Huari Empire reached its maximum extent, at about 800, it collapsed at the centre. Huari was abandoned, as it appears was also Tiahuanaco, which then has been estimated to have had 5,000 to 10,000 people, and Cajamarquilla, a large urban centre near Lima. Thereafter, few signs of urban life occur in the south, except at Pachacamac, until Inca times. Curiously, the abandonment of the cities in the south appears to have coincided with the beginnings of urban settlement on the north coast at Chan Chan, Pacatnamú, and other places.

The fall of Huari and Tiahuanaco

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

THE LATE INTERMEDIATE PERIOD AND THE CHIMU EMPIRE

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

There were few new advances in techniques, and perhaps the most notable was the spread of bronze to the Peruvian coast from northwest Argentina and Bolivia, where tin ore is found and where the manufacture of bronze appears to have originated during the Middle Horizon. A hard alloy of copper and arsenic had been used previously in the centre and north, but the arsenic may well have been a natural impurity in the copper ore. Pottery improved in quality in most areas, though its artistic character was not equal, for instance, to the earlier products of Moche and Nazca. There was a tendency toward standardization and toward reduction in the number of colours, which went with a degree of mass production. The modelling tradition of the north coast revived, but it was dull and lifeless by comparison with that of Moche times and was generally executed in black

The spread of bronze

ware. In other parts, entirely new styles evolved. That of Chancay, on the central coast, is thin, dull red or cream in colour, with rather a dusty-looking cream slip and painted decoration in black. A common shape is an egg-shaped jar with a flaring collar and a pair of small loop handles, which are sometimes turned into a figure by modelling a face on the collar and adding ill-shaped limbs. Bowls and beakers with slightly bowed, almost vertical sides, are other common shapes. The porosity of the ware and the flaky nature of the slip make this pottery inferior in quality to that of other coastal areas. In the south, pottery of an attractive style was made in the Ica Valley. It is hard, well-burnished buff or red ware, covered with painted, textile-derived patterns in black, white, and red, although some vessels also depict small birds and fish. Highly characteristic are bowls with a rounded base meeting the inward-sloping sides at a sharp angle and a thickened rim of triangular section.

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

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

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

Chimú pottery

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

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

The Chimú state originated in the Moche Valley, where its capital, Chan Chan, lies. There are other cities at Farfán and Pacatnamú in the Pacasmayo Valley and at Purgatorio and Apurlé in the Leche and Motupe valleys, which share some features with Chan Chan. All include large walled compounds. Apart from the cities, there were defensive settlements, such as one in the narrow part of the Moche Valley, up which it straggles for five miles, occupying terraced hillsides and side valleys and commanding three of the main canals. A third type of settlement consists of scattered compounds in the midst of large irrigated areas, and there is one example in the Chicama Valley alongside an irrigation canal that took water to Chan Chan. Chan Chan covers an area of about 14 square miles (36 square kilometres), with a central area of about 2.5 square miles containing ten or more large rectangular enclosures sometimes called *ciudadelas* (citadels). These are surrounded by tapering adobe walls, 10 feet thick at the base and about 30 feet high. They range in size from about 400 by 200 yards to 650 by 400 yards. At least six of these citadels have similar plans, and one has recently been studied in detail. It is entered at the north end by a narrow opening and is divided into three parts by high walls. The northern part contains a large entry court, flanked by a kitchen area and several smaller courts, leading to a densely built area of small courts, some of which have a U-shaped structure at one end, while others are filled with small rooms. The U-shaped structures, which do not appear to have been roofed, may have been shrines, and the courts that contain them may have walls covered with mud-plaster reliefs, such as bands of animals, birds, or fish, scrolls or step frets, arranged in a manner reminiscent of Chimú textiles. The central part has a somewhat smaller entrance court leading to several courts occupied by rooms, perhaps storerooms, although nothing was found in them. Another feature of this area is a great burial platform with rows of chambers arranged in three levels. All these features are connected by narrow and tortuous passages. The southern part is an open area, containing one or more *pukíos* (rectangular areas where the ground has been lowered to the water table, either to supply water or to grow plants). *Pukíos* may also be found in other parts of an enclosure. In the spaces between the enclosures, and elsewhere in the city, are large areas of dwellings, irrigated areas, and cemeteries. It is now thought that the *ciudadelas* may have been the dwellings of the ruling classes and their immediate retainers, and it has even been suggested that they were the palaces of successive rulers, maintained by their descendants in the way that those of deceased Incas were maintained in Cuzco. The number of recognizable *ciudadelas* agrees with the number (ten) of known Chimú rulers. This interesting suggestion is further supported by the belief that the Incas learned a great deal from the Chimú after they conquered them, for, not content with carrying Minchançaman off to Cuzco, they established a colony of north-coast workmen there, and Topa Inca appears to have worked out the political organization of the empire at the same time, basing it largely on the Chimú system. Roads between the valleys were always necessary to coastal states and quite vital to the Chimú Empire, and the Incas may have

Chan Chan

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

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

II. Inca civilization

When the Spanish conquistador Francisco Pizarro and his followers first reached Peru in 1527, the Incas governed an empire they called Tawantinsuyu that stretched along the Pacific coast and through the Andes from what is now the northern border of Ecuador to the Maule River in central Chile—a distance roughly equal to that between New York City and the Panama Canal. By conservative estimates, they ruled more than 12,000,000 people who spoke at least 20 unrelated languages. A century earlier, the Incas had controlled little territory beyond their own village in the Cuzco Valley. They formed their empire by conquering and subordinating more than 100 independent ethnic groups—all of which had their own distinctive ways of life and forms of government before they were incorporated into the Inca state.

The Incas kept detailed accounts of their history in order to justify and validate the political and economic advantages they possessed. These traditions were passed down verbally from one generation to the next by a group of professional "memorizers" who either belonged to the Inca elite or were closely tied to it. These traditions were written down by Spanish chroniclers in the 16th century, and they recorded history in terms of the lives and deeds of the various Inca rulers. The generally accepted list of rulers, together with the names applied to the royal corporations, or *panacas*, composed of their descendants is provided in the following table:

Inca Rulers and Royal Corporations

Inca name	Spanish spelling	*panaca* (royal corporation)
Manqo Qhapaq	Manco Capac	Chima
Zinchi Roq'a	Sinchi Roca	Rawra
Lloq'e Yupanki	Lloque Yupanqui	'Awayni
Mayta Qhapaq	Mayta Capac	'Uska Mayta
Qhapaq Yupanki	Capac Yupanqui	'Apu Mayta
'Inka Roq'a 'Inka	Inca Roca	Wika-k'iraw
Yawar Waqaq	Yahuar Huacac	'Awqaylli
Wiraqocha 'Inka	Viracocha Inca	Zukzu
'Inka 'Urqon	Inca Urcon	none
Pachakuti 'Inka Yupanki (1438–71)	Pachacuti Inca Yupanqui	'Iñaqa
Thupa 'Inka Yupanki (1471–93)	Topa Inca Yupanqui	Qhapaq
Wayna Qhapaq (1493–1525)	Huayna Capac	Tumipampa
Washkar 'Inka (1525–32)	Huascar	Washkar
'Ataw Wallpa 'Inka (1532–33)	Atahuallpa	none
Thupa Wallpa (1533)	Topa Huallpa	none
Manqo 'Inka Yupanki (1533–45)	Manco Inca Yupanqui	none (?)
Pawllu Thupa 'Inka (1535?)	Paullu Topa Inca	with Tumipampa
Thupa 'Amaru (1545–72)	Topa Amaru	none

The Incas exercised complete control over the content of all their historical traditions. Consequently, these accounts present an official version of Inca history that emanated from the capital and that was designed to preserve their political regime. Some of the traditions were told publicly because of their propaganda value; the details of others were preserved but kept secret; still others were suppressed entirely because of the potentially harmful information they contained.

THE ECOLOGICAL BACKGROUND OF INCA EXPANSION

The Incas were an Andean people and shared many behaviour patterns and ideals with other ethnic groups in the area. One of the most important of these was the ideal of self-sufficiency. This meant that the members of each ethnic group tried to control all of the resources they needed to sustain themselves throughout the year, so they would not have to rely on other groups to provide them with essential goods and services. In order to achieve this ideal, each group had to cope efficiently with one of the most diverse natural environments in the world. The Andean area is an ecological mosaic of many distinct niches. Since many food resources grow in only one or a few of these niches, there are significant differences in their availability and abundance from one place to another. Corn (maize) is rarely planted above about 9,000 feet (3,000 metres) because the growing season is too short, and potatoes are rarely planted below. There are also seasonal and yearly variations in the availability of many resources.

The Inca ideal of self-sufficiency

Ideally, a group had access to the whole range of food resources that were used in the Andes. This implies that they also had access to the whole range of ecological niches that occur there and that they had to deal effectively with the patterning of resource distribution in the area. To do so, they recognized two important patterns. The first was a repetitive one consisting of a series of resource areas that repeat themselves every 10 to 20 miles (20 to 30 kilometres). It is most apparent on the west side of the Peruvian Andes, where fertile river valleys with abundant food resources are separated from each other by stretches of barren desert and hills. It also occurs in the highlands, where the north–south running river valleys are separated from each other by alpine grasslands. The other pattern is a linear one that consists of a sequence of resource areas that never repeat themselves. This, too, is most apparent on the west side of the Andes because of the steep elevation gradients. Corn and other crops requiring long growing seasons are planted at low elevations; potato fields occur at middle elevations; and grazing lands occur at high elevations, where the fewness of frost-free days prohibits farming. This linear pattern also occurs in the highlands, but is less obvious because of smaller elevation gradients.

The distinction between repetitive and linear arrangements of food resource areas had an important influence on where a group went to obtain the various kinds of food it needed. The food resources that were immediately available to a group living in the upper part of the Cuzco Valley, such as the Incas, were roughly the same ones that were available to groups living in similar environmental settings in other valleys, and they were quite different from the resources that were immediately available to groups living in different environmental settings in the same valley or in other valleys. A group wanting resources that could not be obtained locally had to deal with the linear arrangement of resource areas. Its members could acquire this resource only by gaining access to land in the environmental setting where it occurred. A group wanting something that was in short supply locally had to deal with the repetitive arrangement of resource areas. Its members could acquire additional sources of this resource by gaining access to lands in similar environmental settings in adjacent valleys.

It is useful to view the history of the Incas and their state from the perspective of self-sufficiency because it provides an explanation for much of their activities. Like other Andean peoples, the Incas wanted to be self-sufficient, and they worked out novel ways for achieving this goal. Their origins and early history are largely shrouded in legends that may be more mythical than factual. Their later history, particularly from the reign of Pachacuti Inca Yupanqui (Pachakuti 'Inka Yupanki) onward, is largely based on fact, even though it presents what the Incas wanted people to know. Whether these historical traditions are true, in the sense that they accurately related what happened, is not so important as the fact that the Incas used them to justify their various imperial conquests.

THE ORIGINS AND EXPANSION OF THE INCA STATE

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

Paqarina, the shrines of origin

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

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

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

Mayta Capac

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

The beginnings of external expansion. The fifth emperor, Capac Yupanqui (Qhapaq Yupanki), was appointed ruler by his father before he died. He was apparently not the eldest son but was named emperor because his older brother was considered ugly. Capac Yupanqui was the first Inca ruler to conquer lands outside the Cuzco Valley, although these were only about a dozen miles away. Inca Roca ('Inka Roq'a 'Inka) succeeded his father and subjugated some groups that lived about 12 miles (19 kilometres) southeast of Cuzco. He is mostly remembered in the chronicles for the fact that he fathered a large number of sons, one of whom, Yahuar Huacac (Yawar Waqaq), was kidnapped by a neighbouring group when he was about eight years old. The boy's mother, Mama Mikay, was a Huayllaca (Wayllaqa) woman who had been promised to the leader of another group called the Ayarmaca ('Ayarmaka). When the promise was broken and Mama Mikay married Inca Roca, the Ayarmaca went to war with the Huayllaca and were defeating them. As a peace offering, the Huayllaca agreed to deliver Mama Mikay's son to the Ayarmaca. This tale says a great deal about the way war was waged around the Cuzco Valley at this time; the fact that the Ayarmaca held the boy for several years before returning him to his father suggests that the Incas were no more powerful than several other groups in the area.

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

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

The beginning of permanent conquests

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

During the early 15th century a group called the Chancas was emerging as a political power in the area west of the Inca territory. Presumably, they, too, may have been feeling the effects of diminishing food resources and were trying to maintain their standard of living by acquiring land outside of their home territory. They moved from their place of origin in Huancavelica and conquered the Quechua (K'ichuwa), a large group whose lands lay immediately west of those controlled by the Incas. In about 1438 The Chancas attacked the Incas. One of the major effects of the Chanca invasion was to foment a civil war among the Incas.

Rise of the Chancas

Pachacuti Inca Yupanqui. For some time there had been palace intrigue in Cuzco over who would succeed

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

Victory over the Chancas

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

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

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

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

Murder of Capac Yupanqui

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

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

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

Defeat of the Chimú

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

Pachacuti Inca Yupanqui began rebuilding Cuzco, the political and religious capital of the empire, and at the same time undertook a vast agricultural project over the entire upper end of the Cuzco Valley; rivers were channelled, the valley floor was levelled, and agricultural terraces were built on the surrounding hillslopes. This reclamation project undoubtedly increased the agricultural productivity of the area and involved moving many of the original inhabitants of this part of the valley to other localities for about five years while the work was being completed.

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

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

group by dispersing its members throughout the empire; second, it weakened the ability of an ethnic group to be self-sufficient; and, third, it made it more difficult for the inhabitants of an area to revolt successfully.

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

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

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

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

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

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

Conquest of northern Ecuador

The conquest of northern Ecuador occupied the last years of his life and took place shortly before the Spaniards arrived. He was interested in the area because he had been born in Tumipampas, on the outskirts of Cuenca. During these campaigns, he pushed the frontiers of the Inca Empire to the Ancasmayo River, the present-day boundary between Ecuador and Colombia.

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

Meanwhile, he undertook another expedition in northern Ecuador to wipe out isolated pockets of resistance. During this campaign, he learned that an epidemic was sweeping Cuzco and the surrounding countryside. He left immediately for Quito, on the highroad to Cuzco, to deal with this crisis and arrived there about the same time the epidemic did. The pestilence had spread rapidly from Bolivia and, judging by its description, was either smallpox or measles, both of which were European diseases introduced into South America by the Spanish settlers at La Plata. The disease was probably communicated to the Andean area by the Chiriguano, who had been in contact with the Europeans at La Plata. Whatever the ailment was, Huayna Capac contracted it and died about 1525, without naming a successor in the appropriate manner. This set off another struggle over the throne.

Civil war on the eve of the Spanish Conquest. Huayna Capac's father had begun the custom of marrying a full sister in order to keep the royal bloodline pure and, more importantly, to prevent conflict over succession.

Rule of succession

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

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

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

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

Victory of Atahuallpa

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

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

Imprisonment of Atahuallpa

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

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

INCA CULTURE AT THE TIME OF THE CONQUEST

The story told above largely relates the way the Incas saw their history. The traditions of why they conquered and how they ruled emphasize their own perspectives on the situation that prevailed between 1438 and 1532, rather than those of the peoples they subjugated. If these traditions are accepted at face value, one is led to believe that the Incas brought peace and tranquillity to the Andean area and gave their subjects a divine revelation of how the world should be structured. The events of this period, however, can be viewed from other perspectives —those of the ethnic groups who were conquered or subjugated by the Incas and had to deal with them on a day-to-day basis. They, too, recorded their historical traditions, and enough of them have already been found to present a useful counterbalance to the official accounts of their rulers. By judiciously using these accounts in conjunction with those of the Incas, a much more accurate picture can be gained of how Andean society functioned during the 15th century and the role the Incas played in it.

Social and political structure. The Incas distinguished between those individuals who claimed descent from one of the 12 Inca rulers and belonged to his royal corporation and those individuals who belonged to one of the Inca-speaking ethnic groups from the Cuzco area rather than one of the royal corporations. The former were called "Incas by blood" and formed the highest level of nobility; the emperor always came from their numbers. The latter were called "Incas by privilege," and members of this group were often sent to other parts of the empire to fill responsible administrative positions in the Inca government.

The hierarchical structure

The Inca political system was a hierarchical one and is usually represented as a pyramid. The highest official in the system was the emperor, who not only ruled by divine right but also claimed lineal descent from the sun and was worshipped as divine during his lifetime—at least by the Incas. One of the great weaknesses at this level of the Inca political system was the lack of a detailed rule of imperial succession. As a result, the succession was frequently associated with conflict among the royal corporations, and none of the methods which the Incas devised to eliminate such strife was ever entirely successful.

Immediately below the emperor in the administrative hierarchy were the *'apo*, or prefects, of the four largest territorial divisions of the empire. They belonged to the highest level of Inca nobility and were often close rela-

tives of the emperor. They lived in Cuzco, and their offices were apparently not hereditary. Below them were the provincial governors, the *t'oqrikoq*, who resided in the various provincial capitals and were the lowest government officials that had to be Incas. They had wide judicial powers and administrative responsibilities and directed the activities of the *kurakas*, who formed the lower levels of the Inca bureaucracy.

The *kurakas* governed units composed of 100, 500, 1,000, 5,000, and 10,000 taxpayers. These positions were probably held largely by the traditional leaders of the ethnic groups conquered by the Incas. The Incas solved the problem of dealing with a wide range of political systems and various ways of selecting leaders by making succession to these offices hereditary, subject to the approval of the emperor. The sons of *kurakas*, particularly those responsible for large numbers of taxpayers, were often taken to Cuzco, where they were indoctrinated in the ways of the Inca state. This not only insured the loyalty of their fathers but also provided the Incas with a steady flow of trained personnel for the lower echelons of their bureaucracy. The *kurakas* were exempt from taxes and received special gifts—such as luxury goods usually reserved for the Inca nobility, secondary wives, and retainers—in return for outstanding service to the state.

Administration of the empire. The organization of the population into decimal units administered by the *kurakas* seems to have followed traditional divisions between and within the ethnic groups that had been incorporated into the Inca state. Consequently, the units were monoethnic in composition. This form of organization reflects the Incas' overwhelming concern with collecting taxes that were owed by the ethnic group as a whole rather than by the individuals who belonged to it. Since these taxes were paid in the form of labour, the system was designed to provide the *kurakas* with a way for raising a sufficient number of men to meet the demands of the state. In those areas where data are available, this organizational device did not always correspond to the demographic realities of the situation. Some decimal units contained more persons than they should have, whereas others were undermanned. Such discrepancies were probably of greater concern to the *kurakas*, who had to allocate sufficient manpower to meet the tax obligations, than to the Incas, who received the benefits of this labour.

Labour services

The Incas imposed two kinds of labour taxes on the groups they subjugated. The first was the *mit'a*, which consisted of a specified number of man-days of labour in the army, or in public works projects, or in personal service to the emperor and members of the Inca nobility. The Incas provided all of the necessities required by these individuals during the period of their *mit'a* service. The second was a form of agricultural and pastoral taxation, in which the members of each ethnic group were required to work the lands, or resource areas, and care for the llama and alpaca herds that had been taken from them by the Incas. The products of their labour were used to support the Inca government, the state religion, and the members of any royal corporation that controlled resources in their territory. The Incas controlled varying proportions of the resources that were produced in different localities. In some areas, such as Arapa in the Lake Titicaca Basin, the state religion controlled everything that was produced, whereas in other areas, the largest shares went either to the state government or to one or more of the royal corporations. The remainder of the resources in an area belonged to the ethnic group whose members actually did the work; apparently, however, no group was allowed to keep more resources than its members needed to sustain themselves throughout the year. In times of famine or drought, the Incas distributed food to the affected groups. This food came from the state-owned lands and was stored in warehouses located throughout the empire.

In order to facilitate taxation, the Incas had up-to-date census information. Each year, the *kurakas* counted the number of persons under their jurisdiction and forwarded this information to Cuzco. Individuals were classified according to their physical characteristics into one of 12 age-grades; these categories correspond roughly to the concepts embodied in such English words as infant, child, able-bodied adult male, decrepit old man, and so forth. These data were recorded by knots on pieces of coloured string attached to a main cord. These devices, called quipus, provided the Incas with detailed records of the human resources available in each province of the empire.

The prevention of rebellion. The Incas demanded complete obedience from their subjects and used officials and institutions outside the pyramidal structure to insure that they received it. There was a group of special officials called *tokoyrikoq*, whose activities included making public and secret investigations of the activities of provincial governors and *kurakas* and reporting the findings to the emperor and his councillors. Two of the group's major concerns seem to have been to insure that tax obligations were being met satisfactorily and that rebellions were not being fomented by dissatisfied *kurakas*. The Incas also sent out special judges to punish individuals for crimes that touched on Inca power; these included having relations with one of the chosen women assigned to the temples of the state religion, performing witchcraft against the emperor, and failing to pay labour taxes.

Resettlement of peoples

The resettlement of populations loyal to the Incas in hostile or newly conquered territories was also used to guarantee obedience to the state. The *mitma* groups were obliged to retain their own cultural identities; *i.e.*, customs, language, religion, and distinctive dress. Consequently, they were never really integrated into the local populations where they were settled nor were they placed under the control and jurisdiction of the local *kurakas*, who had to provide them with land and partially support them during their first two years of residence. They were charged with setting proper examples for the native inhabitants of the area and with providing the state with information about their activities. In effect, they were spies and potential soldiers in the event of rebellion. The Inca resettlement program was carried out on a large scale, judging by the fact that the *mitma* groups outnumbered the original inhabitants in many provinces at the time of the conquest.

The Incas tolerated religious differences and permitted the conquered ethnic groups to practice their own religions as long as they also fulfilled their obligations to the state church. But the Incas also used local religion as a vehicle of political control. Once a group was incorporated into the empire, the Incas took its most sacred religious object or idol, called a *wak'a*, to Cuzco, where it was enshrined in one of the Inca temples under the care of the local priests who served it. These *wak'a* were worshipped in Cuzco when a provincial delegation had to come to the capital for some reason. In a very real sense, the Incas were kidnapping *wak'a* and holding them and their priests as hostages. They were safe as long as the members of the group did what the Incas wanted them to do.

The army

The army was also used to extend and enforce Inca authority. Diplomacy was one of the major features of their military system. During the period when they were extending the empire, the Incas would invite representatives from the ethnic groups they planned to incorporate into the state and point out to them the obvious advantages of accepting Inca rule peacefully and the disadvantages of refusing to do so. The fact that the state grew so rapidly under Topa Inca Yupanqui suggests that many groups chose to ally themselves with the Incas, albeit as subordinates, rather than face the consequences of defeat.

The army was composed mainly of troops from ethnic groups that were conquered by the Incas or incorporated peacefully into the state. Every able-bodied man belonging to the appropriate age-grade was subject to military service; and, in all likelihood, most of them spent some time in the army. The only full-time professional soldiers were apparently the high-ranking Inca

officers and the members of the emperor's bodyguard. The army was strong and efficient because of its size, its discipline, and its ability to get supplies regularly. In the organized parts of the empire, supplies were always available in the state warehouses. These were close enough to each other that the troops could always camp at one of them as they moved from one part of the empire to another. When they were subjugating areas outside the empire, their supplies were brought by large pack trains of llamas.

The Incas established garrisons throughout the empire that must have served as deterrents to local rebellions. The garrisons were apparently located in the provincial capitals and were composed of troops that came from other parts of the empire and had no loyalties or ties to the native inhabitants of the province. When the members of an ethnic group revolted, the Incas used troops from these garrisons and from any of the group's traditional enemies to suppress the rebellion.

The final mechanism used by the Incas to prevent rebellions and maintain their authority was fear. If an ethnic group revolted, the army attacked; and the rebels were quickly suppressed. Their leaders, usually the local *kurakas*, were taken to Cuzco, where they were paraded through the streets or trod upon by the emperor as a symbol of their subordinance to him. Often they were imprisoned, tortured, or killed. If they were killed, their skulls were frequently made into cups from which the emperor drank maize beer, or *chicha;* and drums were made from their skins and bodies, as final insults to the rebellious group. The Incas also imposed much heavier demands on the remainder of the group. They apparently confiscated as much as half the remaining arable lands and herds that were being used by the group to maintain its members. This action, of course, placed a tremendous burden on the unsuccessful rebels, since they had few resources to sustain themselves and had to provide labour to work the fields and care for the herds that the Incas took from them.

In spite of all the precautions the Incas took to prevent rebellions, it is clear from their traditions that uprisings did occur. They may have happened with increasing frequency, as the geographical expansion of the empire slowed down after the reign of Topa Inca Yupanqui and the demands of the state, the church, and the royal corporations for access to the available resources in provincial areas continued to grow. This continually increasing burden on the conquered ethnic groups must have been a source of growing dissatisfaction as they saw larger and larger amounts of their resources and labour drawn off by the state. The Incas could not deal with the difficulties, frustrations, and irritations that must have accompanied their growing demands. Too many provincial officials in the lower levels of their bureaucracy belonged to the groups that had to meet them. When conditions became too oppressive and rebellion broke out, these officials sided with their kinsmen against the central government. This dissatisfaction is one of the reasons why so many Andean peoples later became close allies of the Spaniards.

Inca technology. Many of the devices and techniques used by the Incas—for example, the equal-arm balance, bronze, large-scale water management systems, and the quipu, to name only a few—were invented earlier in the Andean area and were adopted by them and their contemporaries. Specific inventions that seem to have been originated by the Incas include mortarless masonry, the plumb bob, and an agricultural implement used for digging (the *taclla*).

The Inca intellectual tradition did not include the idea of generalizing mechanical principles. The crowbar, for example, was used in mining, and the equal-arm balance was used for weighing various commodities. But there is no evidence that the Incas or their Andean contemporaries understood the abstract principle of the lever. Consequently, each invention was the solution to a particular problem, and even the most obvious extensions of these inventions to other problems were never made, as a result.

The spread of technological improvements

Perhaps the Incas' greatest contribution to Andean technology was that they spread many useful tools and techniques from one end of the empire to the other and that they popularized the use of metal tools. Prior to the Inca conquest, metals—gold, silver, copper, and their alloys—were used mainly for ornaments; and tools were made from wood and stone. Bronze tools—crowbars, chisels, axes, knives, and clubheads, to name only a few—became exceedingly common after the conquest.

The Inca road system facilitated the spread of technological innovations throughout the empire and the possibility of governing such a large area from a single capital. Two main roads ran the length of the empire, one along the coast and the other in the highlands. The coastal road began in Tumbes and went at least as far south as Arequipa in southern Peru. The highland road began at the Ancasmayo River on the Colombian–Ecuadorian border and went southward through many of the large population centres in the empire to Ayaviri, where it divided to go around both sides of Lake Titicaca. One branch went down to the coast at Coquimbo, Chile, and followed the shore southward to the area around Santiago. The other branch passed through Chuquisaca, Bolivia, and went southward through northwestern Argentina to the vicinity of Mendoza. The two main arteries were connected to each other and to many of the important population centres by a network of smaller, lateral roads. There were shelters and storehouses, called *tampos,* located at one-day intervals along all of the roads, where travellers could spend the night.

System of weights and measures

The Incas' concern with land and taxes meant that they had to have a system of measurements to determine the extent of their holdings, a calendar for calling up labour and performing ceremonies, and a way of recording numerical information. Inca measurements were based on parts of the human body. Two of the units they used were the span of a man's hand, roughly eight inches (20 centimetres), and the fathom—about 64 inches (160 centimetres). The latter was the standard measurement of land, and a measuring stick of this length served as a legal check. Travelling distance was measured in terms of the pace (about four feet). A larger unit called the *topo* was used along the roads and apparently consisted of 6,000 paces. Area was also measured in terms of the *topo;* however, the size of a *topo* of land is not certain. It may have been an area of 50 fathoms by 20 fathoms (300 by 120 feet). The Incas measured grain in terms of the *phoqca,* which was about 26 quarts, but they apparently had no standard for measuring liquids.

The Inca calendar was an annually self-correcting one. The year began at the time of the December solstice. This was determined by making observations of where the sun rose and set with respect to sets of four towers located on the skylines east and west of Cuzco. These observations were made from a raised platform in the middle of the Great Square in Cuzco. Since the December solstice is the longest day of the year in the Southern Hemisphere, it occurred when the Sun rose and set at the southernmost markers. The year was divided into 12 months of 30 days each; a month was divided into three ten-day weeks. The remaining five days were taken up by the two most important Inca ceremonies—one at the time of the December solstice and the other during the June solstice. The times for planting, harvesting, and ceremonies were determined in terms of the Inca calendar.

Numerical records

The Incas recorded several kinds of information on quipus, which consisted of a main cord and a variable number of pendant cords of different colours. Some of the quipus were clearly used for recording numbers and their sums in terms of the decimal system. The group of knots placed farthest away from the main cord were units; in the next place were tens; and so forth. The absence of a knot in one of these places represented zero. In order to use a quipu that was knotted by someone else, it was necessary to have a description of what was being counted on the various strings. The quipu is still used today in several parts of the Andes for counting herd sizes. While the quipu is well-adapted for recording

numbers, it is a clumsy calculating device; and the chroniclers make it clear that calculations were done with piles of stones or with a device resembling an abacus.

Quipus were also used as a mnemonic device in reciting narratives, genealogies, and liturgical materials. The Incas had a special class of quipu interpreters, called *khipo-kamayoq* (*quipu-camayoc*), who memorized the statistical, historical, and liturgical information collected by the government and recited it when they were called upon to do so. (T.C.P.)

BIBLIOGRAPHY

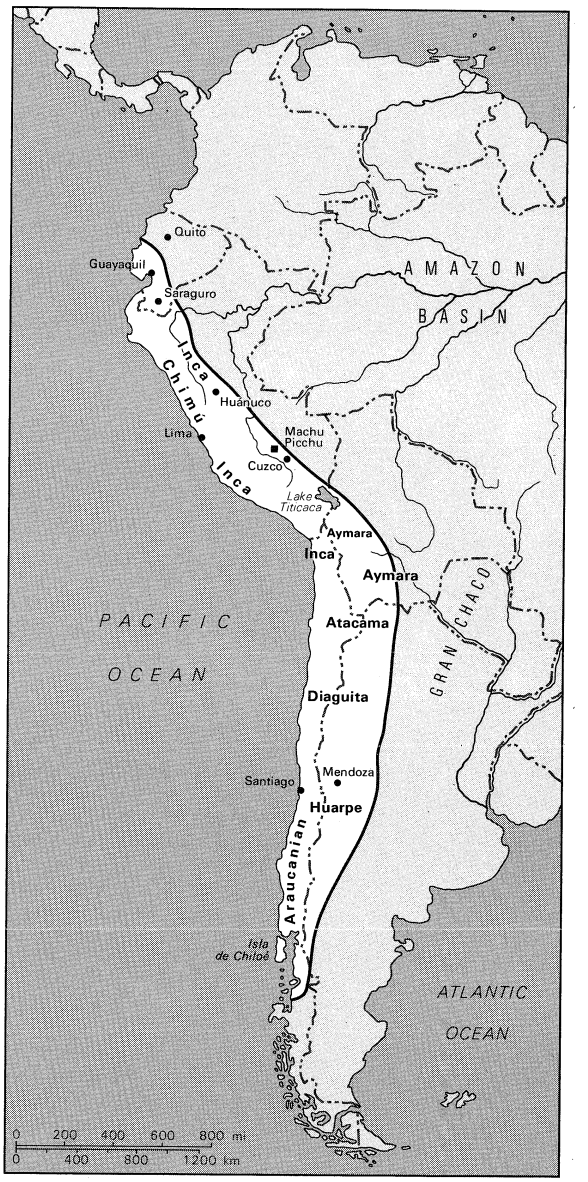
Pre-Inca civilization: WENDELL C. BENNETT and JUNIUS B. BIRD, *Andean Culture History*, 2nd ed. (1960), is out-of-date in many respects, particularly on chronological terms, but still affords valuable information, has a particularly useful section on techniques, and contains good illustration and maps. G.H.S. BUSHNELL, *Peru*, 2nd ed. (1963), also dated, particularly on chronological terms and the Preceramic, is still useful as a general account of the ceramic phases (good illustrations). E.P. LANNING, *Peru Before the Incas* (1967), the most up-to-date general book, is particularly good on the Preceramic and includes useful maps. J. ALDEN MASON, *The Ancient Civilizations of Peru*, rev. ed. (1964), out-of-date, especially on chronological terms and on the Preceramic, has useful information on and illustrations of conditions in Inca times. More specialized are E.P. LANNING, "Early Man in Peru," *Scient. Am.*, 213:68–76 (1965); JOHN HOWLAND ROWE and DOROTHY MENZEL (eds.), *Peruvian Archaeology: Selected Readings* (1967), which contains a very useful chronological table compiled by Rowe and a number of reprints of valuable articles, especially "Form and Meaning in Chavin Art," by Rowe; "Pre-Ceramic Art from Huaca Prieta, Chicama Valley," by J.B. BIRD; "Style and Time in the Middle Horizon," by DOROTHY MENZEL; and "Tiahuanaco Tapestry Design," by A.R. SAWYER; DONALD LATHRAP, *The Upper Amazon* (1970), for information on the Eastern Lowlands; and SEIICHI IZUMI and TOSHIHIKO SONO, *Andes 2 Excavations at Kotosh, Peru, 1960* (1963).

The Inca civilization: Knowledge of the Incas is based primarily either on information collected by 20th-century archaeologists or on accounts that were written after 1532. References to some of the more widely used accounts may be found in the standard work on the Incas written by JOHN HOWLAND ROWE, "Inca Culture at the Time of the Spanish Conquest," *Bull. Bur. Am. Ethnol. 143*, 2:183–330 (1946). Another independent assessment of the chronicles is found in J. ALDEN MASON (*op. cit.*). The idea of community self-sufficiency was first conceived by JOHN V. MURRA and is most clearly stated in his paper "An Aymara Kingdom in 1567," *Ethnohistory*, 15:115–151 (1968). Murra also wrote a series of stimulating papers on the political and economic organization of the Inca Empire. These include "On Inca Political Structure," *Proceedings of the 1958 Annual Spring Meeting of the American Ethnological Society*, pp. 30–41 (1958); "Cloth and Its Functions in the Inca State," *Am. Anthrop.*, 64:710–728 (1962); "Social Structural and Economic Themes in Andean Ethnohistory," *Anthrop. Q.*, 34:47–59 (1961); and "Herds and Herders in the Inca State," in *Man, Culture, and Animals: The Role of Animals in Human Ecological Adjustments*, ed. by ANTHONY LEEDS and ANDREW P. VAYDA (1965). There is currently some controversy over whether the Inca historical traditions should be interpreted as representing real history. JOHN HOWLAND ROWE, "Absolute Chronology in the Andean Area," *Am. Antiq.*, 10:265–284 (1945), best fits the available archaeological evidence. An alternative view is presented by AKE WEDIN in *La cronología de la historia incaica* (1963). Valuable information concerning the problem of succession to the Inca throne is found in MARIA ROSTWOROWSKI DE DIEZ CANSECO, "Succession, Coöption to Kingship, and Royal Incest Among the Inca," *SWest. J. Anthrop.*, 16:417–427 (1960). SALLY FALK MOORE, *Power and Property in Inca Peru* (1958), discusses taxation and the legal and political systems of the Incas.

(G.H.S.B./T.C.P.)

Andean Cultures

Although the Andes mountain chain extends from Venezuela to the southern tip of the continent, it is conventional to call "Andean" only the people who were once part of Tawantinsuyu, the Inca Empire in the Central Andes, or those considerably influenced by it. Even so, Andean is very wide. It encompasses the peoples of Ecuador, including those of the humid coast, many of whose contacts were as frequently with maritime peoples, to both north and south, as with the peoples of the highlands. Most of the populations and civilizations of Bolivia and Peru are Andean in a central, nuclear way, and here again are included the kingdoms of the irrigated desert coast. The peoples who for the past four and a half centuries have occupied the northern highlands of Chile and Argentina also must be included. (For a description of northern Andean peoples, see CENTRAL AMERICAN AND NORTHERN ANDEAN CULTURES.)



Distribution of Andean peoples and cultures.

CENTRAL ANDES

There is a stereotyped image of the Andes showing poverty against a background of bleak, unproductive mountains, where millions insist, against all apparent logic, on living at 10,000 or even 14,000 feet above sea level. Nowhere else have people lived for so many thousands of years in such visibly vulnerable circumstances.

Yet, somehow this perception of the Andean peoples coexists with another, based on the breathtaking stage setting of such archaeological sites as Machu Picchu, the majesty of Inca stone palaces at Cuzco or Huánuco Pampa and such Chimú mud-walled cities as Chan Chan, the beauty of Andean textiles or ceramics in museums the world over, the reported concern of the Inca kings for the welfare of their subjects, and the mostly abandoned large-scale irrigation works or terraces (see also ANDEAN CIVILIZATION, HISTORY OF).

These two visions of Andean peoples and their accomplishments can be reconciled only if it is recognized that what the resources and ecologic potential of an area and a people may be depends on what part of these resources the people use or are allowed to use by their masters. The Andean region was once rich and produced high civilizations because, over millennia, its people developed an agriculture, technologies, and social systems uniquely adapted to the very specialized if not unique ecologic conditions in which they lived.

Economic systems. Since 1532, under European rule, extractive activities, such as silver, tin, and copper mining, for foreign markets have been favoured to the point to which Andean agriculture and the ecologic wisdom in handling productively the extremely high altitudes have been gradually devalued and mostly forgotten. The population of the Central Andes is both less dense and less urban today than it was in 1500. The coastal cities of South America, from Guayaquil to Buenos Aires, are filling with highlanders who have been convinced by four and a half centuries of colonial rule that cultivating at 12,000 feet is too strenuous.

Andean agriculture

Although human occupation began over 20,000 years ago, the beginnings of agriculture and population growth are much more recent. Within the last 8,000 years a specialized desert-and-highland agriculture was developed. There are two significant achievements in the Andean agricultural endeavour. First, given the wide range of geographic circumstances—very high mountains in equatorial and tropical latitudes, a 3,000-mile coastal desert, the Amazon rain forest to the east—there were thousands of quite different ecologic pockets, each with its own micro-environment to be understood and exploited. Dozens of crops, with literally thousands of varieties, were domesticated; most of them remain unknown outside the Andean area. Only the potato has acquired a following elsewhere; and only maize (corn) and possibly cotton were known in the Andean region as well as in the rest of the Americas. It is this multiplicity of minutely adapted crops and the domestication of the alpaca and the llama that made the mountains habitable to millions (the bulk of the population in the Central Andes has always lived between 8,000 and 13,000 feet).

Second, no matter how specialized Andean plants or herds may become, the leap from bare survival to dense populations and civilizations requires something more. The high altitude, with its 200, 250, even 300 frost-threatened nights a year, represents a challenge to any agricultural system. On the high, cold plains, known in the Andes as puna, there are only two seasons: summer every day and winter every night. By alternately using the freezing temperatures of the nocturnal winter and the hot sunshine of the daily tropical summer, Andean peoples developed preserves of freeze-dried meat, fish, and mealy tubers (*charki, chuñu*) that kept indefinitely and weighed much less than the original food. The giant warehouses that lined the Inca highways could be filled with these preserves and used to feed the engineers planning cities and irrigation canals, the bureaucracy and the army, not to mention the royal court, with its thousands of male and female retainers.

Political systems. Even these two technological developments, however, are not enough to characterize and explain the emergence of Andean civilizations. From the intimate knowledge of their environmental conditions, the people developed a set of values that may have started from a desire to minimize risks but that soon was elaborated into an economic and political ideal. Every Andean society—be it a tiny, local ethnic group of 20 to 30 villages in a single valley or a large kingdom of 150,000 souls, such as the Lupaca—tried to control simultaneously a wide variety of ecologic stories up and down the mountainsides; some of them were many days' march from the political core of the nation. If the society was small, the outliers (herders or salt winners above the core; maize, cotton, or coca-leaf cultivators in the warm country below) would be only three or four days away. When the political unit grew large and could mobilize and maintain several hundred young men as colonists, the outliers could be 10 or even 15 days' walk away from the core.

The colonies were permanent, not seasonal establishments. Since more than one highland kingdom or principality would have maize or coca-leaf oases in a given coastal or upland Amazonian valley, there would be not only competition for their control but also coexistence for long periods of time in a single environment of outlying colonies sent out by quite different core societies.

Inca expansion

The Inca state, or Tawantinsuyu as it was known to its own citizens, was perhaps the largest political or military enterprise of all. It reached from Carchi in northern Ecuador to at least Mendoza in Argentina and Santiago in Chile. Its scouts roamed even wider, as recent Chilean archaeology has shown. The Incas expanded and projected on earlier, pre-Incan solutions and adaptations; in the process, many tactics that had worked well on a smaller scale became inoperative; others were reformulated in such ways that their original outline was barely recognizable. For example, they kept an old Andean method of creating revenues for their princes, which involved setting aside acreage for regional authorities and demanding from the conquered peasantry not tribute in kind but rather labour on the field thus set aside. In this way the granary of the peasant household was left untouched; the authority took the risk of hail, frost, or drought decreasing its own revenues.

The Inca state at its zenith did not breach this tradition overtly; the local ethnic groups continued to work the state's acreage and owed nothing from their own larders. But since the needs of kings kept growing, revenues produced on state lands were soon inadequate; acreage could be and was expanded through such public works as irrigation and terracing. A more tangible way was to increase the amount of energy available for state purposes. For some reasons, still insufficiently understood, the kings did not increase productivity by introducing tribute; they preferred to magnify on an imperial scale the patterns of reciprocal obligations and land use familiar to everyone from earlier times.

Beyond the strategic colonies set up on an expanded model, the Incas did not interfere too much with life of the many local groups that they had incorporated into Tawantinsuyu. Most of the cultures that existed in Eucador, Peru, Bolivia, Argentina, and Chile before the Inca expansion can be identified. In fact, because the European invasion beginning in 1532 was mostly concerned with breaking the resistance of the Inca overlords, frequently more is known about the pre-Inca occupants than about Cuzco rule. Inca power was broken and decapitated within 40 years of 1532. The ethnic groups, many of which (like the Wanka or the Cañari) sided with Europeans against the Inca, were still easy to locate and identify in the 18th century. In isolated parts of Ecuador (Saraguro, Otavalo) and Bolivia (Chipaya, Macha) this can still be done today. (J.V.M.)

SOUTHERN ANDES

The area commonly referred to as the Southern Andes is separated from the Central Andes by a broad band of desert covering extreme southern Peru and northern Chile and Argentina. The Southern Andes includes those peoples occupying the mountain regions of northeastern Argentina and northern and central Chile south to Isla de Chiloé. (Tribes farther south are encompassed within the article SOUTH AMERICAN NOMAD CULTURES.) Generally, it is an area of moderate to extreme aridity, with a temperate climate. Agriculture has been an important feature of village life and, in many places, has been sustained by irrigation techniques; however, the agricultural potential—limited by the availability of moisture—has always been much less than that of Peru.

Prior to the advent of farming, the peoples of the area followed a life pattern dependent on hunting, collecting, or fishing, depending upon their geographical location. Plant cultivation, including corn, was introduced at some time after 500 BC, almost certainly diffused to the Southern Andes from Peru; and Inca influence finally appeared in the latter part of the 15th century AD. It is

known from ethnohistoric sources, as well as from the presence of Inca-style fortifications and way stations, that the Cuzco government had established its authority over these peoples of the south. These southern tribesmen—the Atacama of northern Chile, the Diaguita of northwestern Argentina, and others—lived in small, compact towns or villages and practiced irrigation farming and llama and alpaca herding. In general, the cultures of the area may be said to show a marginal and somewhat delayed expression of Andean civilization. These primary influences were also blended with influences from Amazonian cultures by way of the Bolivian lowlands. Socially, for instance, the people were organized into autonomous villages of related kinship groups, much like the tropical forest village farmers and not at all like the highly class-stratified imperial Incas; war was conventionally waged for human trophies, not for territorial conquest. (Ed.)

BIBLIOGRAPHY. The best account of Andean cultures as they appeared to an eyewitness of the early years (16th century) of European rule is PEDRO DE CIEZA DE LEON, *Crónica del Perú* (Eng. trans., *The Incas*, 1959). The earliest periods of human occupation are covered by R.S. MACNEISH, "Early Man in the Andes," *Scient. Am.*, 224:36–55 (1971). For an introduction to Andean archaeology, see JOHN HOWLAND ROWE and DOROTHY MENZEL (eds.), *Peruvian Archaeology* (1967), with selections mainly by American scholars. The role of irrigation in the cultural and political elaboration of coastal kingdoms is well illustrated in PAUL KOSOK, *Life, Land, and Water in Ancient Peru* (1965). Special regional problems are introduced in BETTY MEGGER, *Ecuador* (1966); and DONALD LATHRAP, *The Upper Amazon* (1970). Information on the Southern Andes is included in JULIAN H. STEWARD and LOUIS C. FARON, *Native Peoples of South America*, ch. 10 (1959); and in LOUIS C. FARON, *Hawks of the Sun* (1964).

(J.V.M./Ed.)

Andes Mountain Ranges

A vast series of extremely high plateaus surmounted by even higher peaks, the Andes mountain system stretches for 5,500 miles (8,900 kilometres)—from the southern tip of South America to the continent's northernmost coast on the Caribbean. One of the great natural features of the globe, it separates a small western coastal area from the rest of the continent, affecting deeply the conditions of life within itself and in surrounding areas.

The Andes is not a single line of formidable peaks but a succession of parallel and transverse mountain ranges, or cordilleras, and of intervening plateaus and depressions. Distinct eastern and western ranges—respectively named the Cordillera Oriental and the Cordillera Occidental—are characteristic of most of the system. The western cordillera is within the great belt of crustal disturbances bordering the Pacific and is thus the home of both earthquakes and volcanoes.

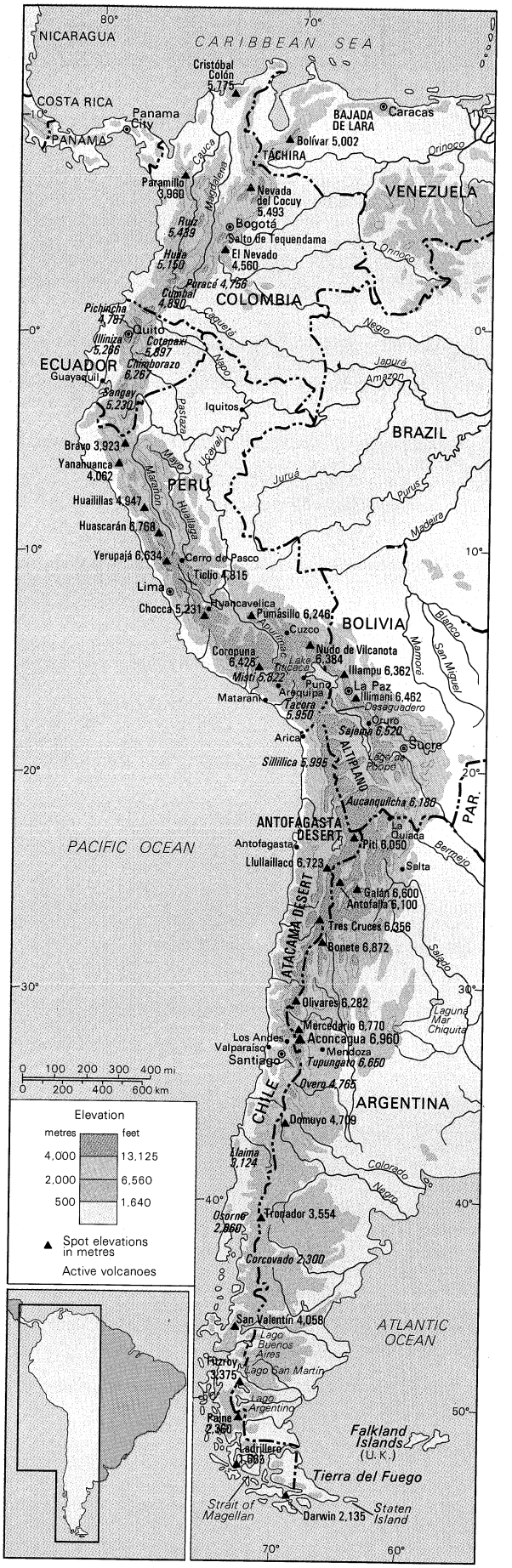
The directional trend of both the cordilleras is generally south–north, but in several places the Cordillera Oriental bulges eastward to form either isolated peninsula-like ranges or such high intermontane plateau regions as the Altiplano (Spanish: "high plateau"), occupying adjoining parts of Argentina, Chile, Bolivia, and Peru.

Some historians believe the name Andes comes from the Quechua Indian *anti* ("east"); others suggest it is derived from the Quechua *anta* ("copper"). It is perhaps more reasonable to ascribe it to the *anta* of the older Aymara language, which connotes copper colour generally. For information on related subjects, see the article SOUTH AMERICA and articles on the several countries crossed by the Andes.

THE NATURAL ENVIRONMENT

Physiography of the Andes. There is no universal agreement about the major south-to-north subdivisions of the Andes system. For practicality, the system is dealt with in this article in terms of six south-to-north subdivisions: Patagonian Andes, Central Andes, Peruvian Andes, Ecuadorian Andes, Colombian cordilleras, and Venezuelan Cordillera.

The Patagonian Andes. The true Andes begin on the mountainous Staten Island (3,700 feet [1,125 metres]) of the Tierra del Fuego archipelago, run to the east



Andes mountain system.

through Isla Grande, where the important ridges, including Darwin, Valdivieso, and Sorondo—all less than 7,900 feet high—then rise to the north beyond the Strait of Magellan. The physiography of this southernmost subdivision of the system is complicated by the presence of the independent Cerro de la Costa.

Numerous transverse and longitudinal depressions and breaches cut this wild and rugged portion of the Andean range, sometimes completely; many are occupied by ice fields, glaciers, rivers, lakes, or fjords. The crests of the mountains exceed 10,000 feet (Fitzroy reaching 11,070 feet [3,375 metres]) north to latitude 46° S but average only 6,500–8,400 feet from latitude 46° to 41° S, except for Tronador (11,657 feet [3,554 metres]). The line of permanent snow is found at 2,300 feet in Tierra del Fuego, 5,000 feet at Volcán Osorno (41° S), and 12,000 feet in Domuyo (36° 38′ S). A line of active volcanoes—including Yate, Corcovado, and Macá—occurs about 40° S to 45° S. Gigantic ice fields are located between Monte Fitzroy and Lago Buenos Aires at both sides of Baker Fjord; the Viedma, Upsala, and other glaciers are born there. Other notable features are the more than 50 lakes found south of 39° S. Those depressions free of water form fertile valleys called *vegas*. The passes are low and easy, reaching 6,000 feet at Pino Hachado, 3,800 feet at Lipinza, and 2,000 feet at Mua Mum.

Peaks in the Patagonian Andes

On both sides, but more especially on the western side, the forests are untamed and magnificent, covering the mountains as high as the snow line, although at the higher altitudes toward the north the vegetation is smaller and less dense. Both Argentina and Chile have created national parks to preserve the area's natural beauty.

North of Lago Aluminé the axis of the cordillera shifts to the east up to a zone of transition between parallels 37° and 35° S, where the geographic aspect and geologic structure change. This marks the northern end of the Patagonian Andes.

The Central Andes. At latitude 35° S the cordillera undergoes a sharp change of character. Its width increases to about 50 miles, and it becomes arid and higher; the passes, too, are higher and more difficult. Glaciers are rare and very high, and there is virtually no vegetation. Lakes disappear north of latitude 39° S, forests, north of 37°. The main range serves as the boundary between Chile and Argentina and also is the divide between rivers flowing to the Pacific and the Atlantic. The last of the southern series of volcanoes, Tupungato (22,309 feet [6,800 metres]) is just east of Santiago, Chile. From there to the peak of Aconcagua, great snowcapped peaks, such as Juncal, rear up. The mighty Aconcagua itself, at 22,834 feet (6,960 metres), is the highest peak in the Western Hemisphere; to its north lies Mercedario (22,206 feet [6,770 metres]), and between them are the passes of Espinacito (16,000 feet) and Patos (12,825 feet). To the south the passes are Pircas (16,961 feet), Bermejo (over 10,000 feet), and Iglesia (13,400 feet). Farther north the passes are more numerous and higher, and the peaks of Bonete, Ojos del Salado, Incahuasi, and Pissis surpass 20,000 feet. The snow line is also much higher.

The peak of Tres Cruces (20,853 feet [6,356 metres]) at latitude 27° S marks the culmination of this part of the cordillera. To the north there is a transverse depression and the beginning of the high plateau called Puna in Argentina, Puna de Atacama in Chile, and Altiplano in Bolivia. The cordillera grows wider as it advances into Bolivia and Peru, where the great plateau is bounded by two ranges: the Occidental and the Oriental.

The beginning of the Altiplano

Northward, to latitude 18° S, the great peaks of El Cóndor, Sierra Nevada, Llullaillaco, Galán, and Antofalla all exceed 19,000 feet. The two main ranges and several volcanic secondary chains enclose depressions called *salares* because of the deposits of salts they contain; Sierra de Calalaste encompasses the big *salares* of Antofalla. Volcanoes of this zone occur mostly on a northerly line along the Cordillera Occidental as far as the Misti (latitude 16° S) in Peru.

At about 18° S the Cordillera Occidental changes to a northwesterly direction to the west side, descending smoothly to the deserts of Atacama and Antofagasta and the southern Peruvian coast.

The eastern range, the Cordillera Oriental, lower and built on a broad bed of lava, is cut and denuded by rivers with steep gradients, fed by heavy rainfall. It has two sections. The southern is 150 miles wide and—with the exception of Chorolque (18,414 feet [5,614 metres])—not very high. The northern, named Cordillera Real, is narrow, with higher peaks and glaciers; and the most important peaks, at over 21,000 feet, are Illimani and Illampu (Nevado de Sorata).

At about latitude 22° S, the Cordillera Oriental penetrates into Bolivia, describing a wide semicircle to the north; the plateau becomes the broadest part of the Andes—the Altiplano de Bolivia.

The Altiplano is one of the largest interior basins of the world—500 miles long, 80 miles wide, and 11,200 to 12,800 feet high; it has no water outlet to the ocean. In about the middle of the plateau there is a great depression, running south to north. Lake Titicaca, the highest navigable lake of the world (110 miles long), is on the northernmost part of the depression; the Río Desaguadero follows the depression, taking the water of Titicaca to the small Lago de Poopó.

The Peruvian Andes. The Cordillera Occidental runs parallel to the coast of Peru, while the Cordillera Real from Bolivia ends in the rough mountain mass of Nudo de Vilcanota at latitude 15° S. From there, two lofty and narrow chains emerge northward, the cordilleras of Carabaya and Vilcanota, separated by a deep gorge; to the west of these and northwest from the city of Cuzco a third range appears, the Cordillera de Vilcabamba. The three ranges are products of erosive action of rivers that have cut deep canyons. West of the Vilcabambas, the Río Apurímac runs in one of the deepest canyons of the hemisphere. In the valley west of the Cordillera de Vilcanota stands the city of Cuzco (10,909 feet [3,326 metres]).

Canyons and passes of the Peruvian Andes

Traditionally, the Peruvian Andes have been described as three cordilleras, which join at Nudo de Vilcanota, Pasco, and Loja (Ecuador). Today, the region is considered as a simple plateau. The *nudos* (knots) are erosive products and hydrographic divides; the cordilleras stand on the plateau and were produced by crustal movements and erosion by the large rivers. The plateau in the south is 60 miles wide and 13,000 feet high; in the north it is narrower and lower (less than 12,500 feet).

The Nudo de Pasco is a large, high plateau. To the west it is bounded by the Cordillera de Huarochirí, on the west slope of which the Río Rímac rises in a cluster of lakes fed by glaciers and descends rapidly to the ocean (15,700 feet in 60 miles). The nearly 15,800-foot-high pass of Ticlio is used by a railroad. On the *nudo* there are many small lakes and ponds, the lake of Junín, about 20 miles long, being the largest.

North of the Nudo de Pasco three different ranges run on the plateau: the Cordilleras Occidental, Central, and Oriental. In the Cordillera Occidental at latitude 10° S a deep, narrow valley (Callejón de Huaylas) separates two ranges—Cordillera Blanca (eastern) and Cordillera Negra (western); the Río Santa runs between them and cuts Cordillera Negra to drain into the Pacific. Cordillera Blanca is a complex highland with permanently snowcapped peaks, some among the highest of the Andes (for example, Huascarán, which is just over 22,204 feet [6,768 metres]). The glaciers that rise here are very often broken off by earthquakes and rush down the slopes, demolishing settlements and vegetation alike. Cordillera Negra, so named because it is without snow, is lower. The two ranges join together at latitude 9° S. The Río Marañón, which runs northward between Cordilleras Occidental and Central at about latitude 6° S, changes its direction to the northeast, penetrating into the region of the *pongos* (narrow transverse gorges that cut the cordillera to reach the Amazon Basin). These include Pongo de Rentema (about one and one-fourth miles long, 200 feet wide), Mayo, Mayasito, Huarcaya, and, the most important, Pongo de Manseriche (seven miles long).

Cordilleras Blanca and Negra

Between Cordillera Central and Cordillera Oriental, the Río Huallaga runs in a deep gorge with few small valleys; it cuts the eastern cordillera in Pongo de Aguirre (latitude 6° S). Cordillera Oriental ends in the Amazon Basin at latitude 5° S.

The permanent snow line is 19,000 feet in Nevado Chanchami (about latitude 16° S) and declines to 15,000 feet in Cordillera Blanca and to 13,000 feet in Huascarán. Permanent snow disappears north of latitude 8° S, the puna ends and the so-called humid puna, or *jalca*, begins. Mountains become wider and smoother in appearance; vegetation changes to heathland and trees. The altitude diminishes, and passes are much lower, as at Cuello de Porcullo (7,000 feet).

The Ecuadorian Andes. A rough and eroded high mass of mountains named Nudo de Loja (latitude 4° S) marks the end of the Peruvian cordillera and the beginning of the Ecuadorian system. This is a large plateau running in a south to north direction and bordered by two chains of high and numerous volcanoes. To the west, in the Cordillera Occidental, geologically recent and relatively low, there is a line of 19 volcanoes, seven of them over 15,000 feet high.

Chains of volcanoes in the Ecuadorian Andes

The eastern border is Cordillera Central (formerly Cordillera Oriental), higher and older, with a line of 20 volcanoes. Five are over 16,000 feet.

The plateau in between (sometimes called Avenida de los Volcanes) is 8,000 feet high and 55 miles wide and is covered by volcanic rocks, andesites, and lava. In some places the accumulation of material forms *nudos* and secondary chains that divide the valley into several basins.

On the external slopes of the cordillera the rainfall is heavy, and vegetation flourishes. The high cordillera is cold and desolate, with peaks covered by mist; such regions are known as *páramos*. The central plateau is partially cultivated and supports the majority of the Ecuadorian population. In the transverse ranges there are some important peaks, including El Relado (13,642 feet [4,159 metres]), Imbabura (15,190 feet [4,630 metres]), and Ñaupán (14,852 feet [4,529 metres]).

A third cordillera has been discovered recently in the eastern jungle of Ecuador; this is the Cordillera Oriental. It seems to be a very old alluvial formation that has been divided by rivers and heavy rainfall into a number of mountain masses, isolated or forming irregular short chains, and covered by luxuriant forest, like the cordilleras of Guacamayo, Galeras, and Lumbaquí. The altitude does not exceed 7,900 feet, except at Cordilleras del Cóndor (13,000 feet) and Cerro Pax (11,000 feet).

The Cordillera Oriental

The Colombian cordilleras. North of the international boundary there is a group of high, snowcapped volcanoes (Azufral, Cumbal, Chiles) known as Nudo de Huaca. Farther north begins the great Macizo Central Colombiano, with the volcanic mass of Nudo de Pasto (latitude 1–2° N), which is the most important Colombian physiographic complex and the centre of the country's river system.

Three distinct ranges, the Cordilleras Occidental, Central, and Oriental, run northward. Cordillera Occidental, parallel to the coast and moderately high (12,800 feet near Antioquia), ends in the mountain complex of the Cordillera de Chocó.

Cordillera Central, or Cordillera del Quindío, is the highest (mean altitude almost 10,000 feet) but the shortest (400 miles) range of Colombian Andes; its last spurs disappear south of the confluence of Cauca and Magdalena (latitude 8° N). Most of the volcanoes of the zone are in this range, including Tolima (17,105 feet [5,215 metres]) and Ruiz (17,844 feet [5,439 metres]) among others; Nevado del Huila (16,896 feet [5,150 metres]) is also in the Cordillera Central. At about latitude 5° 80′ N, the range widens into the plateau of Antioquia, where Medellín stands.

Between Cordillera Central and Cordillera Occidental is a great depression, the Patía–Cauca Valley, divided into three longitudinal plains. The southernmost is the narrow valley of Patía that drains to the Pacific. The middle plain is the highest (8,200 feet) and constitutes the divide of the other two. The northern plain, the largest (15 miles wide and 125 miles long), is the valley of Cauca, which drains by the Rió Cauca to the Magdalena.

Cordillera Oriental has a slight northeast direction and is the widest and the longest of the system. The average altitude is 7,900 to 8,900 feet. North of parallel 3° N it widens, and after the depression of Uribe (6,130 feet) begins the Macizo de Sumapaz (altitude 10,000 to 13,000 feet), where the large Páramo (high plateau) de Sumapaz is located. The volcanoes Puracé, Picos de la Fragua, and Neiva stand south of this *páramo*, the first in the Cordillera Central, close to Popayán, the other two on the Cordillera Oriental. North of Sumapaz the range divides into two, enclosing a large plain 125 miles wide and 200 miles long, often interrupted by small transverse chains that form several peneplains, called *sabanas* (or savannas), commonly containing shallow lagoons and useless bogs. One-third of the population of Colombia lives there. The largest and most populated is the Sabana de Bogotá; other important *sabanas* are: Ubaté, Chiquinquirá, Tunja, and Sogamoso. East of Honda (latitude 5° N) the cordillera is a series of abrupt parallel chains running to the north-northeast: among them Sierra Nevada del Cocuy (18,000 feet) has snowed peaks. Farther north the central ranges end, but the external chains go on and diverge to the north and northeast; the western is Cordillera de Ocaña (6,500 feet), which on its northeastern side has the sierras of Motilones and Perijá; this is the boundary between Colombia and Venezuela and goes as far as parallel 11° N in the Península de la Guajira. The eastern chain bends to the east and penetrates to Venezuela at the Cordillera de Mérida.

Páramos and *sabanas*

The northern end of the Andes

Rivers on the Cordillera Oriental make their ways on the *páramos* and *sabanas* and cut the borders of the plateau through deep gorges; Salto de Tequendama of the Río Bogotá is an example.

On the border of the Caribbean, between Serranía de Perijá and Isla de la Aguja, stands Sierra Nevada de Santa Marta, an isolated triangular massif that rises abruptly from the coast to snowcapped peaks of 18,947 feet (5,775 metres). Geologically speaking, it does not belong to the Andes.

The Venezuelan Cordillera. The Andes of Venezuela are represented by Cordillera de Mérida (280 miles long, 50 to 90 miles wide and about 10,000 feet high); it is separated from Cordillera Oriental of Colombia by the depression of Táchira. It begins in Pico el Cobre, extends first in a northerly direction for 35 miles and then northeast, up to the depression of Barquisimeto, or Bajada de Lara, where it ends. The cordillera is a great uplifted axis (where erosion has uncovered granite and gneiss rocks but where the northwest and southeast sides remain covered by sediments) formed by numerous chains with snow-covered summits separated by longitudinal and transverse depressions—Sierras Tovar, Nevada, Santo Domingo, de la Culata, Trujillo, and others. The Cordillera de Mérida separates the hydrographic regions of the Orinoco and the Caribbean.

To the north, Sierra Falcón and Cordillera del Litoral (called in Venezuela Sistema Andino) do not belong to the Andes but rather to the Guiana system.

Geology. The eastern side of the Andes is in general older than the western. Paleozoic (Cambrian, Silurian, Devonian, and Carboniferous) beds formed from 280,000,000 to over 570,000,000 years ago are found in the eastern Argentinean and Bolivian Andes; Cordilleras Real and Vilcabamba have even more ancient Archean (late Precambrian) as well as Paleozoic rocks, related to the Brazilian Massif. In Peru and Ecuador, the eastern ranges are of ancient gneiss and schists, covered mostly by Mesozoic, or Secondary, sandstones (from about 65,000,000 to 225,000,000 years old). In Colombia, the Cordillera Central shows ancient crystalline rocks (Cambrian to Carboniferous) with intrusion of heat-altered rocks.

The western ranges and the plateaus are mainly of Mesozoic formation, Jurassic and Cretaceous. Rocks of Cretaceous age (up to 136,000,000 years old) are widely

distributed in the Patagonian Andes; in Chile there are Jurassic (up to 190,000,000 years old) and Cretaceous sediments. Puna de Atacama and the west ranges of Bolivia show Mesozoic beds with Cenozoic (Tertiary; 7,000,000 to 65,000,000 years old) volcanic material. In Peru, gray limestone covers large parts of the plateau and the Cordillera Occidental. In Ecuador the Mesozoic beds occur with deposits of volcanic material in the valley. In Colombia the western branch is mostly of Cretaceous beds, as is the Cordillera Oriental.

The entire western coast of South America is included in the great circum-Pacific belt of seismic activity (earthquakes and volcanoes). The great geologic dislocation and the folding of the cordillera started in Cretaceous (Late Mesozoic) time and continued during the Tertiary (early to mid-Cenozoic). During the Mesozoic, the western part of South America was a geosynclinal basin, a vast downwarp in the Earth's surface, covered by the Pacific. Periods of upliftings and subsidences during the Tertiary were accompanied by great volcanic effusions. In the Quaternary (Late Cenozoic), beginning some 2,500,000 years ago, the disturbances continued in a process that is still going on. This involves a continuous raising of the cordilleras and a sinking in the ocean of a great part of the coast, producing an unstable condition caused by the tremendous abrupt difference between the heights of the cordilleras and the great depths of the Pacific. In this geologically recent period, volcanic activity and glaciation produced a period of uniform summit levelling (peneplanation) and denudation by filling depressions with glacial deposits and volcanic material and eroding the overall mountain relief. The modern volcanic system is Quaternary; three main groups were formed: that of the Patagonian and Central Andes (Chile); that between Chile and Peru; and that between Ecuador and Colombia.

Volcanic and earthquake activity in the Quaternary

Soil types and drainage. The complex interchange between climate, geology, topography, and biology that determines soil types and their condition is deeply affected by altitude. In general, Andean soils are relatively young and suffer great erosion by water and winds because of the unimpeded steep gradients of the land.

In the deep south of the Patagonian Andes the formation of soils is difficult; the action of glaciers and winds has left almost-bare rocks in many places. Northerly peat bogs, podzols, and meadow soils, yellow-brown and red-brown, with thick horizons of humus, are found; drainage is poor. In the region of lakes, dark-brown soils occur. They are rich in organic material and well drained. North of latitude 45° S drainage becomes insufficient in surface as well as in profile. In the upper part of the cordillera, soils are formed directly on weathered rocks; in the lower zones reddish-brown soils with gravel and quartz are found; erosion is heavy.

North of parallel 37° S are found primary soils with water deficiency. In the Atacama Desert, where precipitation and vegetation are scarce, so is the organic material; desertic gray soils bearing mineral salts occur, heavily eroded. This type, with little differences, extends on Cordillera Occidental to north of Peru.

The plateau and the east side of the eastern cordilleras show characteristics of soil closely related to altitude, from Bolivia to Colombia.

Zonal soils. In the Andean *páramo* are found rough topography and embryonic soils black with organic material; drainage ranges from very poor to very rapid.

At altitudes between 6,000 and 12,000 feet, red, brown, and chernozem soils occur in soft slopes of valleys and plains. In more poorly drained areas, soils are relatively fertile, being permeable in some places with sandy horizon B (the main layer of the soil profile, extending from the surface down to the bedrock). These soils are economically the most important in Bolivia, Peru, and Ecuador. The *sabana* soils of Colombia are gray-brown, with claypan in certain levels, which makes them impermeable, of very poor drainage.

In the high parts of the mountains are very poor soils, superficial and stony, as represented by Andean rendzina (dark grayish-brown soils) and by brown-black soils. On the east side of the eastern cordilleras, descending to the Amazon Basin, there are thin, dark, humid soils without evolution, suffering great erosion because of heavy rainfall and the steep gradient of the slopes.

Intrazonal soils. These soils include humic clay and solonetz (dark alkaline soils) close to the lakes and lagoons. Also included in this group are soils formed from volcanic ash in Cordillera Occidental, south of latitude 14° S (Chile, Peru, and Bolivia).

Azonal soils. These are alluvial soils, of incomplete evolution and stratified without definite profile, and lithosols (a group of shallow soils consisting of imperfectly weathered rock fragments) that occupy the major part of the Andean massif. Their superficial layer is directly on rocks. They have incomplete profiles on rough slopes. Erosion is heavy. In Colombia, in slopes and gorges, sandy yellow-brown azonal soils are the base of the large coffee plantations.

Climate. In general, temperature increases from Patagonia to the Equator, but combinations of such factors as altitude, proximity to the sea, the cold Peru (Humboldt) Current, heavy or scarce rainfall, and defenses against wind determine an extensive variety of climates. The hottest rain forests, deserts, or deep gorges are often separated from the polar temperatures of the high puna by a few miles. At identical altitudes, however, there is no parallelism between external (to the Pacific or to the Amazon Basin) and internal slopes of the cordilleras; the external slopes are under the influence of either the ocean or the Amazon Basin.

The isotherm (line of equal average temperatures) 10° C (50° F) extends from the Patagonian Andes to Colombia. As mentioned above, the line of permanent snow varies greatly. It is 2,600 feet at the Strait of Magellan, 7,000 feet at 39° S, 20,000 feet at 27° S, 19,000 feet at 16° S, 16,000 feet at 9° S, and 15,000 feet at the tropical Andes.

Precipitation and temperature

Precipitation varies widely. There is heavy rainfall and solid precipitation south of parallel 38° S (180 inches at Lake Quillen). On the Central Andes of Chile and the Atacama Desert, rainfall is scarce, with some precipitation in summer (December, January, February). On the Altiplano de Bolivia and the Peruvian plateau and in the valleys of Ecuador and *sabanas* of Colombia, rainfall is seasonal and not heavy; in some places there is very little. It rains in small amounts on the west side of the Peruvian Cordillera Occidental, somewhat more in Ecuador and Colombia. On the east side of the Cordilleras Orientales (Amazon Basin side), rainfall is usually seasonal and heavy.

Temperature varies with altitude. Up to 4,900 and 8,200 feet the climate is subtropical and tropical, respectively: dry with hot temperatures during the day and mildly warm at night. At heights of between 8,200 and 11,500 feet it is dry, with mild temperatures; but there are big differences between night and day; this realm constitutes the most populated area of the Andes. At 11,500 to 14,800 feet it is dry and cold, with great differences between day and night and between sunshine and shadow, and it is freezing at night. At 13,500 to 15,700 feet (the puna), occurs the climate of the *páramo*, very cold and dry, with subfreezing temperatures. Above 15,700 feet on the peaks and high ridges, climate is polar with very low temperatures and ice-cold winds.

As in all other areas of the world, microclimates (extremely local developments) are widely varied because of the interplay of latitude, altitude, and other factors.

Dependence of life on altitude

Plants and animals. Life on the Andes depends upon altitude; within certain levels, however, the existence of vegetation is determined by climate, watering, and soil and of animal life by the food that is available. The permanent snow line is the upper limit of both. In the Andes, some plants and animals can live at any altitude, and others can only live at certain levels. Cats live only exceptionally above 13,000 feet; white-tailed mice usually do not stay lower than 13,000 feet and live up to 17,000 feet. The Peruvian camelids (llama, guanaco, alpaca, and vicuña) are animals of the Altiplano (11,200 to 12,800 feet), although they can live well at lower altitudes. The

condor is believed to fly up to 26,000 feet. Probably the low barometric pressures of high altitudes are less important for vegetation, but altitude imposes a number of climatic factors, such as temperature, wind, radiation, and dryness of atmosphere, that determine the kind of plants growing at the different parts of the Andes.

In general, the Andes can be divided into latitudinal bands, each with typical main vegetation and fauna; but latitude imposes differences between south and north; and proximity to the Pacific and to the Amazon Basin creates differences between the external and internal slopes of the Cordilleras Occidental and Oriental.

A zone at about parallel 35° S separates two different regions of the Andes. To the south, in the Patagonian Andes, the flora is austral (of southern aspect) instead of Andean. Magnificent forests of araucaria (a pine), oak, coigue (an evergreen used for thatching), chusquea, cypress, and larch occur.

Northerly characteristics are different. The Cordillera Occidental is very dry in the south, slightly humid (with moisture and scarce rainfall) in central and north Peru, and humid with heavy or moderate rainfall in Ecuador and Colombia. Vegetation follows the climatic scheme: in the south, it is poor and desertic; at higher altitudes, steppe vegetation occurs. Animals include the guemul, puma, vizcacha, cuy (guinea pig), chinchilla, camelids, mice, and lizards; birds include the condor, partridge, parina, huallata, and coot. Agriculture is poor. The east side of the Cordilleras Orientales northward from Bolivia has lush vegetation, most of it natural tropical forest; agriculture is difficult because of the steepness of the slopes, but the rich fauna of the jungle is present.

Flora and fauna of the plateau

On the plateau (valleys, plains, ranges, and internal slopes of the cordilleras), life is again closely related to altitude. Tropical palms and eternal snows lie within a few miles of each other; altitude varies from 1,600 feet in deep gorges to more than 20,000 feet in peaks and ridges. Up to 8,000 feet, vegetation is of dry tropical and subtropical climate, and agriculture is important: the great coffee industry of Colombia is mainly located in the warm valleys of this zone. Between 8,200 and 11,500 feet lies the most populated zone of the Andes; some of the major cities of the Andean countries are there, and it supports the main part of Andean agriculture. Temperatures vary from warm in the valleys to moderate low (down to 50° F [10° C]) on the plains, *sabanas*, and slopes; and there is seasonal rainfall and water from rivers. This zone is also suitable for livestock and poultry.

At 11,500 to 13,400 feet relief is usually rough and difficult for agriculture; in Colombia this zone is *páramo* and sub-*páramo*, with seasonal rainfall; in Ecuador rain is abundant; in Peru and Bolivia the *páramo* has from moderate to scarce rainfall.

At 13,400 to 15,700 feet (puna) vegetation is formed by plants that resist the cold temperature and the freezing during the night; above 16,000 feet, vegetation is almost absent.

HUMAN ECOLOGY

The inhabitants. Man's presence on the Andes is relatively recent; the oldest human bones found have been 10,000 to 12,000 years old, although human habitation may be much older. The lack of oxygen due to the altitude, especially above 12,000 feet, is so physiologically disturbing that it imposes deep adaptative changes even in the cells of the body. The highest altitude in the Andes at which man has been found as a permanent resident is 17,100 feet (shepherds in the south of Peru) and, as a temporary worker, 18,500 to 19,000 feet (Carrasco Mine, in Aucanquilcha, Atacama Desert, Chile).

From Patagonia to the southern limits of the Bolivian Altiplano, the Andes are unpopulated; a few small groups of shepherds and farmers live in the lower slopes and *vegas* of the cordillera.

Northward, the Andes hold the largest part of the population and the majority of the most important cities of Bolivia, Peru, Ecuador, and Colombia. The main populated area is localized between 8,200 and 11,500 feet. In Colombia 11,500 feet is the upper limit of human life, and above 11,500 feet population is sparse; whereas in Peru and Bolivia, most of the inhabitants live above 10,000 feet (4,000,000 Peruvians live between 10,000 and 13,000 feet, and 800,000 live above 13,000 feet).

Indians and mestizos

Half of the population of Bolivia are Aymara and Quechua Indians; the other half are mestizos (or cholos, mixed); in the Lake Titicaca district live the small and primitive group of the Urus. Population is mainly distributed between the high *páramos*, where, except for a semi-nomad population of shepherds, the principal occupation is mining, and lower narrow valleys, where the people engage in agriculture.

In Peru, mining is the most important human activity above 11,500 feet, but the great part of the Andean population is dedicated to agriculture and raising sheep, goats, llamas, and alpacas; more and more people are engaged in industry and commerce. In the south, around Lake Titicaca, there is a group of Aymara Indians; but the largest native population is Quechua.

The inhabitants of the Andes of Ecuador are mainly Quechuas and mestizos; in the south there are small groups of Cañaris and, in the north, of Aymara descendants (the Salasacas). Agriculture is the main occupation; some native groups are dedicated to ceramics and weaving.

In Colombia the largest part of the population lives between 5,000 and 10,500 feet. Only 7 percent are native Indians: a nucleus of Chibchas reside on the Altiplano of the Cordillera Oriental; Paeces and Guambianos on the Cordillera Central; Pastos and Quillacingas on the southern mountains. At about 3,000 to 6,500 feet, the zone of coffee plantations is the most densely populated (259 inhabitants per square mile). Agriculture is the main occupation, followed by cattle raising.

Transport. The Andes are a formidable barrier for communication, with great effects on the economical and cultural development of the region. Production centres are, in general, far away from seaports, and the mountainous character of the land makes the construction and maintenance of railways and roads difficult and expensive. Most of the railways have been built to transport mining products. Agriculture requires roads because valleys are small and widely separated, making railways too expensive.

Pack animals

There is still a large use of pack trails between small communities and between farms and markets. Horses, donkeys, and mules are widely used; in Colombia the ox and in Peru and Bolivia the llama are also transport animals.

Railways are few and the internal systems little developed. There are two international railways between Chile and Argentina: the first, Los Andes–Mendoza, through the pass of Juncal, of narrow gauge, was opened in 1910 and connects Valparaíso and Buenos Aires; the second, Antofagasta–Salta (the former in Chile, the latter in Argentina), was opened in 1948.

Bolivia is connected with its neighbours by four railways: La Paz–Buenos Aires, crossing the Puna from Villagra to La Quiaca; La Paz–Oruro–Antofagasta, through Portezuelo de Ascotán; La Paz–Arica, crossing a point 14,000 feet high; and La Paz–Lake Titicaca, which continues on to Arequipa and Matarani (Peru).

In Peru there are two internal but important railways: the Puno–Cuzco and the Lima–Cerro de Pasco and Huancavelica, the highest in the world, crossing Ticlio at an altitude of 15,797 feet (4,815 metres).

The main railroad in Ecuador is the Quito–Guayaquil and, in Colombia, the Ferrocarril del Atlántico, from Bogotá to the Caribbean coast, opened in 1961.

The construction of roads on the Andes is very active; thousands of miles are opened every year. The Pan-American Highway is built in the mountains in Colombia and Ecuador. (Locally known as the Carretera Bolivariana, it connects Caracas, Bogotá, and Quito.) Various routes in Peru, Bolivia, Argentina, and Chile are included in the system.

Air transport is also important; it is especially developed in Colombia due to the difficulties of overland communication.

Economic resources. *Agriculture and livestock.* Agricultural operations on the Andes are difficult, and the results are relatively poor. The water supply is inadequate; a large part of the plateau is dry or receives little and irregular seasonal rainfall. Temperatures of the high plains are too cold, and crops are subject to freezing. The terrain is rough, and soils are not well developed; and where fertile valleys do occur, they are narrow and small.

Agricultural patterns

Some agricultural products of the Andes, however, can be grown in sufficient quantity to be exported: the main items are coffee (especially Colombia), cacao, coca, tobacco, cotton, and cinchona (quinine). There are possibilities of increasing the amount of arable land area by irrigation, but not to a great extent. The east side of the eastern cordilleras, facing the Amazon Basin, is in fact being developed, although its steep and rugged nature makes development difficult.

The natural pastures of the plateau are put to good use for cattle raising, and possibilities exist for further development of this activity. Colombia exports cattle, and Peru has a good milk-canning and livestock industry. Sheep, goat, llama, and alpaca raising are widespread in Peru and Bolivia, with both countries exporting sheep and alpaca wool.

Mining. The mining industry of the Andes is one of the most important of the world. Mining is especially extensive in the south. The industry includes mining for copper in Chile and Peru; tin in Bolivia; silver, lead, and zinc in Bolivia and Peru; gold in Peru, Ecuador, and Colombia; platinum and emeralds in Colombia; bismuth in Bolivia; vanadium in Peru; and coal and iron in Chile, Peru, and Colombia. Petroleum is extensively distributed along the whole eastern side of the Andes.

FUTURE PROSPECTS

The development of the Andean countries is linked to the existence of the Cordillera de los Andes, both because of its resources (whether potential or in use) and because of the fact that it constitutes a tremendous obstacle for transportation and communication.

Colombia, Ecuador, Peru, Bolivia, and Chile signed an agreement in 1969 known as the Andean Pact (Pacto Andino) or Cartagena Pact (Pacto de Cartagena), which aims at establishing economic and cultural unity. Among specific measures aimed at are the abolition of customs barriers, the joint exploitation of resources, and the coordination of industrial development. Agrarian reform is another matter of concern.

Other possibilities for future development lie in the use of the rivers, which in the mountains can scarcely be used for either irrigation or transportation because most of them run through steep gorges. In the early 1970s Peru was planning to use the big mountain rivers flowing eastward to the Amazon Basin to irrigate its desert coast by diverting the flow to the west by drilling tunnels. Bolivia, Peru, Colombia, and Ecuador either had already established, or else were constructing, large hydroelectric power plants that utilized the steep gradient of the Andean rivers.

BIBLIOGRAPHY. Among the modern reference sources, two kinds of works may be noted. There are highly specialized studies, such as HANS KINZL and ERWIN SCHNEIDER, *Cordillera Blanca, Perú* (1950), perhaps the most interesting scientific work on the Central Andes of Peru; HUMBERTO FUENZALIDA VILLEGAS, "El geosinclinal andino y el geosinclinal de Magallanes," *University of Chile, School of Geology*, no. 5 (1964); and the work of GUSTAV STEINMANN, *Geologie von Perú* (1929), who has studied the Peruvian Andes but whose book covers the whole mountain system. H.P. MOON, *The Geology and Physiography of the Altiplano of Peru and Bolivia* (1939), is still very important. There also are broad general studies the coverage of which includes treatment of biogeography and population: HUMBERTO FUENZALIDA VILLEGAS, "Orografía," "Clima," "Hidrografía," "Suelos," "Biogeografía," in *Geografía económica de Chile*, vol. 1, ch. 2, 4–7 (1965), provides good coverage of that portion of the Andes; PIERRE DENIS, *Amérique du Sud: les pays andins* (1927; Spanish trans., 1928), is a standard work on the Andes of Venezuela; FEDERICO A. DAUS, *Geografía de la República Argentina*, 10th ed., 2 vol. (1960); and CARLOS PENAHERRERA DEL AGUILA, *Geografía general del Perú*, vol. 1 (1969), provides interesting overall views of those portions of the Andes in the countries named. Among 19th- and early-20th-century studies, the wide coverage given in ISAIAH BOWMAN, *The Andes of Southern Peru* (1916, reprinted 1968), is justly renowned. Also of significance are MARTIN CONWAY, *The Bolivian Andes* (1901), by the climber of Illimani; ALFRED HETTNER, *Die Kordillere von Bogotá* (1892; Spanish trans., *La Cordillera de Bogotá*, 1892, reprinted 1966), which gives a summary of his studies on the Central Mountain System of Colombia; EDWARD WHYMPER, *Travels Amongst the Great Andes of the Equator* (1892; ed. by F.S. SMYTH, 1949), by a botanist and Alpinist who was the first to climb the volcano Chimborazo in Ecuador; and TEODORO WOLF, *Geografía y Geología del Ecuador* (1892), a splendid description of the high sierras of that country.

(M.T.V.)

Andhra Pradesh

Andhra Pradesh is one of the 21 states of the Republic of India. Located in the eastern part of the Indian subcontinent, it has an area of 106,272 square miles (275,244 square kilometres) and a population of about 43,000,000; it is the fifth largest state in area and the fifth largest in population in the Indian Union. It is bounded by the states of Tamil Nadu (formerly Madras) to the south, Mysore to the west, and Mahārāshtra, Madhya Pradesh, and Orissa to the north; to the east it is bounded by the Bay of Bengal. The state capital is Hyderābād, which has a population of about 1,600,000.

The state derives its name from the Andhra people, who have inhabited the area since antiquity, and who have developed their own language, Telugu. The state in its present form came into existence in 1956 as a result of the demand of the Telugu-speaking Andhras—who form the largest linguistic group in India, with the exception of the Hindi-speaking peoples—for a separate state; at that time they numbered more than 30,000,000. Although the state is primarily agricultural, there is some mining and a small amount of industry.

History. Though Sanskrit writings dating to around 1000 BC contain references to a people called "Andhras" living south of the central Indian mountain ranges, historical evidence of the Andhras dates from the times of the Mauryan dynasty, which ruled in the north in the 3rd century BC. The great Mauryan emperor Aśoka (*c.* 265–238 BC) sent Buddhist missions to the Andhras in the south. Around the 1st century AD, the Sātakarṇi, or Sātavāhanas, one of the most renowned of the Andhra dynasties, came to power. Its members ruled over almost the whole of the Deccan Peninsula and even established trade relations with Rome. They were patrons of diverse religions and were also great builders: their principal city, Amarāvatī, contained Buddhist monuments inaugurating a new style of architecture. Experts ascribe parts of the famous paintings in the Ajantā caves of the Deccan to the Andhra painters of that period. Buddhism prospered under the Andhras and in their capital flourished the great Buddhist university of antiquity, where Nāgārjuna (*c.* AD 150–250), the founder of the Mahāyāna school of Buddhism, taught. The ruins of the university at Nāgārjunakoṇḍa still attest to its past glory. In the 11th century, during the reign of the eastern Cālukyas, a dynasty who unified most of the Andhra area, the first of the Telugu poets, Nannaya, began translating the Sanskrit epic, the *Mahābhārata*, into Telugu, thus inaugurating the birth of Telugu as a literary medium. The dynasty of the Kākatīyas of Warangal in the 12th and 13th centuries extended Andhra power militarily and culturally; it was during their regime that the commercial expansion of the Andhras toward Southeast Asia reached its peak. By this time, however, the Muslims had established themselves in the north, and their invasion of the south led to the fall of Warangal in 1323. But the rise of the kingdom of Vijayanagar, to the southwest of Warangal, arrested the expansion of the Muslim power for some time. Acclaimed as the greatest kingdom in Andhra history and as one of the greatest in Indian history, Vijayanagar, under the rule of its great king Kṛṣṇa Deva Rāya, who reigned from 1509 to 1529, became synonymous with military glory, economic pros-

The mission from Aśoka

perity, good administration, and artistic splendour. During this period, Telugu literature flourished. The formation of an alliance between the various neighbouring Muslim principalities, however, led to the fall of Vijayanagar in 1565, leaving the Muslims in control of the Andhra areas.

Arrival of the British and French

In the 17th century British and French traders began to involve themselves in Indian politics. As a result, successive Niẓāms (rulers) of Hyderābād, seeking to consolidate their kingdom against rivals, obtained first French and later British support. In exchange for their help, the British acquired from the ruling Niẓām the coastal Andhra districts lying to the north of Madras and later the hinterland districts. Thus, the major part of the Andhra country came under British rule. Part of the Telugu-speaking areas, known as the Telingana region, remained under the Niẓām's dominion. The French, for their part, acquired a few towns in the Andhra country.

During the 19th century, which witnessed the rise of Indian nationalism, the Andhras came to the forefront of the movement. Leaders such as Kandukuri Veerasalingam pioneered in social reform. In the struggle against British rule, Andhra leaders played decisive roles. Pride in their historical and linguistic achievements led them to demand a separate province. Simultaneously, a movement was also organized to unite the Telugu-speaking peoples living under British rule and those under the Niẓām's administration. The Andhras' demand for separate statehood became so insistent that when the central government refused to accede to the demand, a local leader, Potti Sreeramulu, fasted to death in 1953 to dramatize the issue. The government finally acceded to the people's request by creating in 1953 the Andhra state, which included the Telugu-speaking districts of the former Madras state, thus paving the way for the formation of linguistic states throughout India in 1957. The erstwhile state of Hyderābād, which had joined independent India in 1949, was split up and its nine Telugu-speaking districts (Telingana) were joined to the Andhra state in 1956 to form the new state of Andhra Pradesh.

Demand for Andhran statehood

The landscape. The state has three main physiographic regions: the coastal plain to the east, extending from the Bay of Bengal to the mountain ranges; the mountain ranges themselves, the Eastern Ghāts, which form the western flank of the coastal plain; and the plateau to the west of the Ghāts.

The coastal plain runs almost the entire length of the state and is watered by several rivers, flowing from west to east through the hills into the Bay. The deltas formed by the most important of these streams—the Godāvari and the Krishna—comprise the central part of the plains, an area of fertile alluvial soil.

The Eastern Ghāts, which do not form a continuous range, being broken by the great river valleys, are part of a larger mountain system extending from central India to the far south and running parallel to the east coast. They have highly porous soils on their flanks.

The plateau to the west of the ranges is composed of gneissic rock (a foliated rock formed deep in the earth under conditions of heat and pressure); it has an average elevation of 1,600 feet above sea level. As the result of erosion, this is a region of graded valleys, with red, sandy soil and isolated hills, which support a thorny vegetation. Black soil is found in certain parts of the area.

Coastal vegetation

Vegetation and animal life. Near the coastal plains mangrove swamps and palm trees are found, while the deltas constitute a rich agricultural belt where food crops, fruit, and tobacco are grown. The soil on the plateau and the uncertain rainfall allow only dry crops to be raised, although with the extension of irrigation, rice and sugarcane are also being grown. Of the state's total area, 24 percent is forested. The forests consist of both moist deciduous and dry forest vegetation; teak, rosewood, bamboo, and wild fruit trees are found. Common trees such as the neem (a tree which produces an aromatic oil), banyan, mango, and peepal (a fig tree sacred to Buddhists) also occur, as well as flowering plants—jasmine, rose, and many local varieties. Animal life, apart from common domestic types (dogs, cats, and cattle), includes tigers, leopards, bears, and deer, which abound in the hills and forest areas.

Climate. A summer that lasts from March to June, a July-to-September season of tropical rains, and a winter from October to February constitute the three seasons. The range of maximum and minimum temperatures is from 74° F to 82° F (23° C to 28° C) and from 50° F to 53° F (10° C to 12° C), respectively. Summers are very warm in the coastal plains, with temperatures in places reaching 108° F (42° C). On the plateau, summers are cooler and winters colder. Rainfall is to a large extent caused by the southwest monsoon (rain-bearing) winds; some places receive a maximum of 55 inches of rain and others only 20 inches. In coastal areas, rainfall is heavy; in some areas on the plateau, especially in the north and west, it is sparse.

Population. The state's population at the time of the 1971 census was about 43,000,000. There are 98 women for every 100 men. The average density at the time of the 1971 census was 408 inhabitants per square mile, with wide variations in different districts; Hyderābād, for example, had a high concentration with 940 per square mile. The state's extensive forest areas contain a population of 1,300,000, including some small tribal groups such as the Thenchu.

Languages and religions

The state's cultural pattern is revealed by the distribution of its languages and religions. Telugu is spoken by 86 percent of the population, and about 7 percent speak Urdu, the language of the Muslim population, whose members are concentrated in the Telingana area. The remaining groups consist of people speaking border-area languages, including Tamil, Kannada, and tribal languages such as Lambadi. About 88 percent of the population are Hindus, 8 percent are Muslims, and 4 percent are Christians. The Christians live in urban centres and coastal areas, and the Muslims are concentrated in the west.

More than 8,000,000 people (17 percent) are urban dwellers and a third of these are concentrated in the industrial and manufacturing areas around the three main cities—Hyderābād, Vishākhapatnam, and Vijayawāda. With increasing industrial development, these cities are becoming linked with neighbouring urban areas, forming agglomerations. Some important industries and several smaller enterprises, such as sugar factories and service industries, are scattered in medium and smaller urban areas. There are 13 towns with populations of more than 100,000 each.

Literacy

The literacy rate is about 25 percent—33 percent among men but only 16 percent among women. In the Hyderābād, Krishna, and Godāvari districts literacy is higher than elsewhere.

Government and administration. Andhra Pradesh is a constituent unit in the Indian Union, and the state's constitution, which provides for a parliamentary form of government, is contained in the Indian Constitution itself. A governor, appointed by the federal government, is the executive head of the state administration, but the real power is in the hands of a cabinet of ministers responsible to the state legislature. The state has a bicameral legislature consisting of a Legislative Assembly of 287 members elected by adult suffrage from territorial constituencies and a Legislative Council indirectly elected by university graduates, local bodies, and members of the Legislative Assembly.

The administration is conducted by various departments, each under the direction of a minister, assisted by permanent civil servants. The State Secretariat at Hyderābād supervises the administration of the state's 21 districts. Local administration in each district is the responsibility of a district collector. Rural local government has been democratically decentralized by the introduction of a system in which local authorities operate at the village, block (a unit consisting of a group of villages), and district levels. Urban areas are governed by municipal bodies.

Features of local government

A special feature of state government is the regional

committee for Telingana (which consists of the nine Telugu-speaking regions that were formerly a part of Hyderābād); the duty of the committee is to ensure that the views of the people of Telingana are given adequate consideration. The committee was established when the region was joined to Andhra Pradesh in 1956, in order to protect Telingana interests, since the area was economically and educationally less advanced than the Andhra areas.

The state judiciary is headed by a High Court, located in Hyderābād; the High Court has original jurisdiction in some cases and exercises appellate and administrative control over the district and lower level courts. The High Court is itself subject to the appellate authority of the Supreme Court of India in certain matters. Within the state are located many federal instrumentalities, including the railways administration, the Income Tax Agency, and other establishments. A number of defense establishments are located at Hyderābād, and Vishākhapatnam is the headquarters of the East Coast Naval Command.

The economy. *Agriculture.* Agriculture is the primary sector of the state's economy, engaging 70 percent of the state's total work force. About 29 percent of its area is devoted to the production of food grains and 50 percent of the state's income is attributable to agriculture. The state produces over 4,500,000 tons of rice annually, making it the second-largest rice-producing area in the country. Andhra Pradesh is also known for such cash crops as tobacco and sugarcane, in addition to chilies, oil seeds, and pulses (leguminous crops, such as peas, beans, and lentils).

Principal crops

The state's rivers, particularly the Godāvari and the Krishna, account for its agricultural importance; for a long time their benefits were restricted to the coastal districts, which had better irrigation facilities. Since independence, great efforts have been made to tap the waters of these and other rivers for the benefit of dry regions, such as Rayalaseema and Telingana. A multipurpose project at Nāgarjuna Sagar, diverting the waters of the Krishna, was begun in 1955 to irrigate an additional 3,000,000 acres, thereby augmenting rice and sugarcane production. Half of the total investment for development is allotted to agricultural irrigation. The state's forest areas annually yield high quality timber and forest produce.

Minerals. Among the state's mineral resources are mica, manganese, barite, and high-grade coal. Low-grade iron ore is to be found in the southern parts of the state. Production at the once world-famous diamond mines, where the Koh-i-noor diamond and other famous stones were found, is being renewed. Quartz, limestone, and graphite are also found. For the exploitation of its mineral resources the state has established a mining corporation.

Industry. Andhra Pradesh is industrially underdeveloped. Since independence, industries such as shipbuilding, aeronautics, and plants producing electrical equipment, machine tools, and drugs have been established by the central government in the Vishākhapatnam and Hyderābād areas. Private enterprises include chemical, textile, and cigarette manufacturing, and light industry; many of these enterprises are located in the Vijayawāda-Guntūr region. Mining and manufacturing industries, however, account for less than 10 percent of the state's income. The state is asking for greater federal investment in industrialization, and the central government sanctioned the construction of a mammoth steel plant at Vishākhapatnam, where there is easy access to raw materials and port facilities. The phenomenal increase in recent years in power generated by hydroelectric and thermoelectric projects augurs well for industrialization and irrigation; installed capacity in the early 1970s amounted to 1,000,000 kilowatts. With its long coastline and many rivers, the state also has a modest but expanding fishing industry.

Federal investment in industrialization

Employment. The working force includes nearly 42 percent of the population, three-quarters of whom are engaged in agriculture. As much as 15 percent of the population, however, are agricultural labourers who are subject to seasonal unemployment. Labour migration to urban areas is common, but employment opportunities in the trade and industrial sectors are limited, absorbing only 5 percent of the population. Efforts are being made to find additional employment for the rural population by reviving cottage industries and locating light industries in rural areas.

Transportation. There are three civilian airports—at Hyderābād, Vijayawāda, and Vishākhapatnam. The state has over 5,000 miles of rail track. The rail system serves a major part of the state and connects it with other parts of India. There is an extensive road system of nearly 55,000 miles, of which 1,400 miles form part of the three national highways connecting all parts of the country. Other roads are maintained by state and local governments. Bus transportation is mostly nationalized and offers facilities for express travel between various cities.

The river canals in coastal areas, especially the saltwater Buckingham Canal running parallel to the coast from the Krishna River to Madras, are used for cargo transportation. Vishākhapatnam is a major seaport used by international shipping. A large open harbour is under construction there.

Social conditions. *Health.* Medical and health facilities, under the supervision of the Director of Medical and Health Services, have expanded rapidly in recent years. Under the Primary Health Centres scheme, medical help, both curative and preventive, is being brought to rural areas. An increase in the number of in-patient dispensaries and in the number of beds available in hospitals has augmented urban medical facilities. The state's principal hospitals, such as the Osmania Hospital at Hyderābād and the King George Hospital at Vishākhapatnam, for example, have been improved; and new institutes, including those for treating particular diseases, have been opened. There is also a family-planning program. Medical aid is free to low-income groups, and several medical-insurance schemes cover various categories of employees. Medical education has been expanded, with the number of medical colleges increasing to 14 by 1970.

Education. The state's educational system provides for ten years of schooling followed by a two-year junior college course leading to undergraduate and postgraduate education. Primary education has been free and compulsory since 1961. High school education, primarily the concern of local bodies, is also free. In the decade of the 1960s, expenditures on secondary education more than doubled. Privately run high schools also exist, mostly in urban areas. There are more than 120 junior colleges. The state has three universities—at Hyderābād, Waltair, and Tirupati—and an agricultural university at Hyderābād. Affiliated to each university are a number of colleges, totalling nearly 200 altogether. Postgraduate instruction and research facilities are available at the three university centres as well as at other postgraduate centres. Technical education is receiving special attention in order to meet the demands of industrialization. Various industrial-training institutes provide for vocational training, while the engineering colleges of the three universities train advanced technical personnel.

Higher education

Social welfare. Before independence, social welfare work was mainly undertaken by private agencies. But the magnitude of the problem and scarcity of resources, both organizational and financial, led the government to accept primary responsibility in this field. Public investment in social welfare amounts to more than three-quarters of all that is spent on planning. There are social-welfare programs for disabled and handicapped persons, scheduled castes (formerly called "untouchables"), and tribal or other groups that are not fully integrated into the social structure. Social programs include those for education, economic advancement, health, and housing. Scholarship programs are available for all scheduled caste, tribal, and other disadvantaged groups in all educational institutions that receive substantial financial assistance from state and federal agencies. In addition

to the policy of reserving places for these groups in educational institutions and in employment, other ameliorative measures include housing schemes and land-distribution schemes. Progress toward social advancement has, nevertheless, been rather slow. A separate government department for women's welfare exists. There are also many private social organizations, including the Andhra Mahila Sabha, an organization that for 30 years has worked for the cause of women's welfare.

Cultural life and institutions. The Andhras' contribution to India's cultural heritage is substantial. Their legacy in art and architecture has already been noted. The *kuchipudy* style of dance is unique in the Indian tradition, while South Indian music has derived much from Andhra roots. Many of India's major composers are Andhras, and Telugu is the language of most of the compositions. Telugu, one of the four literary languages of the Dravidian family, occupies a prestigious place among Indian languages, being renowned for its antiquity as well as for its mellifluous quality. Andhras are prominent in the modern Indian literary renaissance, and their writing reflects the impact of the contemporary revolution in literary forms and expression. There are more than 25 daily newspapers in the state and many periodicals in English, Telugu, and Urdu. Muslim culture in the Telingana region further enriches the state's cultural diversity.

Before independence, arts and literature thrived mostly under the sponsorship of private organizations, many of which still function. The state has also created three autonomous academies to revive, popularize, and promote fine arts, dance, drama, music, and literature.

Folk culture

The conscious pursuit of culture is more an urban than a rural phenomenon, for cultural performances, literary meetings, and religious discussions occur mostly in towns or cities. In rural areas, folk culture predominates. Cultural evolution in different parts of the state under different historical circumstances resulted in the occurrence of recognizable variations in dialect, in the caste structure, and in habits, thus diversifying the folk arts. Folk media such as minstrel-ballad performances, puppet shows, and mythological storytelling, are indigenous to the area; use of these media in social and political communication is also common.

The penetration of the mass media, especially of radio, to rural parts has helped to bring an awareness of classical culture to the rural folk, and of folk arts to the urban population. There are four broadcasting stations—located at Hyderābād, Vijayawāda, Vishākhapatnam, and Cuddapah. Andhra Pradesh is the second largest movie-making state in the country.

BIBLIOGRAPHY. The GOVERNMENT OF INDIA, *Census of India—1961*, vol. 2 (1966), contains information on population, urbanization, and industrialization trends in the state. The GOVERNMENT OF ANDHRA PRADESH, *Fourth Five-Year Plan* (1965), provides information on the state's economy. V. JAGANNADHAM, *Social Welfare Organisation* (1967), and the Central Social Welfare Board's monthly organ, *Social Welfare*, contain material on the organization and work of the social welfare programs in the state. The *Andhra Pradesh Almanac* (annual) includes very useful information to the general reader on various aspects of the state.

(R.V.R.C.R.)

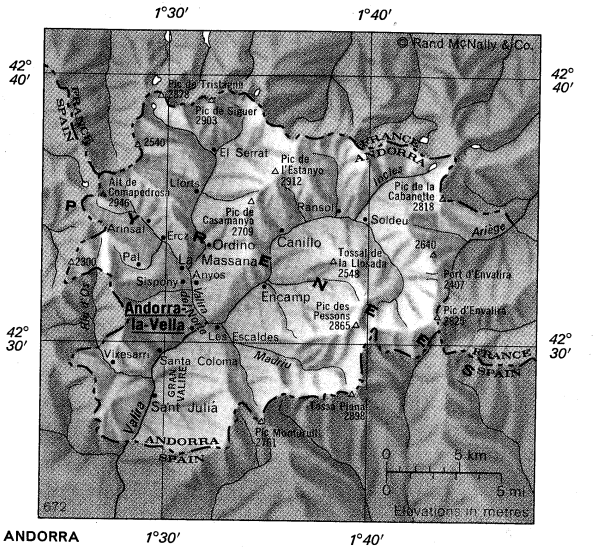
Andorra

An autonomous principality for more than 1,100 years, Andorra (officially, in Catalan, Principat d'Andorra) is the largest of the small states of Europe, which include Monaco, San Marino, and Liechtenstein. Its 179 square miles (464 square kilometres) comprise a landscape of high peaks and deep valleys on the southern slopes of the Pyrenees, bounded on the south and west by Spain and on the north and east by France.

Historical background

Andorra's survival as an independent entity since the Dark Ages of European history is among its most interesting aspects. Its independence is ascribed traditionally to the grant of the lands of the counts of Urgel, in Spain, which was made by a son of Charlemagne, Louis the Pious, to the Bishop of Urgel in AD 819. Dual allegiance to co-princes, one in Spain and one in France, originated in the late 13th century, and, with interruption only during the French Revolution, Andorra has been governed jointly by the Bishop and the French head of state, each of whom inherit jurisdiction with their office.

Since the 1950s, however, Andorra has been affected greatly by a profound change from its traditional pastoral and farming economy to one based on commerce and tourism. Immigration has nearly quadrupled its population, to more than 20,000, and the face of its communities and countryside has been altered to accommodate the new attentions from and to the outside world.



The landscape. Altitudes in Andorra's typically Pyrenean valley-and-peak landscape range from about 3,000 to 9,600 feet (about 900 to 2,900 metres). The valley of the Valira River bisects the country, splitting into a Y shape at the capital, Andorra-la-Vella (from Spanish Andorra la Vieja, Andorra the Old) and dividing Andorra into three natural regions: the valleys of eastern Valira, of northern Valira, and of the Gran Valira.

The climate is generally dry, with heavy rainfalls in spring and autumn and abundant snow that remains for several months in the highlands. Temperature varies greatly in relation to altitude. Climatic conditions favour the growth of pine and birch forests on the mountains. The flora is varied; on the great natural pastures grow rhododendron, many varieties of lilies, and, notably, the narcissus. The fauna, which is less varied, consists principally of wild boars, deer, partridge, and hare, while in the rivers trout abound.

In the highlands are situated the small villages that are typical of Andorra, with their houses of granite, wood, and slate. In the larger villages in the valleys, the recent economic growth has left its imprint in the form of modern buildings that contrast sharply with the traditional architecture. The largest villages are Andorra-la-Vella and Les Escaldes, with its hydroelectric station and sulfur springs.

The people. Although written records of Andorra exist from as early as the 3rd century BC, prehistoric remnants attest to human occupation of the valleys since far earlier times. Until the advent of modern road systems, the population had been isolated and protected in the inaccessible valleys since the invasions of the Muslims in the 8th century AD and their subsequent expulsion by Charlemagne, in 803.

The present population comprises about 30 percent Andorrans native by descent, 60 percent Spanish, and 5 percent French, with the remainder members of some 35 various nationalities. This blending, together with the continual influx of tourists, gives Andorra an international character.

Historic, traditional, and geographic factors lend the principality a strong affinity with the region of Catalonia.

Andorra, Area and Population

	area		population	
	sq mi	sq km	1954 census	1971 census
Municipalities (*comuns*)				
Andorra (Andorra-la-Vella)	23	59	...	14,000
Canillo	47	121	...	400
Encamp	29	74	...	2,000
La Massana	24	61	...	1,000
Ordino	34	89	...	400
Sant Juliá	23	60	...	3,000
Total Andorra	179*	464	6,000	21,000†

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

in northern Spain. Andorra's official language is Catalan, its institutions are based in Catalonian law, and a large proportion of the Spanish immigrants are Catalan. The religion of the majority of Andorrans is Roman Catholicism, and the principality is part of the diocese of Seo de Urgel (See of Urgel).

Economy and transportation. With only about 4 percent of the land cultivable, the traditional economy revolved about the pasturing of sheep and the harvesting of timber and tobacco. Industry is limited to processing the latter two products and to exploiting the considerable hydroelectric resources.

Impact of commerce and tourism

In the modern economy, commerce is encouraged by favourable customs agreements with France and Spain, by the absence of personal income taxes, and by the economic disparity between Spain and more industrialized nations—which accounts in large measure for the heavy Spanish immigration. Above all, Andorra has become an international marketplace for the goods of many of the heavily industrialized nations of the world. Tourism has grown through the attraction of the mountains, the marketplace, and the excellent facilities for winter sports. By the early 1970s, hotel accommodations had increased more than fivefold since 1950. The economic expansion has prompted the establishment of several banks associated with French and Spanish investment firms. No national monetary unit exists, however, and both French and Spanish currencies are used.

The physical geography of Andorra prohibits air transportation. No railway system exists, but good roads link Andorra with France and Spain and permit access by automobile to the most secluded spots even when there is snow.

Government and services. Andorra's co-allegiance to the Bishop of Urgel and the president of France is acknowledged by annual payment of a token tribute, or *questia*, alternately to the Bishop of Urgel and the president of France.

Delegation of governmental powers

The co-princes represent Andorra internationally and exercise jointly all legislative, executive, and judicial powers. These powers are, in actuality, granted to permanent delegates and judges, or *veguers*. The *veguers* are primarily responsible for public order; and the police, the principality's only armed force, are under their jurisdiction. The *veguers*, together with a judge of appeal appointed by the co-princes, form the Tribunal of Courts (Tribunal de Corts), the supreme court for criminal justice. Civil justice is administered, in the first instance, by two civil judges appointed by the *veguers*. The country is divided into six municipalities, or *comuns*, governed by communal boards under the direction of a *consol* elected by the people. The Council General of the Valleys is the supreme administrative body but has no legislative power. It comprises 24 councillors, four elected from each *comú*. In 1971 women were permitted to vote for the first time in Andorran elections.

Such public services as education, postal services, and telegraph are provided by the Spanish and French governments. Others are the responsibility of the Council General: telephone service, fire brigade, and social security. In 1969 a mandatory social insurance went into effect, covering illness, disability, and old-age pensions. Poverty is nonexistent.

Cultural life. Andorra's social and legal traditions are recorded in the *Manual Digest* (1748) of usages and the *Politar andorrà* (1763), which serve as references in legal proceedings and for cultural history.

Splendid samples of pre-Romanesque and Romanesque religious architecture remain, notably in the Lombardian style. There are also remarkable Romanesque paintings attributed to the Master of Santa Coloma and the Master of Andorra (12th and 13th centuries). Altarpieces and altar decorations known as retables and triptychs remain from the Gothic era. Popular music and dances are usually connected with the celebration of village feasts and religious ceremonies.

There is no Andorran press as such, but the monthly *Andorra Magazine* was established in 1970 in Les Escaldes. Two radio stations are in operation, and televised programs from Spain and France can be received.

Prospects. Andorra is a historical curiosity. Despite the transformations that the other European countries have undergone over the centuries, Andorra has managed to preserve its mountain tranquillity and its traditions as well as its independence and viable political system since feudal times. Tourism appears to offer Andorra prosperity through the foreseeable future, but the country faces two major challenges: the consolidation of its economic expansion on an autonomous basis (above all by continuing to improve tourist facilities) and the evolution of certain public institutions—*e.g.*, the judicial and legislative systems and the lack of autonomy in education and communications—that have become inadequate because of the country's growth.

BIBLIOGRAPHY. J.M. GUILERA, *Una Història d'Andorra* (1960), is an objective general history; while B. RIBERAYGUA ARGELICH, *Los Valles de Andorra, recopilación histórica* (1949), is a scholarly compilation useful for reference on specific topics. Social and economic aspects are covered in J.M. VIDAL Y GUITART, *Instituciones políticas y sociales de Andorra* (1949); and "Les Problèmes actuels des Vallées d'Andorra," *Publications de l'Institut d'études politiques de Toulouse* (1970). A. PUIGOIOL, *Andorra, el meu país* (1963), presents the main facts about Andorra in a popular style. J.A. BRUTAILS, *La Coutume d'Andorre* (1965), covers customs and ways of life. The strong influence of geography on Andorran life is well elucidated in S. LLOBET, *El medio y la vida en Andorra* (1947). Works in English include YOUNGMAN CARTER, *On to Andorra* (1963); *Spain, Portugal, Andorra, Gibraltar* in "The Illustrated Library of the World and Its Peoples" (1963); and CHARLES L. SHERMAN, *The Five Little Countries of Europe* (1958).

(N.Ma.)

Andrea del Sarto

The reputation of the Florentine painter and draftsman Andrea del Sarto has not yet recovered from the discovery, made in the 19th century, that the conventional characterization of his works—"faultless"—could be used as a term of reproach. His drawings are eloquent witness to only one ambition: to make each picture more perfect in terms of composition and craftsmanship than the last. Until the patient pursuit of perfection is thought, once more, to be a proper ideal for a naturally gifted artist, his importance is best presented as that of a great teacher. Among his pupils and followers were most of the significant Florentine painters of the first half of the 16th century—Rosso Fiorentino, Pontormo, Francesco Salviati, and Giorgio Vasari, for example—and it is very largely through his example that the tradition of Florentine art was transmitted (but transformed) through to the end of the Renaissance and was able to embrace the stylistic innovations made around 1500 by Leonardo da Vinci and Michelangelo.

The suspicion that, while he was worthy and sound, Sarto nevertheless lacked some essential creative fire can be traced back to malicious accounts written by Giorgio Vasari, a 16th-century biographer and artist for whom Sarto was a disgrace to the profession. Vasari's specific complaint was that Sarto lacked ambition, and in Vasari's terms the complaint was probably true, for the oppor-

tunities for worldly success that Raphael enjoyed could have been his, too, but he did not take them.

Early life and influences

The painter's name was Andrea d'Agnolo (family name probably Lanfranchi), and he was born in Florence on July 16, 1486; his family had been craftsmen and tradesmen, and his father was a tailor (hence "del Sarto"; Italian *sarto*, "tailor"). Little of real interest is known about his life, probably because it was for the most part uneventful. He was notably short in stature and known to his friends as Andreino. With two brief exceptions, his working life was spent in Florence. It is possible that Sarto was apprenticed to a goldsmith and a wood-carver before settling with the painter Raffaellino del Garbo soon after 1500. Raffaellino was a respectable and, in these years, still competitive representative of the traditions of late-15th-century painting, but his influence not unnaturally disappeared rather quickly as Sarto looked about him at the works of Leonardo, Michelangelo, and Fra Bartolommeo. He began to produce independent work about 1506—not precociously. Almost immediately he began a long association with a new church and convent of SS. Annunziata (for which he did frescoes in 1509–14 and 1525), and he moved to a workshop near it in or about 1511. There, for five or six years, he shared the experiences and sometimes commissions of a major sculptor, Jacopo Sansovino, which led him to an increasingly and, in the end, exceptionally solidly structured style. These were the years in which Il Rosso and Pontormo were his pupils, and it may fairly be said that around 1513–14 the leadership in Florentine painting passed from Fra Bartolommeo's workshop to his.

In 1517 or 1518 Sarto married Lucrezia del Fede, a widow whom he had, according to her testimony, used as a model for several years; she brought him property and a useful dowry. In 1518 he was summoned by the king of France, Francis I, to Fontainebleau, where he was preceded by a reputation based upon pictures made for export. It is unlikely that he found the life of a court artist congenial, and he remained for a year or less without beginning any major commission. Soon after his return, his connections with the Medici family (powerful since their return to Florence from exile in 1512) led to the most significant contract of his career—for part of the

Decoration of the Villa Medici

decoration of the Villa Medici at Poggio a Caiano, near Florence. The patron was in fact the Pope, Leo X, whom Sarto almost certainly visited in Rome in 1519–20; but the project, the only one that ever offered Florentine artists the scope that Raphael had in the Vatican Palace, collapsed when the Pope died in December 1521. Sarto's fresco "Tribute to Caesar" is a fragment now incorporated into a much later scheme of decoration.

In 1520 Sarto began to build himself a house in Florence, which was later inhabited and modified by several other painters; it was a substantial property without being a palace. By 1523 he had a manservant as well as apprentices. Throughout his life he was content to work, when it suited him, for nominal fees, for no remuneration at all, or for only part of a fee offered to him, probably because he was in comfortable circumstances. He would paint for a carpenter or a king. A plague in 1523–24 drove Sarto and his wife to seek security in the Mugello, a valley north of Florence, but the interruption was brief. After the expulsion of the Medici, once more, in 1527, he worked for the republican government of Florence. His "Sacrifice of Isaac," intended as a political present to Francis I, was painted in this period. After the siege of Florence by imperial and papal forces, he succumbed to a new wave of plague and died in his house on September 28, 1530; he was buried in SS. Annunziata.

Importance of the Scalzo frescoes

Andrea del Sarto's most striking monument is the grisaille (gray monochrome) series of frescoes on the life of St. John the Baptist in the Chiostro dello Scalzo, Florence. Begun about 1511, the work was not completed until 1526, and almost all of it was painted by his own hand, so that it reads like an artistic autobiography covering the greater part of his career. His portraits of his wife, Lucrezia (*c.* 1513–14, Prado, Madrid; and *c.* 1522, Staatliche Museen Preussischer Kulturbesitz, Berlin), can be supplemented by many others disguised as Madonnas (*e.g.*, the celebrated "Madonna of the Harpies") just as his self-portraits in the Uffizi and in the National Gallery of Scotland at Edinburgh (both *c.* 1528) can be extended by several others, more or less hidden in his paintings from 1511 onward. A badly damaged pair of circular portraits of Andrea and Lucrezia at the Art Institute of Chicago appear to be signed (toward 1530).

Sarto's style is marked throughout his career by an interest, exceptional among Florentines, in effects of colour and atmosphere and by sophisticated informality and natural expression of emotion. In his early works such as the "Marriage of St. Catherine," the search for the expression of animation and emotion led to an ecstatic and nonidealistic style that proved immensely attractive to a younger generation of painters. Restraint increasing with maturity did not inhibit the achievement of such passionate later works as the "Pietà" (*c.* 1520), but the mood is always intimate and never rhetorical. In the 1520s, his style, as a result of the influence of Michelangelo or of artistic events in Rome, became perceptibly more ideal, more polished, and approximates what may properly be called a grand manner in the last of the Scalzo frescoes, the "Birth of the Baptist" (1526). From first to last, Sarto's integrity as a craftsman, his sheer professionalism, is impressively consistent; and it is characteristic of him that he refused to have his works engraved. His real quality is most vividly revealed, however, where there were no rewards of publicity, in his drawings.

MAJOR WORKS

"Madonna" (*c.* 1508; Galleria Nazionale, Rome); "Five Episodes from the Life of S. Filippo Benizzi" (1509–10; SS. Annunziata, Florence); "Noli Me Tangere" (1510; Uffizi, Florence); "The History of St. John the Baptist" (fresco series; *c.* 1511–26; Chiostro dello Scalzo, Florence); "Baptism of Christ" (*c.* 1511; Chiostro dello Scalzo, Florence); "The Archangel Raphael, Tobias, St. Leonard, and a Donor" (1511; Kunsthistorisches Museum, Vienna); "Procession of the Magi" (1511; SS. Annunziata, Florence); "Annunciation" (1512; Pitti Palace, Florence); "Marriage of St. Catherine" (1512–13; Gemäldegalerie, Dresden); "Holy Family" (*c.* 1513; Hermitage, Leningrad); "Birth of the Virgin" (1513–14; SS. Annunziata, Florence); "Portrait of a Young Woman" ("Lucrezia del Fede"; 1513–14; Prado, Madrid); "Christ the Redeemer" (*c.* 1515; SS. Annunziata, Florence); "Holy Family" (1515; Louvre, Paris); "The Stories of Joseph the Jew" (1515–16; Pitti Palace, Florence); "Madonna with Child and St. John" (1516; Borghese Gallery, Rome); "The Virgin and Child with St. John the Baptist and Two Angels" (*c.* 1517; Wallace Collection, London); "Madonna of the Harpies" (1517; Uffizi, Florence); "Portrait of a Young Man" (1517–18; National Gallery, London); "Disputation on the Trinity" (1517–18; Pitti Palace, Florence); "Holy Family" (1517–18; Louvre, Paris); "Charity" (1518; Louvre, Paris); "Pietà" (*c.* 1520; Kunsthistorisches Museum, Vienna); "Madonna della Scala" (1522; Prado, Madrid); "Assumption of the Virgin" (1522–23; Pitti Palace, Florence); "Pietà" (1524; Pitti Palace, Florence); "The Last Supper" (completed 1527; S. Salvi, Florence); "Sacrifice of Abraham" (*c.* 1527; Cleveland Museum of Art); "Holy Family with the Infant St. John" (*c.* 1527; Metropolitan Museum of Art, New York); "Madonna with Six Saints" (1528; Pitti Palace, Florence); "Charity" (1528; National Gallery of Art, Washington, D.C.); "Abraham's Offering" (1528; Gemäldegalerie, Dresden); "Self-Portrait" (*c.* 1528; National Gallery of Scotland, Edinburgh); "The Sacrifice of Isaac, Prisoner of an Angel" (*c.* 1529; Prado, Madrid); "Portraits of the Artist and His Wife, Lucrezia" (*c.* 1530; Art Institute of Chicago).

BIBLIOGRAPHY. SYDNEY J. FREEDBERG, *Andrea del Sarto*, 2 vol. (1963); and JOHN K. SHEARMAN, *Andrea del Sarto*, 2 vol. (1965), are well-illustrated monographs containing references to earlier writings; the selected bibliography in the former is especially useful. Each contains a catalogue raisonné of paintings, the second has in addition a separate catalog of drawings and register of documents. There are significant differences of opinion as to attributions, chronology, and general assessment, but at present it seems to be generally believed that the truth lies somewhere between them.

(J.K.Sh.)

Anesthetic

Anesthetics are used to cause insensitivity to pain during surgical operations. Although anesthetics have similar effects in that all depress the central nervous system, they

have varied physical and chemical properties; in addition, their mechanisms of action generally are not yet well understood. Anesthetics are general and local, or regional, in their effects. General anesthetics, which cause a loss of consciousness that occurs in several stages, usually are administered in combinations consisting of several specific drugs so that insensitivity to pain (analgesia), loss of consciousness, and muscle relaxation are effective. General anesthetics include inhalation anesthetics—*i.e.*, gases such as nitrous oxide and cyclopropane and volatile liquids such as diethyl ether and halothane. In addition, short-acting barbiturates may be injected into a vein to cause general anesthesia. Neuromuscular blockers, or muscle relaxants, include curare. Local or regional anesthetics are used to anesthetize only the region of the body on which surgery is to be performed; lidocaine hydrochloride (Xylocaine) is a commonly used local anesthetic.

Historical background. After the first demonstration in 1846 of the safe and effective use of diethyl ether as an anesthetic during surgery in man, it was suggested that anesthesia, derived from a Greek word meaning lack of feeling, be used to describe the phenomenon. For centuries before the 1840s, relief from the pain of crudely performed operations had been provided by various means—*e.g.*, deliberate asphyxiation, application of pressure on nerves, freezing of the part to be operated on, inhalation of fumes from extracts of narcotic plants, or alcoholic stupefaction. About 1800 the Englishman Sir Humphry Davy noted the anesthetic effects of nitrous oxide on himself; within a short time, carbon dioxide was used to depress animals. The successful use of anesthesia during surgery was the result of the work of several doctors in the U.S. between 1842 and 1847. In 1842 Crawford W. Long of Georgia removed a small tumour from the neck of a youth who was under the influence of ether; in 1844 a Connecticut dentist gave a partially successful demonstration of the use of nitrous oxide for tooth extraction. In mid-October of 1846 at the Massachusetts General Hospital in Boston, another dentist, William T.G. Morton, administered diethyl ether to a young man as a doctor removed a tumour from the jaw of the unconscious patient. Several months later, an Englishman, John Snow, became the first physician to devote full time to the practice of anesthesia; he administered not only ether but also chloroform, which was given to women to make them insensitive to the pain of childbirth.

After the introduction of anesthetics, physicians were able to perform surgery more slowly and carefully on patients insensitive to pain; at the same time, physicians began to associate certain diseases with specific pathological changes in organs and, because anesthesia was available, to suggest their surgical removal. That general anesthesia is hazardous, however, was shown by sudden deaths, which occasionally occurred during chloroform anesthesia. The discovery in 1884 that cocaine produced insensitivity to pain if applied to the eye, therefore, was followed rapidly by the development of local anesthesia. Administration of anesthesia now is performed safely by physicians who have specialized in this branch of medicine. During surgery, complicated apparatus is required not only to administer anesthetics but also to observe, measure, and treat adverse changes in circulation, respiration, kidney function, and blood that may result from the anesthetic and the surgical procedures.

Uptake and mechanism of action of anesthetics. When a gas such as nitrous oxide or a vapour such as ether is inhaled, it quickly reaches the air cells or alveoli in the lungs. A certain amount of the anesthetic, depending on its solubility and concentration, dissolves in blood passing through the lungs and is carried in the bloodstream to the brain. The amount of anesthetic circulating with the blood produces several distinct stages of anesthesia. The first stage is analgesia, an absence of the sense of pain; the second stage is mental or physical excitement. In the third stage, called the surgical stage, complete loss of consciousness is accompanied by muscle relaxation.

Stages of anesthesia

If too much anesthetic is administered and the blood concentration is excessive, respiration and circulation may become depressed, or stop entirely, as the anesthetic affects specific brain centres. It is necessary, therefore, to know the level or depth of anesthesia at all times; if it is found that inhalation anesthesia, for example, has produced a deep state of unconsciousness, the concentration of anesthetic administered to a patient is decreased until the excess anesthetic is expired through the lungs and its concentration in the blood decreases. If a solution of an anesthetic such as a barbiturate is injected into a vein (intravenous administration), the anesthetic reaches the brain by way of the bloodstream, as does an inhaled anesthetic; recovery from injected anesthesia, however, depends on its rate of removal from the brain by the blood and also the rate either of its transformation in the liver or of its excretion by the kidneys in urine.

The exact mechanism by which anesthetics affect the activity of the central nervous system (brain and spinal cord) has not yet been established with certainty. One effect of anesthetics is that they interfere with transmission of nerve impulses and alter electric activity; in fact, anesthesia can be produced by applying alternating current to a region of the skull over a specific area of the brain (electronarcosis). The unconsciousness produced by anesthetics has been attributed to several factors: changes in certain properties of nerve cell membranes; suppression of the activity of certain enzymes (organic catalysts); and solubility of the anesthetic in lipids, which comprise part of a nerve cell membrane. It has also been proposed that anesthetics interact with water molecules in the brain to form very small hydrated crystals; such crystals could influence physical factors involved either in the transmission of an impulse along a nerve cell or at a synapse, the point at which an impulse passes from one nerve cell to another. In the hindbrain, an effect of anesthetics is to block the transmission of nerve impulses in pathways that carry pain sensations and in those that regulate sleep and arousal.

Toxicity of anesthetics. A distinction between anesthesia and coma is that the loss of consciousness produced by an anesthetic is completely reversible if it is administered properly; a coma that results from poison or disease may not be reversible. In general, anesthetics do not cause extensive tissue damage, although specific ones may have adverse effects. Chloroform may damage the liver and depress the circulatory system; because these effects occur with relative frequency, chloroform seldom is used. Other halogenated anesthetics (*i.e.*, those that contain chlorine, bromine, or fluorine) also show this tendency; repeated administration of halothane, for example, may cause liver damage, while methoxyflurane may affect the kidney tubules in a way that results in excessive loss of water from the body. After prolonged exposure to nitrous oxide, a general toxic effect—failure of the bone marrow to manufacture white blood cells—may occur. Malformation of a fetus may occur if a woman is exposed to anesthesia early in a pregnancy. Some of the toxic effects of anesthetics probably result from their transformation in the body to form more toxic compounds.

General anesthetics. There has been a trend away from the use of anesthetics such as ether and cyclopropane, largely because they are flammable and have undesirable side effects. Chloroform has been replaced by a series of nonflammable halogenated volatile liquids. Intravenous anesthetics with actions on specific brain regions, especially those that affect consciousness, the sensation of pain, and muscle relaxation, are continuously sought in research concerned with anesthetics.

Inhalation anesthetics. Both gases and volatile liquids used as inhalation anesthetics should have certain characteristics. Volatile anesthetics, liquids that are extremely potent and relatively soluble in blood, should vaporize readily and should have an effective vapour pressure and an appropriate solubility in blood and tissues. All inhalation anesthetics should be stable, inert, nonflammable compounds that are easy to synthesize and identify. They should be sufficiently potent so that high concentrations are not required to produce unconsciousness. In addition,

inhalation anesthetics should cause minimal toxicity, have an acceptable odour, and be free from side effects such as nausea and vomiting. No inhalation anesthetic has yet been developed that fully meets these criteria.

Nitrous oxide

Nitrous oxide is the only inorganic compound used as an anesthetic. A sweet-smelling, nonflammable gas that provides excellent relief from pain, nitrous oxide lacks potency; as a result, high concentrations of the gas are necessary to produce anesthesia. Ethylene, a gas similar to nitrous oxide but more potent, is no longer used because it is flammable. Although nitrous oxide causes rapid loss of consciousness, a patient may remain in the second stage of anesthesia, agitated or emotionally uninhibited, crying or laughing—hence its common name, laughing gas. Nitrous oxide is administered in combination with intravenous anesthetics (barbiturates) that produce unconsciousness and with neuromuscular blockers that relax muscles. Most general anesthetics are administered in this way, with one agent (*e.g.*, barbiturate) responsible for loss of consciousness, another (*e.g.*, nitrous oxide) providing analgesia, and a third (*e.g.*, curare) causing muscle relaxation.

Diethyl ether (ether) is a volatile, flammable liquid of restricted anesthetic use because of its several undesirable characteristics; *e.g.*, it has a suffocating and obnoxious odour, may cause nausea and vomiting after surgery, and does not cause immediate loss of consciousness. Ether, however, is potent, provides good analgesia and muscle relaxation, and improves circulation in early stages of anesthesia. A volatile liquid with a pungent odour, divinyl ether (Vinethene) is a seldom used anesthetic that was developed to combine the favourable properties of ethylene and ether. After this potent, flammable substance is administered, however, alterations in the time or force of the heartbeat may occur. Cyclopropane, introduced in 1934, is a potent, flammable gas with an odour of garlic. Since it forms an explosive mixture with oxygen, cyclopropane must be administered in a closed system; a patient rebreathes the gas from a reservoir bag after it has been freed of carbon dioxide by passage through soda lime. Both loss and recovery of consciousness occur rapidly, but patients often become excited as they awaken.

Halothane and methoxyflurane

Halothane (Fluothane) was developed in the 1950s to incorporate most of the attributes of an ideal anesthetic. Since breathing the potent, nonflammable vapour causes circulatory and respiratory depression, the liquid is vaporized in special devices to provide safe concentrations. Although breathing may be affected, the tubes (bronchi) that lead to the lungs are not constricted, as with cyclopropane, so that halothane is safe for asthmatic patients. In addition, dilation of blood vessels allows adequate circulation to tissues even though there is a decrease in blood pressure. If deep anesthesia is required, loss of consciousness and analgesia are produced with an injection of a barbiturate and inhalation of a mixture of nitrous oxide, oxygen, and halothane; muscles are relaxed by using neuromuscular blocking drugs.

Methoxyflurane (Penthrane) is a potent, nonflammable, volatile liquid with a fruitlike odour. Loss of consciousness occurs slowly because the vapour tends to dissolve in the rubber apparatus from which the patient inhales the anesthetic. Because it has a high solubility in body tissues, methoxyflurane vapour is retained in the body for a prolonged period of time, resulting in slow recovery of consciousness. Analgesia and muscle relaxation are obtained readily with low concentrations, however, so that respiratory and circulatory depression can be avoided. In addition to its possible toxic effects on kidney tubules, methoxyflurane may also cause liver damage. Fluroxene (Fluoromar), a vapour of low flammability, combines properties of both halothane and methoxyflurane but is less potent.

Ethyl chloride is a highly flammable, volatile liquid of great potency; it readily depresses circulation and causes heart irregularities. When ethyl chloride is applied to the skin it evaporates rapidly and freezes the tissues, so that minor surgery (*e.g.*, incision of an abscess) can be performed. In addition to its use in inducing general anesthesia, ethyl chloride sometimes is sprayed on the skin as a treatment for muscle pain and soreness.

Intravenous anesthetics. Intravenous injection of an anesthetic is a more rapid way to produce unconsciousness than inhalation of a gas or vapour. In the 1930s, barbituric acid derivatives were developed that enter the brain from the blood much more quickly than barbiturates commonly taken by mouth to induce sleep (see SEDATIVE-HYPNOTIC DRUGS). After injection into the blood, a short-acting barbiturate circulates to the brain and causes loss of consciousness within seconds. Consciousness is regained when the barbiturate leaves the brain and circulates to other parts of the body. Because they are stored in fatty tissues and only slowly eliminated from them, short-acting barbiturates are present in low concentrations in the blood stream for a prolonged period of time; as a result, sedation persists in the form of a hangover. Unfortunately, short-acting barbiturates do not influence the sensation of pain or provide muscle relaxation; therefore, they are used in combination with an inhalation anesthetic, usually nitrous oxide, and a neuromuscular blocking drug. Three short-acting barbiturates commonly used in intravenous anesthesia are thiopental sodium, thiamylal, and methohexital sodium.

Short-acting barbiturates

Neuromuscular blocking agents. Curare is a name for a variety of poisons that were used for hundreds of years by hunters in South America to paralyze prey. In 1942, a preparation of a purified curare poison was given intravenously for the first time to relax the muscles of a patient before surgery. Known collectively as neuromuscular blockers, these drugs interfere with the transmission of nerve impulses from nerve to muscle; they paralyze voluntary muscles (*e.g.*, those of arms and legs) but not cardiac muscle or smooth muscles of the intestines and blood vessels. The neuromuscular blockers are of two kinds: those that compete with, and therefore block, the action of acetylcholine, a compound that ordinarily is released by nerves to initiate muscle contraction, and those that mimic the action of acetylcholine by affecting a specific region of a muscle membrane so that a sustained action resembling a state of paralysis, rather than a normal muscle contraction of short duration, occurs.

Curare

Gallamine triethiodide and d-tubocurarine chloride compete with acetylcholine and therefore are called competitive neuromuscular blockers; they are effective from 20 to 30 minutes, or until their action is terminated by removal from the muscle to the bloodstream and distribution to other tissues. Although some destruction of these drugs occurs in the liver, the majority of each is excreted in urine. The action of these drugs also can be stopped by injection of another drug, neostigmine, which allows acetylcholine to accumulate so that normal muscle activity can occur. Succinylcholine chloride causes a sustained effect that lasts three to eight minutes; the action ends when the drug is broken down by an enzyme called pseudocholinesterase that is present in blood. Several factors (*e.g.*, body temperature) influence the extent and duration of paralysis produced by neuromuscular blockers.

Local anesthetics. Local or regional anesthetics are applied to body surfaces and injected near nerves in order to prevent sensations of pain associated with surgery, injury, or disease. Local anesthesia may be preferred to general anesthesia because the patient usually is conscious and therefore cooperative during surgery and requires less nursing care after the surgery has been performed. Local or regional anesthetics interfere with chemical and physical changes associated with initiation and transmission of nerve impulses. Reactions (*e.g.*, loss of consciousness, heart action) to a local anesthetic may occur if large amounts of it are absorbed into the bloodstream from the point of injection; these reactions usually can be avoided, however, if known toxic limits of an anesthetic are not exceeded during administration. In addition, the hormone epinephrine may be added to the anesthetic solution to constrict blood vessels around the site of the injection; as a result, the anesthetic enters the

bloodstream more slowly. Other toxic reactions to local anesthetics include allergic responses, local irritation, and tissue destruction at the site of injection.

Criteria for an acceptable local anesthetic include complete reversibility of action, lack of irritation upon injection; minimal toxic effects, rapid onset and appropriate duration of action, ease of sterilization without loss of activity, and easy synthesis at low cost. Local anesthetics may be divided into two classes. In one class are drugs, such as cocaine, procaine hydrochloride (Novocaine), and tetracaine hydrochloride, that enter the bloodstream and are detoxified there by pseudocholinesterase. The second class includes dibucaine hydrochloride (Nupercaine), lidocaine hydrochloride (Xylocaine), and mepivacaine (Carbocaine). After these drugs enter the bloodstream, they may be detoxified in part in the liver or excreted unchanged in urine.

Types of local anesthetics

The application of certain local anesthetics to the eye or to mucous membranes of nose, mouth, throat, trachea, or esophagus is known as topical anesthesia. Nerve endings in these regions are easily anesthetized. A topical anesthetic usually is used, for example, prior to removal of a foreign body from the eye. Injection of a local anesthetic beneath the skin, called infiltration anesthesia, is used prior to surgical removal of skin tumours or suturing of lacerations. In nerve block anesthesia, an anesthetic is injected into a nerve to anesthetize a small region of the body; *e.g.*, injection of the ulnar nerve at the elbow anesthetizes the little finger, while injection of an anesthetic into the mandibular nerve of the lower jaw allows painless tooth extraction. Injection of anesthetics into larger nerve groups anesthetizes a larger part of the body; *e.g.*, injection at the armpit or just above the collar bone in the neck, called brachial plexus anesthesia, anesthetizes the entire arm. In addition, the closer to the spinal cord an anesthetic is injected, the more widespread is the anesthesia obtained. A local anesthetic solution injected between the vertebrae around the spinal cord membranes (peridural anesthesia) or directly into the cerebrospinal fluid in the lumbar region of the spine (spinal anesthesia) permits painless surgery to be performed on the legs, abdominal wall, or abdomen.

Procaine hydrochloride

Cocaine, an alkaloid compound found in the plant *Erythroxylum coca*, was the first anesthetic used for several types of regional anesthesia; *i.e.*, topical, infiltration, nerve block, and spinal anesthesia. Cocaine still is used for topical anesthesia of the nose and throat; its blood-vessel constricting action helps to produce a bloodless surgical area. Because it is addictive, difficult to sterilize, and causes a high incidence of toxic reactions, however, cocaine has been replaced to a great extent by procaine hydrochloride. First used in 1905, procaine hydrochloride is not a topical anesthetic, but its use became widespread because it is less toxic, easier to synthesize and sterilize, and has a shorter duration of action than cocaine. Epinephrine prolongs procaine action, which usually lasts about an hour. Tetracaine hydrochloride is a potent anesthetic used for topical and spinal anesthesia; its effects last one and a half to two hours. Dibucaine hydrochloride is used largely to produce topical anesthesia of mucous membranes and to treat, among others, the pain of sore throat. Two similar anesthetics are lidocaine hydrochloride and mepivacaine; both are used for topical and injection anesthesia. Mepivacaine has blood-vessel constricting properties, and both drugs are detoxified more slowly than is procaine hydrochloride. Lidocaine hydrochloride is a widely used local anesthetic. Deliberate intravenous injection is widely practiced to treat abnormal rhythms of the heart.

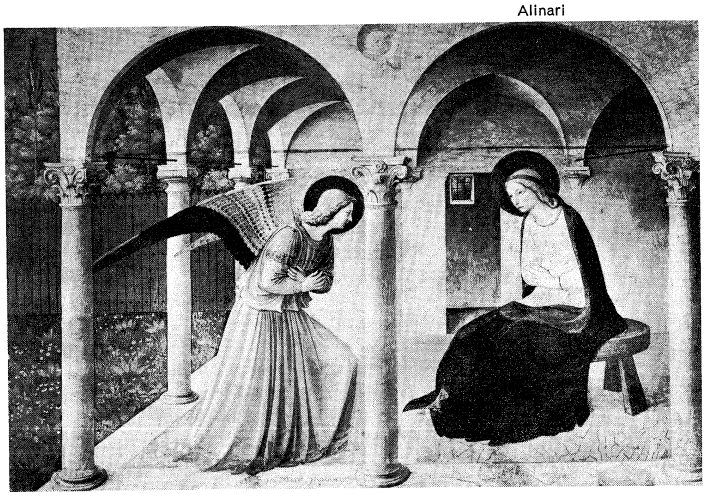
BIBLIOGRAPHY. J. ADRIANI (ed.), *Labat's Regional Anesthesia: Techniques and Clinical Applications*, 3rd ed. (1967), a clearly written and well illustrated text on regional anesthesia, written and revised by two men well known for their original contributions to the field; R.D. DRIPPS, J.E. ECKENHOFF, and L.D. VANDAM, *Introduction to Anesthesia*, 3rd ed. (1967), an informative and readable text recommended for the beginner in the field; T.E. KEYS, *The History of Surgical Anesthesia* (1945), a concise, authoritative, and readable account of the history of anesthesia, including general, regional, and intravenous methods; L.D. VANDAM, "Uptake and Transport of Anesthetics and Stages of Anesthesia," in J.R. DI PALMA (ed.), *Drill's Pharmacology in Medicine*, 3rd ed. (1965), a fairly readable scientific discussion of uptake and distribution of anesthetics in the body, with notes on the determination of neurological signs and stages of anesthesia; W.D. WYLIE and H.C. CHURCHILL-DAVIDSON, *A Practice of Anaesthesia*, 2nd ed. (1966), an authoritative source used by professionals in the field.

(L.D.V.)

Angelico, Fra

The 15th-century Italian painter Fra Angelico seems to reach into the poetic essence of things and re-create the tangible world in a profoundly religious light. His works, which reflect a strong classical influence, are now seen to belong within the framework of the Renaissance Florentine style, but they also embody such a serene religious attitude that critics once placed him among the artists of the preceding medieval period. From his own day to the present, appreciation of his works has fluctuated markedly, centring at some times on their religious qualities and at other times on their artistic qualities, which are equal to those of the greatest of European artists.

Alinari



"The Annunciation," fresco by Fra Angelico, 1438–45. In the Museo di San Marco, Florence.

S. Domenico period. He was born at Vicchio, near Florence, probably about 1400 (it has only recently been discovered that he was not born in 1387 or 1388, as the 16th-century biographer Giorgio Vasari wrote; this information has resulted in substantial changes in the chronology of his works). Baptized Guido di Pietro, he gained a reputation as a painter under this name by 1417. In that year he became associated with a miniaturist of the late Gothic tradition, Battista Sanguigni, who was later his assistant. Sometime between 1420 and 1422, he became a Dominican monk and resided in the monastery of S. Domenico at Fiesole, taking the name of Fra Giovanni da Fiesole. At Fiesole, he was probably influenced by the teachings of Giovanni Dominici, the militant leader of the reformed Dominicans; the writings of Dominici defended traditional spirituality against the onslaught of Humanism. Angelico was also influenced by his fellow monk St. Antoninus Pierozzi, who became the archbishop of Florence when Fra Angelico refused the post and who may have consolidated his faith and inspired some of his compositions.

Influence of Dominici and Pierozzi

Angelico was probably trained by the greatest painter and miniaturist of the Gothic tradition, Lorenzo Monaco, whose influence may be seen in the clear, painstaking delicacy of execution and the vibrant luminosity that seem to spiritualize the figures in Angelico's paintings. These qualities are apparent in two small altarpieces in the Museo di San Marco (or dell'Angelico) in Florence: one with the "Madonna of the Star" and the other with "The Annunciation."

Angelico's "Deposition" for Sta. Trinita in Florence was once attributed to Lorenzo Monaco, who had begun it before he died in 1425. Monaco had divided it into a triptych and executed the pinnacles. Angelico, however, made it a unified altarpiece with a vast landscape dominated by a varicoloured hill town. It is perhaps an imaginative evocation of Cortona, where Fra Angelico spent some time and where important works of his are to be found. Against that background are sharply outlined human figures in interconnected groups; their features are so delicately traced that attempts have been made to identify them as portraits. These arrangements of figures attest to Angelico's deep knowledge of the formalism that characterized the art of the early Renaissance.

Two strands were interwoven in Angelico's life at Fiesole: the pious life of a monk and continuous activity as a painter. Vasari described him as "saintly and excellent," and, not long after his death, he was called *angelico* ("angelic") because of his moral virtues. This subsequently became the name by which he is best known, often preceded by the word *beato* ("blessed").

Angelico did not remain absorbed in prayer in his monastery, however; he knew and followed closely the new artistic trends of his time, above all the representation of space by means of perspective. In works such as the large "Last Judgment" and "The Coronation of the Virgin" in the Uffizi, for example, the human figures receding toward the rear themselves create a feeling of space similar to that in the paintings of Angelico's great Florentine contemporary Masaccio. The earliest work by Angelico that can be dated with certainty is a triptych of huge dimensions that he painted for the linen merchants' guild (or Arte dei Linaiuoli; hence its name, the "Linaiuoli Altarpiece"); it is dated July 11, 1433. Enclosed in a marble shrine designed by the Florentine sculptor Lorenzo Ghiberti, this altarpiece represents the Virgin and son facing forward, monumentally, and, surrounding them in a minor key, charming angels, developing the motif of the "Madonna of the Star." The group has affinities with the Florentine Maestàs (*i.e.*, Madonna and child enthroned in majesty) of the 14th century, but the influence of Masaccio may be seen in the formalism of the construction, extending in a somewhat strained manner to the four saints painted on the two folding shutters. Angelico finished the work with a predella, or narrow strip of paintings at the bottom: these include "The Adoration of the Magi" and "The Martyrdom of St. Mark," which are lucid and compact in their narrative and have a strictly defined perspective, a technique that is even more effective in the small painting depicting the naming of John the Baptist.

Earliest known dated work

In 1436, Angelico was commissioned to paint an altarpiece for the Brotherhood of Sta. Maria della Croce al Tempio, which he completed by December of that year. In the serene "Lamentation," he executed figures in silent contemplation surrounding the dead Christ. Angelico was inspired by a famous painting by the 14th-century master Giottino (now in the Uffizi), but he expanded the subject to include more figures in a more complex arrangement, and he set them within a melancholy landscape, extending across the long walls of Jerusalem, with a leaden overhanging sky. In this painting Angelico included figures from sacred tradition and persons who had existed historically, probably to point up the historical continuity of devotion to the Redeemer.

Also in the 1430s, Angelico painted one of the most inspired works of the Florentine Renaissance, "The Annunciation," now in the Museo Diocesano of Cortona, an altarpiece significantly superior to his two other paintings on the same subject. It shows the Garden of Eden with Adam and Eve being driven out by the Angel, yet also under the sway of the radiant messenger and pure maiden who are portrayed in the space of a Renaissance-style portico. The predella is skillfully divided into stories of the Virgin Mary, naturalistically portrayed, especially the Visitation, which has a realistic panorama. Angelico always followed reality closely, even when he used a miniaturist technique. Occasionally, he resorted to medieval techniques, such as a gold background, in deference to the taste of those who commissioned the work, but his figures still emerge quite distinctly from the panels, in the Renaissance manner, revealing the painter's increasingly sure and harmonious pictorial idiom. Angelico's "Annalena Altarpiece," also of the 1430s, is, so far as is known, the first *sacra conversazione* (*i.e.*, "sacred conversation," a representation of the Holy Family) of the Renaissance.

Composition of "The Annunciation"

Years at the monastery of S. Marco. Angelico remained in the Fiesole monastery until 1439, when he entered the monastery of S. Marco in Florence. There he did most of his work as a mural painter. S. Marco had been transferred from the Sylvestrine monks to the Dominicans in 1436, and the rebuilding of the church and its spacious monastery began around 1438, from designs by the Florentine architect and sculptor Michelozzo. The construction was generously subsidized by the Medici family. Angelico was commissioned around 1438 by Cosimo de' Medici the Elder to execute the altarpiece, for which he again painted a *sacra conversazione*. When the church was consecrated at Epiphany in 1443, the altarpiece must have dominated the place of worship. Angelico portrayed the Virgin and child raised high on a throne, with saints on either side receding into space; among them are the two patron saints of the Medici, Cosmas and Damian. This work, one of the most compelling Fra Angelico ever created, ends in a dense grove of cypresses, palms, and pines against a deep but toneless sky. His figures seem cleansed of any human passion and to have supreme serenity of spirit. A predella, showing eight little legends of the two Medicean saints separated by a Pietà (Virgin Mary holding the body of Christ) completed the work. Unfortunately, these paintings are now scattered among various museums. The narrative in all the scenes is more organized and simplified than in his previous work, with creative touches that he was later to carry forward in his mural painting.

On the walls of the monastery of S. Marco in Florence are the paintings that mark the high point of Angelico's career. In the chapter hall, he executed a large "Crucifixion" that seems akin to the "Moralities" of the 14th century, which urged detachment from worldly vanities and salvation through Christ alone. In addition to the three crucified figures against the sky, Angelico painted groups of ritual figures, rhythmically arranged, with a chorus of martyrs, founders of religious orders, hermits, and defenders of the Dominican order (whose genealogical tree is depicted beneath this striking scene), as well as the two Medicean saints. Thus, in the comprehensiveness of this work, Fra Angelico developed a concept that was barely suggested in his earlier altarpieces. He portrayed the exaltation of the Redeemer in many other paintings in the monastery's first cloister and in its cells. In one corridor he executed an Annunciation that broadened the pattern of his earlier one in Cortona and, beyond it, a *sacra conversazione*, bathed in lucid light. In the cells, he proclaimed devotion to Christ crucified in at least 20 examples, all related to monastic life. The pictorial work in these narrow spaces is intricate, probably the work of numerous hands directed by the master, including Benozzo Gozzoli, the greatest of Fra Angelico's disciples, and Zanobi Strozzi, another pupil better known as a miniaturist, as well as his earliest collaborator, Battista Sanguigni. The hand of Fra Angelico himself is identifiable in the first ten cells on the eastern side. Three subjects merit particular attention: a Resurrection, a coronation of the Virgin, and, especially, a gentle Annunciation, presented on a bare white gallery, with St. Peter Martyr in prayer, timidly facing the group, his coloured habit contrasting with the delicate two tones of pink in the garments of the Virgin and the Angel. The cells, originally hidden from public view because of monastic vows of seclusion, reveal the secret joy of the painter-monk in creating figures of purity to move his fellow monks to meditation and prayer. The images in these paintings are the lyrical expressions of a painter who was also their prior.

The high point of Angelico's career

Stay in Rome. At the end of 1445 Fra Angelico was called to Rome by Pope Eugene IV, and he remained

there until about 1450. In the summer of 1447, however, he had undertaken to decorate the chapel of S. Brizio in the cathedral of Orvieto. Angelico's assistants, above all, Gozzoli, worked closely with him on two canvases, crowded with figures, in this chapel. These canvases of Christ the Judge, amid the hierarchy of angels, and the chorus of the prophets, respectively, were only partially executed by Angelico; they were continued more than 50 years later by Luca Signorelli.

Frescoes executed in Rome

In Rome, the frescoes that Angelico executed in a chapel of St. Peter's (*c.* 1446–47), in the chapel of the Sacrament in the Vatican (not before 1447), and in the studio of Pope Nicholas V (1449) have all been destroyed. But the Vatican still possesses his decorative painting for the Cappella di Niccolò V. There, he painted scenes from the lives of SS. Stephen and Lawrence, along with figures of the Evangelists and saints, repeating some of the patterns of the predella on his altarpiece of S. Marco. The consecration scenes of St. Stephen and that of St. Lawrence are both set in solemn cathedral interiors, and the almsgiving of St. Lawrence is set against the background of a temple. In this scene particularly, Angelico imbued the poor and afflicted who surround the deacon-saint with a serenity that purifies them and illuminates them with an inner light, rendering them equals of the blessed figures on the altarpieces. At the same time, the organization of these works and the rendering of architecture in them mark the culmination of his development as a Renaissance artist.

Around 1450 Fra Angelico returned to Florence, where, still a monk, he became prior of the monastery of S. Domenico in Fiesole (1450–*c.* June 1452). His most notable work of this time was the cycle of 35 paintings of scenes from the life of Christ and other subjects, for the doors of a silver chest in the sanctuary of the church of SS. Annunziata in Florence. These works, which have been extensively repainted, are probably distant echoes of the destroyed paintings in the Cappella di Niccolò V. Although the authenticity of these works is disputed, the "Massacre of the Innocents," "Flight into Egypt," and "Presentation in the Temple" seem to be Angelico's because of the bright spontaneity of the slender figures, as well as the spatiality of the surroundings and the landscape. Such traits derived from the artist's vast experience in mural painting. In most of these little pictures, however, there is a kind of disconnectedness and weariness, indicating the hand of pupils, whose art was a far cry from Fra Angelico's ineffable poetry. There is still a certain monumental tone in the late altarpiece he executed in the monastery of Bosco ai Frati in the Mugello (now in the Museo di San Marco, Florence). With the completion of this altarpiece and several other minor works, Fra Angelico's fertile artistic labours drew to a close.

In 1453 or the following year, Fra Angelico again went to Rome, where he died on February 18, 1455, in the Dominican monastery in which he had stayed during his first visit to Rome. It was close by the church of Sta. Maria della Minerva, where his tomb remains an object of veneration.

Influence

In addition to the influence he had on his followers, Fra Angelico exerted a significant influence in Florence, especially between 1440 and 1450, even on such an accomplished master as Fra Filippo Lippi. As a monk, Fra Angelico was lauded in writings of the 15th century and later, some of which bestowed a legendary halo on him. As a painter, he was acclaimed as early as 1438 by the contemporary painter Domenico Veneziano. Vasari, in his section on Angelico in *Lives of the Most Eminent Italian Painters, Sculptors and Architects*, was inaccurate in his biographical data but correctly situated Fra Angelico in the framework of the Renaissance. Vasari characterized him in terms that remained standard until the end of the 18th century, when writers of the Neoclassical period, using judgments of a philosophical and didactic nature, placed him out of his time and even in the 14th century, thus making him an artist of transition. Almost all modern art critics, however, place him again within the framework of the Renaissance. With classical measure, Fra Angelico embodied a deeply religious attitude.

MAJOR WORKS

"The Dead Christ with Saints" (1425–28; Courtauld Institute Galleries, London); "Madonna of the Star" ("Madonna della Stella," 1425–30; Museo di San Marco, Florence); "Madonna with Child and Two Angels" (1425–30; van Beuningen Collection, Vierhouten, The Netherlands); "The Coronation of the Virgin" (1428; Uffizi, Florence); "Madonna and Child" (1428–30; S. Domenico, Fiesole); "The Martyrdom of St. Peter" (1429; Museo di San Marco); "The Madonna of Humility" (1430–33; National Gallery of Art, Washington, D.C.); "The Last Judgment" (*c.* 1430–33; Museo di San Marco); "The Coronation of the Virgin" (1430–35; Louvre, Paris); "Linaiuoli Altarpiece" (1433; Museo di San Marco); "Sta. Trinita Altarpiece" (*c.* 1433; Museo di San Marco); "The Death and Assumption of the Virgin" (*c.* 1434; Isabella Stewart Gardner Museum, Boston); "Virgin and Child with Angels, Saints and Donor" (*c.* 1435; Museum of Fine Arts, Boston); "The Annunciation" (*c.* 1436; Prado, Madrid); "Lamentation" (1436; Museo di San Marco); "The Annunciation" (*c.* 1436; Museo Diocesano, Cortona); triptych (*c.* 1436; Museo Diocesano, Cortona); "Triptych: Madonna and Child with Saints and Angels" (*c.* 1437; Galleria Nazionale dell'Umbria, Perugia); "Madonna and Child Enthroned Between SS. Cosmas and Damian, Peter Martyr, John the Evangelist, Lawrence, and Francis" ("Annalena Altarpiece," begun 1437; Museo di San Marco); "Virgin and Child Enthroned with Angels Between SS. Lawrence, John the Evangelist, Mark, Cosmas, Damian, Dominic, Francis and Peter Martyr" (1438–40; Museo di San Marco); frescoes at San Marco, Florence (1438–50); frescoes, Cappella di Niccolò V (1448–50; Vatican, Rome); 35 paintings for the doors of a silver chest in SS. Annunziata, Florence, Museo di San Marco.

BIBLIOGRAPHY. JOHN POPE-HENNESSY, *Fra Angelico* (1952); *Mostra delle opere del Beato Angelico nel V Centenario della morte* (1955), exhibition of Fra Angelico's works on the 500th anniversary of his death; MARIO SALMI, *Il Beato Angelico* (1958), with a critical bibliography; STEFANO ORLANDI, *Beato Angelico* (1964), in Italian; ELSA MORANTE and UMBERTO BALDINI, *L'opera completa dell'Angelico* (1970), an important catalog.

(M.Sal.)

Angels and Demons

Throughout the history of religions, both ancient and modern, primitive and sophisticated, varying kinds and degrees of beliefs have existed in various spiritual beings, powers, and principles that mediate between the realm of the sacred or holy—*i.e.*, the transcendent realm—and the profane realm of time, space, and cause and effect. Such spiritual beings when regarded as benevolent are usually called angels in Western religions; those viewed as malevolent are termed demons. In other religions—Eastern, ancient, and primitive—such intermediate beings are less categorical, for they may be benevolent in some circumstances and malevolent in others.

NATURE AND SIGNIFICANCE

Angels. The term angel, which is derived from the Greek word *angelos*, is the equivalent of the Hebrew word *mal'akh*, meaning "messenger." The literal meaning of the word angel thus points more toward the function or status of such beings in a cosmic hierarchy rather than toward connotations of essence or nature, which have been prominent in popular piety, especially in Western religions. Thus, angels have their significance primarily in what they do rather than in what they are. Whatever essence or inherent nature they possess is in terms of their relationship to their source (God, or the ultimate being). Because of the Western iconography (the system of image symbols) of angels, however, they have been granted essential identities that often surpass their functional relationships to the sacred or holy and their performative relationships to the profane world. In other words, popular piety, feeding on graphic and symbolic representations of angels, has to some extent posited semidivine or even divine status to angelic figures. Though such occurrences are not usually sanctioned doctrinally or theologically, some angelic figures, such as Mithra (a Persian god who in Zoroastrianism became an angelic mediator between heaven and earth and judge

The meaning of angels in functional terms

and preserver of the created world), have achieved semidivine or divine status with their own cults.

In Zoroastrianism there was a belief in the *amesha spentas*, or the holy or bounteous immortals, who were functional aspects or entities of Ahura Mazdā, the Wise Lord. One of the *amesha spentas*, Vohu Manah (Good Mind), revealed to the Iranian prophet Zoroaster (6th century BC) the true God, his nature, and a kind of ethical covenant, which man may accept and obey or reject and disobey. In a similar manner, about 1,200 years later, the angel Gabriel (Man of God) revealed to the Arabian prophet Muḥammad (5th–6th century AD) the Qur'ān (the Islāmic scriptures) and the true God (Allāh), his oneness, and the ethical and cultic requirements of Islām. The epithets used to describe Gabriel, the messenger of God—"the spirit of holiness" and "the faithful spirit"—are similar to those applied to the *amesha spentas* of Zoroastrianism and the third Person of the Trinity (Father, Son, and Holy Spirit) in Christianity. In these monotheistic religions (though Zoroastrianism later became dualistic) as also in Judaism, the functional characteristics of angels are more clearly enunciated than their ontological (or nature of Being) characteristics—except in the many instances in which popular piety and legend have glossed over the functional aspects.

Various religions, including primitive religions, have beliefs in intermediary beings between the sacred and profane realms, but the belief is most fully elaborated in religions of the West.

Changes and variations in the meaning of the term demon

Demons. The term demon is derived from the Greek word *daimōn*, which means a "supernatural being" or "spirit." Though it has commonly been associated with an evil or malevolent spirit, the term originally meant a spiritual being that influenced a person's character. An *agathos daimōn* ("good spirit"), for example, was benevolent in its relationship to men. The Greek philosopher Socrates, for example, spoke of his *daimōn* as a spirit that inspired him to seek and speak the truth. The term gradually was applied to the lesser spirits of the supernatural realm who exerted pressures on men to perform actions that were not conducive to their well-being. The dominant interpretation has been weighted in favour of malevolence and that which forbodes evil, misfortune, and mischief.

In primitive religions, spiritual beings may be viewed as either malevolent or benevolent according to the circumstances facing the individual or community. Thus, the usual classification that places demons among malevolent beings is not totally applicable in reference to primitive religions.

The positions of spiritual beings or entities viewed as benevolent or malevolent may, in the course of time, be reversed. Such has been the case in the ancient Indo-Iranian religion, from which evolved early Zoroastrianism and the early Hinduism reflected in the Vedas (ancient Aryan hymns). In Zoroastrianism the *daevas* were viewed as malevolent beings, but their counterparts, the *devas* in ancient Hinduism, were viewed as gods. The *ahuras* of Zoroastrianism were good "lords," but in Hinduism their counterparts, the *asuras*, were transformed into evil lords. In a similar manner, Satan, the prosecutor of men in the court of God's justice in the Old Testament book of Job, became the chief antagonist of Christ in Christianity and of man in Islām. Many similar transformations indicate that the sharp distinctions made between angels as benevolent and demons as malevolent may be too simplistic, however helpful such designations may be as indicators of the general functions of such spiritual beings.

CELESTIAL AND NONCELESTIAL FORMS: RELATIONSHIPS OF BELIEFS IN ANGELS AND DEMONS TO VIEWS OF THE COSMOS

Because man is a being much concerned with boundaries —*i.e.*, what makes him different from other animate beings, what makes his community (and thus his world) different from other communities (and other worlds)— his view of the cosmos has influenced his understanding of what are called angels and demons. The cosmos may be viewed as monistic, as in Hinduism, in which the cosmos is regarded as wholly sacred or as participating in a single divine principle (Brahman, or Being itself). The cosmos may also be viewed as dualistic, as in Gnosticism (an esoteric religious dualistic belief system, often regarded as a Christian heretical movement, that flourished in the Greco-Roman world in the 1st and 2nd centuries AD), in which the world of matter was generally regarded as evil and the realm of the spirit as good. A third view of the cosmos, generally found in the monotheistic religions of Judaism, Zoroastrianism, Christianity, and Islām, centred on a tripartite universe: celestial, terrestrial, and subterrestrial. This third view has influenced Western man's concepts of angels and demons as well as his scientific and metaphysical concepts.

Man's concern with ways of relating to the cosmos

Relationship to views of a tripartite cosmos. In the biblical, Hellenistic (Greco-Roman cultural), and Islāmic worlds of thought, the terrestrial realm was a world in which man was limited by the factors of time, space, and cause and effect. The celestial realm, generally composed of seven heavens or spheres dominated by the seven then-known planets, was the realm of the divine and the spiritual. The subterrestrial realm was the area of chaos and the spiritual powers of darkness. At the highest level of the celestial sphere was the ultimate of the sacred or holy: *e.g.*, Yahweh, the God of Judaism, whose name was so holy it should not even be spoken; Bythos, the unknowable beginning beyond beginnings of Gnosticism; the heavenly Father of Christianity, known through his Logos (the divine Word, or Reason, Jesus Christ); and Allāh, the powerful, the almighty, and the sublime God of Islām.

The role of angels as celestial beings

In order to reveal the purpose and destiny of man—the highest being of the terrestrial realm—the ultimate of the celestial sphere enabled man, according to such views, to come to a knowledge of who he is, what is his origin, and what is his destiny through celestial messengers—angels. The message, or revelation, was usually focussed on the identity of the source of the revelation— *i.e.*, the ultimate being—and on the destiny of man according to his response. Because of a cosmic rift in the heavenly sphere prior to the creation of the world or the announcement of the revelation, angels, depending on their relationship to the Creator, might attempt to deceive man with a false revelation or to reveal the truth about man's true nature (or identity), origin, and destiny. Angels who attempted to pervert the message of the ultimate celestial being in order to confuse man's understanding of his present boundary situation as a terrestrial being or his destiny as a supraterrestrial being—though not always termed demons—are malevolent in function. Included among such malevolent angels are the devil of Christianity and Judaism or Iblīs (the Devil) of Islām, who, in the form of a serpent in the biblical story of the Garden of Eden—according to later interpretations of the story—attempted to disrupt man's understanding of his creaturely boundaries, or limitations. He did this by tempting man to eat the fruit of the tree of knowledge of good and evil so that he might become like God (or the divine beings of the heavenly court). In Zoroastrianism, the Evil Spirit (Angra Mainyu, later Ahriman) attempted—through subservient spirits such as Evil Mind, the Lie, and Pride—to deceive terrestrial man so that he would choose a destiny that was subterrestrial—punishment in a chasm of fire.

In the aftermath of the 16th-century Copernican revolution (based on the theories of the Polish astronomer Copernicus), in which man's view of the cosmos was radically altered—*i.e.*, the Earth was no longer seen as the centre of the cosmos but, instead, merely as a planet of a solar system that is a very small part of a galaxy in an apparently infinite universe—the concepts of angels and demons no longer seemed appropriate. The tripartite cosmos—heaven above, Earth in the middle, and hell below—appeared to be an anachronism.

With the emergence of modern Western psychology and psychoanalytical studies in the 19th and 20th centuries, however, the underlying principles of beliefs in angels

Influence of psychology on angelology and demonology

and demons have taken on new meanings. Many Christian theologians have found some of the concepts of psychoanalysis helpful in reinterpreting the meanings of underlying primitive and traditional beliefs in angels and demons. The tripartite cosmos was re-mythologized into a tripartite structure of the personality—the superego (the restrictive social regulations that enable man to live as a social being), the ego (the conscious aspects of man), and the id, or libido (a "seething, boiling cauldron of desire that seeks to erupt from beneath the threshold of consciousness"). Thus, demons—according to this reinterpretation—might well be redefined as projections of the unregulated drives of man that force him to act only according to his own selfish desires, taking no account of their effects on other persons. From a social point of view, demons might also be defined as the environmental and hereditary forces that cause man to act, think, and speak in ways that are contrary to the well-being of himself and his community. A modern French writer, Denis de Rougemont, has maintained in his book *The Devil's Share* that the devil and the demonic forces that plague the modern world can be well documented in modern man's return to barbarism and man's inhumanity to man. In the 2nd century AD, Clement of Alexandria, a Christian philosophical theologian, pointed toward a psychological interpretation of demonic forces by stating that man was often captivated by the inner appetitive drives of his passions and bodily desires. The Freudian "myth" of the human personality and other psychological studies have thus initiated a new dimension in the study of angels and demons. Medieval iconography, which graphically depicted angels and demons as hybrid creatures that often defied even the most vivid imaginations of the persons who viewed them, has been supplanted by psychological, psychoanalytical, and modern mythological symbolism coupled with theological reflection.

Angels as aids or hindrances to man's goal of salvation

Relationship to views of a dualistic cosmos. In religious traditions that have viewed the cosmos in a dualistic fashion, such as Gnosticism, angels were believed to be celestial beings who controlled certain spheres through which a soul was to pass as it freed itself from the shackles of its material existence. Knowledge of these angels and their names was a necessary prerequisite for achieving eventual union with the ultimate spiritual reality. Included among various lists of the seven angels ruling the seven planetary spheres are Gabriel, Adonai (Lord), Aariel (lion of God), and others. The angel of the creation of the world of matter, Yahweh (sometimes called the Demiurge, the Creator), was evil, in the Gnostic view, not only because he was the Creator but also because he tried to keep spiritual men from knowing their true origin, nature, and destiny.

Manichaeism, a dualistic religion founded in the 3rd century AD by Mani, an Iranian prophet, like Gnosticism divided the world into two spheres—Goodness (Light) and Evil (Darkness). These two principles are mixed in the world of matter, and the object of salvation is to unmix the material and the spiritual so that one may achieve a state of absolute goodness. Highest in the celestial hierarchy are the 12 light diadems of the Father of Greatness and the Twelve Aeons, the "firstborn"—angelic figures that are divided into groups of threes, surrounding the Supreme Being in the four quarters of the heavens. Because the Devil, the Prince of Darkness, desires the advantages of the Kingdom of Light, in an ensuing battle between the celestial forces Light and Darkness are mixed, and the world of matter and spirit is created. Unaware of his spiritual nature and constantly tempted by the demons of the Prince of Darkness, man is eventually led to understand his true nature through the activity of angelic beings called the Friends of the Lights and the Living Spirit and his five helpers: Holder of Splendour, King of Honour, Light of Man, King of Glory, and Supporter.

Relationship to views of a monistic cosmos. Those who view the cosmos as basically monistic—such as Hinduism, Jainism, and Buddhism—generally have no belief in angels, who function mainly as revealers of the truth. This function is performed by other beings, such as *avatāra*s (incarnations of the gods) in Hinduism, *tīrthaṅkara*s (saints or prophets) in Jainism, or *bodhisattva*s (Buddhas-to-be) in Buddhism. Because such personages generally are viewed more in terms of exemplifiers of the holy life than as conduits of a revelation (except in the case of several *avatāra*s and *bodhisattva*s), they are not to be regarded in terms of the typical Western conceptions of angelic beings. These religions do, however, have widespread beliefs in demons.

Belief in demons as common to all religious or mythological views about the cosmos. Belief in demons is not connected with any particular view of the cosmos. Demons have a very wide geographical and lengthy historical role as spiritual beings influencing man in his relationship to the sacred or holy. They may be semihuman, nonhuman, or ghostly human beings who, for various reasons, generally attempt to coerce man into not attaining his higher spiritual aspirations or not performing activities necessary for his well-being in the normal course of living. The ancient Assyrian demon *rabiṣu* apparently is a classic prototype of a supernatural being that instilled such a fear in men that their hair literally raised from their bodies when confronted with knowledge of the *rabiṣu*'s presence.

Powers of demons in the West

In 17th-century Europe, various demons were cataloged according to their powers to entice men to indulge in what were called their basic instincts or desires. Included in such lists were nightmare demons, demons formed from the semen of copulation, and demons who deceived persons into believing that they could perform transvections (nocturnal flights to sites of sabbats, alleged rites of witchcraft). According to some authorities in the 20th century (as well as early Christian polemicists), the alleged demons noted by the prevailing religions of the world are the former gods or spiritual beings that succumbed to or were overpowered by the dominant doctrinal views of a conquering people. Thus, the Teutonic, Slavic, Celtic, or Roman gods either were reduced to demonic antagonists of Christ, his saints, or his angels or were absorbed by the cults of Christian saint figures. Followers of the ancient but no longer influential deities were often subjected to persecution as advocates of witchcraft, especially in Christian Europe (see also WITCHCRAFT).

TYPES OF ANGELS AND DEMONS

Angels and demons, as noted earlier, have been categorized as benevolent, malevolent, or ambivalent or neutral beings that mediate between the sacred and profane realms.

Functions of angels

Benevolent beings. Benevolent beings, usually angels but sometimes ghosts of ancestors or other spiritual beings that have been placated by sacrifices or other rituals, assist man in achieving a proper rapport with God, other spiritual beings, or man's life situations. Angels, for example, not only act as revealers of divine truths, but they also are believed to be efficacious in helping man to attain salvation or special graces or favours. Their primary function is to praise and serve God and do his will. This is true of angels in both Christianity and Zoroastrianism, as well as in Judaism and Islām. As functional extensions of the divine will, they sometimes intervene in human affairs by rewarding the faithful and punishing the unjust or by saving the weak who are in need of help and destroying the wicked, who unjustly persecute their fellow creatures. In the intertestamental book of Tobit (an apocryphal, or "hidden," book that is not accepted as canonical by Jews and Protestants), the archangel Raphael (God Heals), for example, helps the hero Tobias, the son of Tobit, on a journey and also reveals to him magic formulas to cure his father's blindness and to counteract the power of the demon Asmodeus.

Angels also have been described as participants in the creation and the providential continuance of the cosmos. Clement of Alexandria, influenced by Hellenistic cosmology, stated that they functioned as the movers of the stars and controlled the four elements—earth, air, fire,

and water. Many angels are believed to be guardians over individuals and nations. The view that there are guardian angels watching over children has been a significant belief in the popular piety of Roman Catholicism. Angels are also regarded as the conductors of the souls of the dead to the supraterrestrial world. In the procreation of men, angels are believed to perform various services. This is especially noticeable in the instances of angels announcing the births of divine figures or special religious personages, such as Jesus and John the Baptist in the New Testament.

Though the function of angels is of primary significance, theological reflection and popular piety have placed much emphasis on the nature of angels. In early Judaism angels were conceived as beings in human form: the angel who wrestled with the patriarch Jacob, as recorded in the book of Genesis, was in the form of a man. In Judaism of the Hellenistic period (3rd century BC to 3rd century AD), however, angels were viewed as noncorporeal spiritual beings who appeared to man in an apparitional fashion. Their spiritual nature had been emphasized earlier by Old Testament prophets, such as Ezekiel and Isaiah, in their visionary descriptions. The cherubim and seraphim, two superior orders of angels, are described as winged creatures that guard the throne of God. The use of wings attached to various beings symbolizes their invisible and spiritual nature, a practice that can be traced back to the ancient Egyptians, who represented the battling sun-god Horus of Edfu as a winged disk. In Christian iconography the spiritual nature of angels has been almost universally represented—until the 20th century—by winged human figures. Their spirituality and, therefore, their noncorporeality led to various kinds of speculation among theologians and common people about the nature of the appearances of angels, which has been recorded in both Scripture and legends based on popular piety. Some theologians, such as Augustine in the 4th and 5th centuries, stated that angels, who have ethereal bodies, may be able to assume material bodies. This problem, however, has not been solved to the satisfaction of later theologians.

Malevolent beings. Malevolent beings—demons, fallen angels, ghosts, goblins, evil spirits in nature, hybrid creatures, the *daevas* of Zoroastrianism, the *narakas* (creatures of hell) of Jainism, the *oni* (attendants of the gods of the underworld) in Japanese religions, and other such beings—hinder man in achieving a proper relation with God, the spiritual realm, or man's life situations. Some angels are believed to have fallen from a position of proximity to God—such as Lucifer (after his fall called Satan by early Church Fathers) in Judaism, Christianity, and Islām—because of pride or for attempts to usurp the position of the Supreme Being. In their fallen condition they then attempt to keep man from gaining a right relationship with God by provoking men to sin. Some medieval scholars of demonology ascribed to a hierarchy of seven archdemons the seven deadly sins: Lucifer (Pride); Mammon (Avarice); Asmodeus (Lechery); Satan (Anger); Beelzebub (Gluttony); Leviathan (Envy); and Belphegor (Sloth). Besides tempting men to sin, the fallen angels, or devils, were believed to cause various types of calamities, both natural and accidental. Like the demons and evil spirits of nature in primitive religions, the fallen angels were viewed as the agents of famine, disease, war, earthquakes, accidental deaths, and various mental or emotional disorders. Persons afflicted with mental diseases were considered to be "demon possessed."

The function of demons

Though the functions of demonic figures, like those of fallen angels, is of major significance, the nature of demons has been of concern to theologians and persons infused with popular piety. Like angels, demons are regarded as spiritual, noncorporeal beings, but they have been depicted in religious iconography as hybrid creatures with horrifying characteristics or as caricatures of idols of an opposing religion. In the early church, for example, there was a belief that pagan idols were inhabited by demons. The horrifying aspects of demons have been represented in the woodcuts of medieval and Reformation artists and in the masks of shamans, medicine men, and priests of primitive religions—either to frighten the believer into behaving according to accepted norms or to ward off ritualistically the power of the demonic forces loose in the terrestrial or profane realm.

Ambivalent or neutral beings. Ambivalent or neutral spiritual beings are usually not found in Western religions, which usually divide the inhabitants of the cosmos into those who are either allied with or in opposition to the Supreme Being. Islām, however, classifies spiritual beings into angels (*malā'ikah*), demons (*shāyaṭīn*), and *jinn*, or genies. This last category includes spiritual beings that might be either benevolent or malevolent. According to legend, the *jinn* were created out of fire 2,000 years before the creation of Adam, the first man. Capable of both visibility and invisibility, a *jinn* could assume various forms—either animal or human—and could be either a help or a hindrance to man. By cunning, a superior use of intellect, or magic, a man might be able to manipulate a *jinn* for his own benefit.

The role of the *jinn*

Various minor nature spirits—such as the spirits of water, fire, mountains, winds, and other spirits recognized in primitive religions—are generally neutral, but, in order to keep them that way or to make them beneficial to man, proper sacrifices and rituals must be performed.

VARIETIES OF ANGELS AND DEMONS IN THE RELIGIONS OF THE WORLD

Intermediate beings between the sacred and profane realms assume various forms in the religions of the world: celestial and atmospheric beings; devils, demons, and evil spirits; ghosts, ghouls, and goblins; and nature spirits and fairies.

In Zoroastrianism, Judaism, Christianity, and Islām. In the Western religions, which are monotheistic and view the cosmos as a tripartite universe, angels and demons are generally conceived as celestial or atmospheric spirits. In the popular piety of these religions, however, there is a widespread belief in ghosts, ghouls, goblins, demons, and evil spirits that influence man in his terrestrial condition and activities. The celestial beings may be either benevolent or malevolent, depending on their own relationship to the Supreme Being. On the other hand, the demons and evil spirits that generally influence man in his role as a terrestrial being (rather than in his destiny as a supraterrestrial being) are viewed in popular piety—and somewhat in theological reflection—as malevolent in intent.

Angels are generally grouped in orders of four, six, or seven in the first ranks, of which there may be several. The use of four, which symbolically implies perfection and is related to the four cardinal points, is found in Judaism, Christianity, and Islām. Early Zoroastrianism, much influenced by the astronomical and astrological sciences of ancient Iran, coordinated the concept of the seven known planetary spheres with its belief in the heptad (grouping of seven) of celestial beings—*i.e.*, the *amesha spentas* of Ahura Mazdā: Spenta Mainyu (the Holy Spirit), Vohu Mana (Good Mind), Asha (Truth), Ārmaiti (Right Mindedness), Khshathra (Kingdom), Haurvatāt (Wholeness), and Ameretāt (Immortality). In later Zoroastrianism, though not in the *Gāthās* (the early hymns, believed to have been written by Zoroaster, in the Avesta, the sacred scriptures), Ahura Mazdā and Spenta Mainyu were identified with each other, and the remaining bounteous immortals were grouped in an order of six. Over against the bounteous immortals, who helped to link the spiritual and the material worlds together, was the counterpart of the Holy Spirit, namely Angra Mainyu, the Evil Spirit, who later became the great adversary Ahriman (the prototype of the Jewish, Christian, and Islāmic Satan), and the *daevas*, who were most likely gods of early Indo-Iranian religion. Allied with Angra Mainyu against Ahura Mazdā were Akōman (Evil Mind), Indrā-vāyū (Death), Saurva (a *daeva* of death and disease), Nāñhaithya (a *daeva* related to the Vedic god Nāsatya), Tauru (difficult to identify), and Zairi (the personification of Haoma, the sacred drink

The orders of ranks of angels and demons in Zoroastrianism

related to the sacrifices of both *ahuras* and *daevas*). Among other demonic figures is Aēshma (violence, fury, or the aggressive impulse that consumes man)—who may well be the demon Asmodeus of the book of Tobit, Āz (Concupiscence or Lust), Mithrāndruj (He Who Lies to Mithra or False Speech), Jēh (the demon Whore, created later by Ahriman to defile the human race), and many others (see also ZOROASTRIANISM AND PARSIISM).

The angels and demons of Judaism and Christianity

Angelology and demonology in Judaism became more highly developed during and after the period of the Babylonian Exile (6th–5th centuries BC), when contacts were made with Zoroastrianism. In the Old Testament, Yahweh is called the Lord of hosts. These hosts (Sabaoth) are the heavenly army that fights against the forces of evil and performs various missions, such as guarding the entrance to Paradise, punishing evildoers, protecting the faithful, and revealing God's Word to man. Two archangels are mentioned in the canonical Old Testament: Michael, the warrior leader of the heavenly hosts, and Gabriel, the heavenly messenger. Two are mentioned in the apocryphal Old Testament: Raphael, God's healer or helper (in the book of Tobit), and Uriel (Fire of God), the watcher over the world and the lowest part of hell (in II Esdras). Though these are the only four named, seven archangels are noted in Tob. 12:15. Besides the archangels, there were also other orders of angels, the cherubim and seraphim, which have been noted earlier.

Under the influence of Zoroastrianism, Satan, the adversary, probably evolved into the archdemon. Other demons included Azazel (the demon of the wilderness, incarnated in the scapegoat), Leviathan and Rahab (demons of chaos), Lilith (a female night demon), and others. To protect themselves from the powers of the demons and unclean spirits, Jews influenced by folk beliefs and customs (as with Christians later) often carried charms, amulets, and talismans inscribed with efficacious formulas (see also JUDAISM; JEWISH MYTH AND LEGEND).

Christianity, probably influenced by the angelology of Jewish sects such as the Pharisees and the Essenes as well as of the Hellenistic world, further enhanced and developed theories and beliefs in angels and demons. In the New Testament, celestial beings were grouped into seven ranks: angels, archangels, principalities, powers, virtues, dominions, and thrones. In addition to these were added the Old Testament cherubim and seraphim, which, with the seven other ranks, comprised the nine choirs of angels in later Christian mystical theology. Various other numbers of the orders of angels have been given by early Christian writers: four, in *The Sibylline Oracles* (a supposedly Jewish work that shows much Christian influence); six, in the *Shepherd of Hermas*, a book accepted as canonical in some local early Christian churches; and seven, in the works of Clement of Alexandria and other major theologians. In both folk piety and theology the number has generally been fixed at seven. The angels receiving most attention and veneration in Christianity were the four angels mentioned in the Old Testament and the Apocrypha. Michael became the favourite of many, and in the practice of his cult there was often some confusion with St. George, who was also a warrior figure.

Demonology experienced a renewal in Christianity that probably would have been acceptable in Zoroastrianism. Satan, the archenemy of the Christ; Lucifer, the fallen Light Bearer; and the originally Canaanite Beelzebub, the Lord of Flies (or, perhaps, Beelzebul, the Lord of Dung), mentioned by Jesus, are all devils. The concept and term devil are derived from the Zoroastrian concept of *daevas* and the Greek word *daibolos* ("slanderer" or "accuser"), which is a translation of the Jewish concept of Satan. As a singular demonic force or personification of evil, the devil's chief activity was to tempt man to act in such a way that he would not achieve his supraterrestrial destiny. Because demons were believed to inhabit waterless wastelands, where hungry and tired persons often had visual and auditory hallucinations, early Christian monks went into the deserts to be the vanguard of God's army in joining battle with the tempting devils. They often recorded that the devil came to them in visions as a seductive woman, tempting them to violate their vows to keep themselves sexually pure, both physically and mentally.

Worship of demons in the West

During certain periods in Christian Europe, especially the Middle Ages, worship of demons and the practice of witchcraft brought about the wrath of both church and people on those suspected of practicing diabolical rites, such as the Black Mass. One formula from the Black Mass (the mass said in reverse and with an inverted crucifix on the altar) has survived in popular magic: "hocus-pocus," an abbreviated form of "Hoc est corpus meum" ("This is my body"), the words of institution in the Eucharist, or Holy Communion. Witchcraft and sorcery have been closely associated with demonology in the thought of Christianity, especially in the West.

In the second half of the 20th century, in connection with a renewed interest in the supernatural, there has been evidence of a revival of demon worship and black magic, although this has generally been restricted to small cults that have proved to be quite ephemeral (see also CHRISTIANITY; CHRISTIAN MYTH AND LEGEND).

Angelology and demonology in Islām are closely related to similar doctrines in Judaism and Christianity. Besides the four throne bearers of Allāh, four other angels are well-known: Jibrīl (Gabriel), the angel of revelation; Mīkāl (Michael), the angel of nature, providing man with food and knowledge; ʿIzrāʾīl, the angel of death; and Isrāfīl, the angel who places the soul in the body and sounds the trumpet for the Last Judgment. Demons also contend for control of men's lives, the most prominent being Iblīs (the Devil), who tempts mortal man, or Shayṭan, or Satan (see also ISLAM; ISLAMIC MYTH AND LEGEND).

In the religions of the East. As noted earlier, the function of angels in Eastern religions was carried by *avatāras*, *bodhisattvas*, and other such spiritual beings who were extensions of God or the sacred. Belief in demons was and is very widespread, influencing various rituals and practices to counteract the forces that are hostile to man and nature. In Hinduism, the *asuras* (the Zoroastrian *ahuras*) are the demons who oppose the *devas* (the gods). Both vied for the *homa*, or the *amṛta* (the sacred drink that gives power), but the god Viṣṇu (the preserver), incarnated as a beautiful woman (Mohinī), aided the gods so that they alone would drink the *amṛta*, thus giving them power over the demons. Among the various classes of Hindu *asuras* (demons) are *nāgas* (serpent demons); Ahi (the demon of drought); and Kaṃsa (an archdemon). Demons that afflict men include the *rākṣasas*, grotesque and hideous beings of various shapes who haunt cemeteries, impel men to perform foolish acts, and attack *sadhus* (saintly men), and *piśacas*, beings who haunt places where violent deaths have occurred. Buddhists often view their demons as forces that inhibit man from achieving Nirvāṇa (bliss or the extinction of desire). Included among such beings are Māra, an arch tempter, who, with his daughters, Rati (Desire), Rāga (Pleasure), and Tanhā (Restlessness), attempted to dissuade Siddhārtha Gautama, the Buddha, from achieving his Enlightenment. As Mahāyāna (Greater Vehicle) Buddhism spread to Tibet, China, and Japan, many of the demons of the folk religions of these areas were incorporated into Buddhist beliefs. The demons of Chinese religions, the *kuei-shen*, are manifested in all aspects of nature. Besides these nature demons there are goblins, fairies, and ghosts. Because the demons were believed to avoid light, the Chinese who were influenced by Taoism and folk religions used bonfires, firecrackers, and torches to ward off the *kuei*. Japanese religions are similar to Chinese religions in the multiplicity of demons with which men must contend. Among the most fearsome of the Japanese demons are the *oni*, evil spirits with much power, and the *tengu*, spirits that possess man and that generally must be exorcized by priests (see also HINDUISM; HINDU MYTHOLOGY; BUDDHISM; BUDDHIST MYTHOLOGY; JAINISM; CHINESE RELIGION; CHINESE MYTHOLOGY; JAPANESE RELIGION; JAPANESE MYTHOLOGY).

Types of demons in Eastern religions

In primitive religions. The spiritual beings of primitive or preliterate religions of Asia, Africa, Oceania, and the Americas are generally viewed as malevolent or benevolent according to circumstances rather than because of their inherent nature. Eshu, a god of the Yoruba of Nigeria, for example, is looked upon as a protective, benevolent spirit as well as a spirit with an evil power that may be directed toward one's enemies. These beings possess what is called mana (supernatural power), a Melanesian term that can be applied both to spirits and to persons of special status, such as chiefs or shamans. In primitive religions, the spirits of nature are generally venerated in return for certain favours or to ward off catastrophes, much in the manner of the religion of ancient Rome. Ancestor gods abound, and thus the ghosts of the dead must be placated, often with the performance of elaborate rites (see also PRIMITIVE RELIGION; SHAMANISM; ANCESTOR WORSHIP).

Conclusion. Though traditional beliefs in angels and demons have been questioned among those cultures affected by Western science and technology, reinterpretations of such beliefs, under the influence of psychological studies and the study of myth in the history of religions, have been of significance to theological reflection. By viewing angels and demons functionally, rather than in terms of their natures, modern man may discover that he has a greater kinship than he has generally realized with men of previous or different cultures in his attempt to gain an advantageous rapport with the transcendent, social, and psychological realms that he faces in everyday life.

BIBLIOGRAPHY. C. JOUCO BLEEKER and GEO WIDENGREN (eds.), *Historia Religionum: Handbook for the History of Religions*, vol. 1, *Religions of the Past* (1969), and vol. 2, *Religions of the Present* (1971), contains helpful sections on the role of angels and demons in chapters on the various religions, as well as a very usable bibliography. J.B. NOSS, *Man's Religions*, 4th ed. (1969), is a good one-volume work that contains useful sections on angels and demons in the chapters dealing with the various religions and also helpful general bibliographies. GUSTAV DAVIDSON, *A Dictionary of Angels, Including the Fallen Angels* (1967); and RUSSELL H. ROBBINS, *The Encyclopedia of Witchcraft and Demonology* (1959), containing useful information for the general reader, are Western oriented with illustrations and selected bibliographies. R.C. ZAEHNER, *The Dawn and Twilight of Zoroastrianism* (1961), has excellent sections on the role of angels and demons in Zoroastrianism and their relationship to Hindu spiritual beings. ROBERT M. GRANT, *Gnosticism and Early Christianity*, 2nd ed. (1966), contains useful sections relating angelic and demonic figures of Judaism, Christianity, and Zoroastrianism to Gnostic speculation. DENIS DE ROUGEMONT, *La Part du diable* (1944; Eng. trans., *The Devil's Share*, 1945), develops the theme, for the general reader, that demonic forces are still at work in the modern world.

(L.F.)

Angiosperm

The angiosperms, more generally known as the flowering plants, constitute the largest and most conspicuous group of modern plants. They predominate in number of species—perhaps up to 250,000—and in number of individuals, and constitute the dominant vegetation of the Earth's surface today. Even in coniferous forests, in moss and lichen tundra, and in fern and cycad thickets, angiosperms often play a conspicuous role. The taxonomic disposition of the flowering plants is somewhat unsettled; some earlier classifications list them as a class, Angiospermae (of the division Spermatophyta), but more recent systems raise the group to the level of a division (roughly comparable to the "phylum" rank among animals), called Anthophyta, or more recently, Magnoliophyta.

GENERAL FEATURES

Angiosperms inhabit the most diversified environments and may be found wherever higher plants can survive. They have evolved a remarkable variety of life forms, a much greater variety than is to be found in the other divisions of the plant kingdom. They range in life span from a few weeks to several thousand years. Angiosperms provide man with food, shelter, clothing, drugs, and ornamental plants. The most important of all sources of food are such angiosperms as the cereal grasses (especially rice; wheat; and maize, or corn), sugarcane, potatoes, tuberous vegetables (yams, taro, etc.), and fruits.

Vegetable fats, oils, and masticatories are also of great value. Besides providing nourishment for man, fats and oils are used extensively in technology, medicine, and the perfume industry. Spices are supplied by such angiospermous plants as pepper, laurel, clove, cinnamon, vanilla, and others. Plants containing stimulants are of great importance, especially tea, coffee, and cacao. A special place in man's life has been reserved for plants containing narcotics that are widely used in medicine (*e.g.*, cocaine and morphine). Textile plants (*e.g.*, cotton and flax) are still very important for the manufacture of clothing. But besides textiles for clothing, vegetable fibres serve a wide variety of purposes. Many trees yield useful woods, a number of which are of high commercial value. Chemicals of various kinds are also derived from wood. More than 10,000 species of flowering plants are used in medicine. The number of species used in industry is considerably less, but some of them, especially the tropical rubber trees, are exceptionally important. Finally, angiosperms ornament the land and provide the most natural human environment.

Diversity of the flowering plants

Flowering plants present an immense diversity in habit, form, and size, ranging from the minute, almost microscopic duckweed (the smallest known angiosperm) to gigantic trees like eucalyptuses (see Figure 1). Most angiosperms grow erect upon the ground, but many of them are trailing, climbing, or physically supported on other plants. The majority are terrestrial. Most of them contain chlorophyll and can manufacture their own food from raw materials, but some (as saprophytes) absorb their food from decaying plants or (as parasites) obtain their food from other living plants. Members of several families are carnivorous, like the sundew (*Drosera*), obtaining a part of their food from insects and other small animals.

Although many angiosperms are woody (trees and shrubs), the majority are herbaceous. It is now widely accepted that the arborescent habit is in general primitive in angiosperms and that all herbaceous plants originated from woody ancestors; the exceptions include woody representatives of the goosefoot family (Chenopodiaceae) and all of the aborescent monocots (palms, screw pines, bamboos, and others). Herbaceous angiosperms have more advanced characters than woody angiosperms and are more plastic in terms of evolution. In herbs, the reproductive phase comes much earlier and with a minimum expenditure of material for construction of vegetative organs, while production of seeds is at a maximum in comparison to the vegetative mass. Furthermore, herbaceous species commonly expand their geographic range more rapidly than trees. Herbs have expanded rapidly over the Earth and play an important role in global vegetation. The development of herbaceous angiosperms has had special significance for the evolution of herbivorous mammals and ground-feeding birds.

LIFE CYCLE

The typical angiosperm plant body—with its leaves, stems, and roots—is the conspicuous phase of the angiosperm life cycle and is called the spore-bearing (sporophyte) generation; each of its living cells has at least two complete sets of chromosomes, twice as many as the cells of the gamete-bearing (gametophyte) generation that follows. During flowering this plant produces microspores, which develop into male gametophytes, and megaspores, which develop into female gametophytes.

As in gymnosperms, both male and female gametophytes are reduced to microscopic proportions. The male gametophyte is ultimately shed as the pollen grain; the female gametophyte, the embryo sac, is retained on the sporophyte in the ovule. If suitable pollen grains are trapped on the receptive stigmatic surface of a flower, they may germinate and one may fertilize the egg nucleus

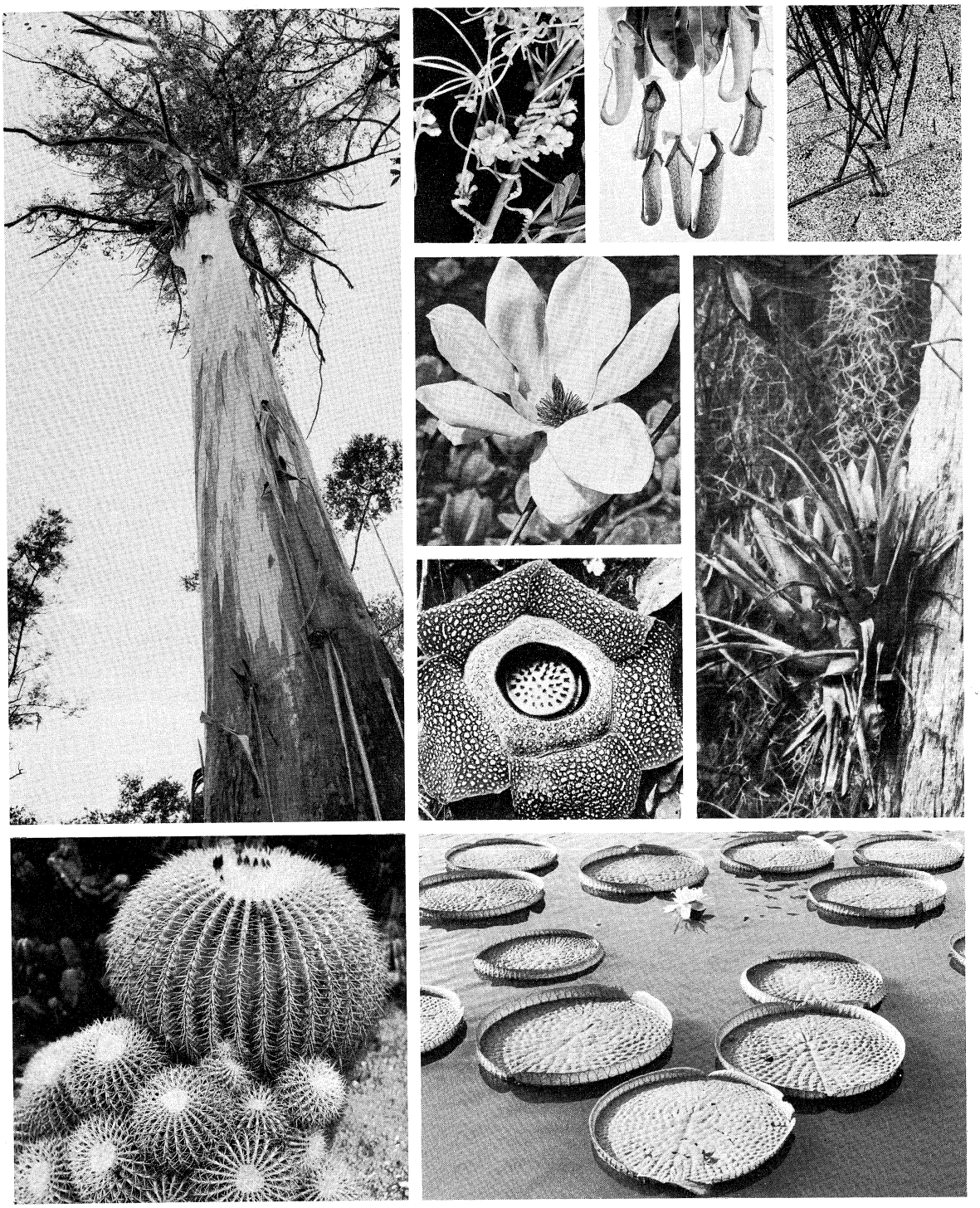


Figure 1: *Diversity among flowering plants.*
(Top left) Australian eucalyptus tree (*Eucalyptus regnans*, family Myrtaceae), largest of the flowering plants. Top row: (left) Dodder (*Cuscuta*, family Convolvulaceae), a climbing parasite; (centre) Old World pitcher plant (*Nepenthes*), a photosynthetic plant that captures insects; (right) duckweed (*Lemna*, family Lemnaceae), one of the smallest of the flowering plants. Centre: (above) Blossom of saucer magnolia (*Magnolia soulangeana*, family Magnoliaceae), a primitive flowering plant; (below) Sumatran rafflesia flower, world's largest known flower. The flower of this parasitic plant, which has no green parts, is bright red and fleshy. (Centre right) A bromeliad, an epiphytic air plant that grows upon (but is not parasitic on) trees and shrubs. (Bottom left) Golden barrel cactus (*Echinocereus grusonii*, family Cactaceae), a xerophyte. (Bottom right) Victoria water lily (*Victoria regia*), a hydrophyte.
(Top left) Douglass Baglin; top row: (left) Ray Kriner—Grant Heilman, (centre) W.H. Hodge, (right) Dennis Brokaw; centre: (above) A.W. Ambler—National Audubon Society; (below) W.H. Hodge; (centre right) Russ Kinne—Photo Researchers; (bottom left) W.H. Hodge; (bottom right) John H. Gerard

in the embryo sac. A zygote is formed and the ovule develops into the seed. The seed ripens and awaits germination to become the conspicuous sporophyte plant body. Thus the cycle is complete.

Pollination and fertilization. The vast majority of angiosperms are normally characterized by cross-pollination, or the transfer of pollen from one flower or plant to another, but self-pollination, or the transfer of pollen from anther to stigma of the same flower is not uncommon. Cross-pollination has the advantage of producing new combinations of existing genes, and the offspring that arise from a cross are often superior in vigour

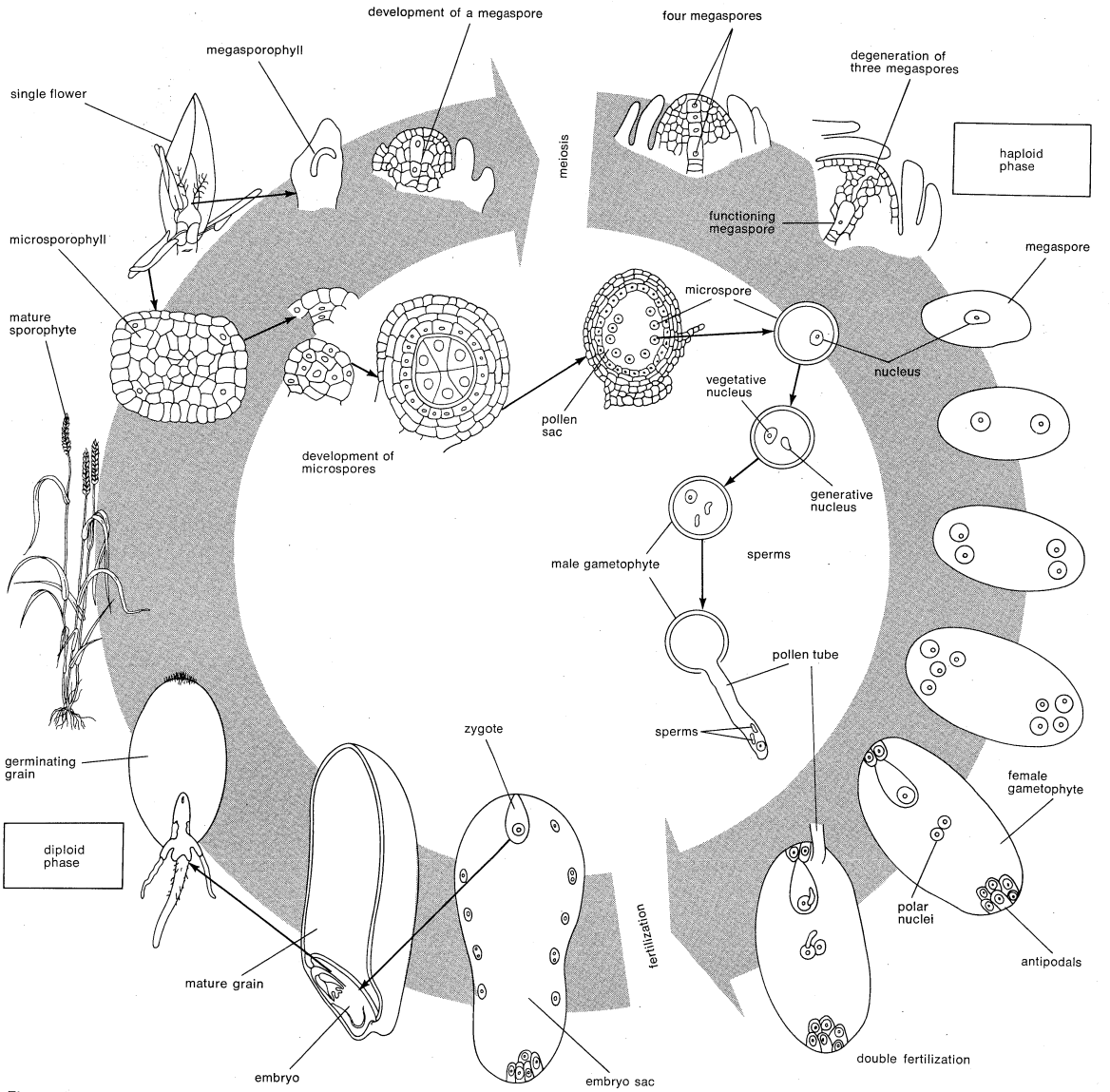


Figure 2: Life cycle of a representative angiosperm.

to those in which both gametes came from the same plant. A great variety of adaptations, both in flower behaviour and structure, prevent or restrict self-pollination. One of the commonest and most effective of these mechanisms is self-sterility (see POLLINATION).

Shortly after the pollen-grain reaches the stigma (see Figure 2), it germinates and the emerging pollen tube grows down the inside of the style and into the cavity of the ovary. The tube nucleus is carried near the tip of the pollen tube, followed by the generative cell. While the tube is developing (or earlier, even before the pollen grain is shed), the generative cell divides into two male gametes. In most cases the pollen tube enters the ovule through a tiny opening, the micropyle. Less commonly the pollen tube penetrates the basal (chalazal) region of the ovule.

After its entrance into the female gametophyte, the tip of the pollen tube ruptures, discharging the two male gametes into the cytoplasm of the female gametophyte. One of the male gametes penetrates the egg and unites with the female nucleus, thus effecting fertilization. The resulting zygote normally develops into the embryo. The other male gamete unites with two other nuclei (the polar nuclei) of the female gametophyte (or with a nucleus formed by prior fusion of the polar nuclei), thus forming the endosperm nucleus. The endosperm nucleus gives rise to a food storage tissue called endosperm. The participation of each of the two male gametes in a fusion process is unique to the angiosperms, and is called double fertilization.

Double fertilization

Embryo and seed development. The development of the embryo from the zygote begins with rapid nuclear divisions accompanied by the formation of cell walls. The early stages of embryo development within the seed are similar in both classes of the angiosperms. But further development proceeds in different ways: in dicotyledons a lateral seed leaf (cotyledon) arises on each side of the stem primordium, while in the monocotyledons only one of two primordial cotyledons develops and is displaced toward the apex. The developmental process clearly indicates that the monocotyledonous embryo arose from the dicotyledonous type by way of a distinct alteration in early stages of development. In dicotyledons the embryo is typically differentiated into four main parts: the radicle, or the embryonic root; the hypocotyl, or the transition region from root structure to stem structure; the cotyledons, attached to the upper end of the hypocotyl; and the plumule, or the bud of the embryo, inserted between the cotyledons (see SEED AND FRUIT).

The primary endosperm nucleus divides to form the endosperm, or food storage tissue of the seed, which serves to nourish the developing embryo and often, later, the germinating seed. From the evolutionary point of view the endosperm is a "new" structure characteristic of the angiosperms only. Often, however, the endosperm degenerates before the seed is mature, and the food is

stored in the thickened cotyledons (as in beans, peas, and other legumes) or in other parts of the embryo, especially the hypocotyl.

The seed of a primitive angiosperm is characterized by abundant endosperm and a minute, undifferentiated embryo. The seeds of more advanced groups usually have scanty endosperm or none at all and a large, differentiated embryo.

The nucellus is generally resorbed during development of the seed, but in some groups it persists and gives rise to a food storage tissue called perisperm. The integument develops into the seed coat (testa), a protective covering for the embryo.

The ripened seeds are enclosed in a fruit, which is exclusively an angiosperm organ specialized for protection and dispersal of the seeds. A fruit is a modified gynoecium (mainly ovary), often with adherent parts of the flower.

Fruits may be divided into two major types: dry and fleshy. Dry fruits may be further subdivided into two groups depending upon the number of seeds and whether the fruit opens at maturity (dehiscent) or remains closed (indehiscent). Fleshy fruits are indehiscent. When a group of fruits develops from a cluster of distinct but closely packed flowers, as in the mulberry and pineapple, it is called a multiple fruit. This conventional classification of fruits, though empirical, is useful for the classification and identification of angiosperms (see SEED AND FRUIT). But for phylogenetic studies it would be more useful to have an evolutional classification of fruits based on fundamental features of the flowers from which the fruits are derived.

FORM AND FUNCTION

Vegetative organs. Like other vascular plants, angiosperms have specialized conducting tissues. The xylem serves in the conduction of water and minerals, and also provides strength. The phloem serves to conduct dissolved food and minerals. In a tree the wood is the xylem and the bark is the phloem. Only the inner (live) bark serves in conduction, and in a large tree only the outer part of the wood serves in conduction. The xylem of most angiosperms has specialized conducting tubes called vessels, but in certain types (aquatics, parasites, saprophytes, and succulents) the vessels have been secondarily lost during evolution. A few primitive angiosperms, however, never had vessels; those include members of the families Winteraceae, Trochodendraceae, and Tetracentraceae, among others. The essential cell type in phloem is the sieve element. In most angiosperms the sieve elements are joined end to end to form sieve tubes, with the sieve areas concentrated onto the sieve plate at the end. Furthermore, each sieve tube is in longitudinal contact with another kind of specialized cell, the companion cell. Gymnosperms, in contrast, have a simpler type of phloem, lacking companion cells, and with the sieve areas not concentrated onto terminal sieve plates. In some of the more primitive angiosperms, such as *Austrobaileya*, the phloem is not very different from that of gymnosperms.

Leaves

The leaves of angiosperms are endlessly variable in size, shape, and structure. Their arrangement on the stem may be alternate, opposite, or whorled. The alternate arrangement is primitive, opposite- and whorled-leaf arrangements having been derived in the course of evolution. The leaves of primitive angiosperms are mostly simple, entire, pinnately veined, leathery, and smooth. The simple, entire leaf with pinnate venation is primitive. Leaves shed in the winter or in the dry season are secondary and have originated from "evergreen" leaves. The deciduous habit may have evolved primarily in the tropics as an adaptation for survival during the dry season and permitted such forms to spread from the tropical zone to more seasonal belts. But most deciduous woody forms have originated in the course of angiosperm dispersal from the lower to the higher latitudes.

Leaves and other aerial green parts have specialized pores called stomata (singular stoma) in the epidermis. These openings are spaces between two specialized crescent-shaped epidermal cells known as guard cells. In many angiosperms the stomata are associated with two or more cells that differ in shape from the other epidermal cells. These distinct cells are called subsidiary, or accessory, cells of the stoma. A few main types of stomatal complexes based on number and arrangement of subsidiary cells have been recognized among the angiosperms. The stomatal complexes are providing an increasingly important source of taxonomic characters.

Two types of root structure are apparent among angiosperms, corresponding in general to the division into monocotyledons and dictyledons (see the Table below).

Reproductive organs. The outstanding structure in the life history of angiosperms is the flower, a highly modified and specialized reproductive shoot in which the spore-bearing organs or sporophylls (stamens and carpels) are borne on a more or less shortened axis (receptacle) and are usually surrounded by a perianth (sepals and petals). A flower that has both stamens and carpels is termed bisexual (or, according to more accurate terminology, amphisporangiate). If either the stamens or the carpels are absent, the flower is unisexual (or, more accurately, monosporangiate) and is either male (staminate or microsporangiate) or female (carpellate or megasporangiate). When a plant bears either staminate or carpellate flowers, it is called dioecious; when it bears both staminate and carpellate flowers, it is called monoecious. There are no essential differences between the flower of angiosperms and the strobilus of gymnosperms. It is usually said that angiosperms differ from gymnosperms in having the ovules (potential seeds) borne within a closed cavity, the ovary, formed by the carpel (or carpels). But this difference is not clear-cut; the carpel is open at the time of pollination in some primitive angiosperms, and in some gymnosperms (*Araucaria*) the seed is enclosed in the seed scale. (For details of floral structure, see REPRODUCTIVE SYSTEMS, PLANT.)

The receptacle is a stem with internodes much shortened or suppressed. The most primitive flowers have an elongated receptacle with an indefinite and variable large number of floral parts arranged spirally upon it. In the course of flower evolution the receptacle has become shorter, and the parts fewer, fixed in number and arranged in whorls. These flowers are characteristic of the advanced angiosperms.

Sepals and petals

The sepals (collectively called the calyx) are usually green and more or less leaflike or bractlike, but they may be reduced. Morphologically, they are specialized leaves. The main function of the calyx is the protection of the inner and more delicate floral parts as these develop in the bud. In some genera (*Trollius*, *Lilium*, *Tulipa*, *Allium*, and others) sepals are much like the petals and scarcely distinguishable except by position and by vascular anatomy.

The petals (collectively called the corolla) usually alternate with the sepals and differ from them in being usually more delicate in texture, larger in size, and of some colour other than green. In some primitive angiosperms (Magnoliales, Illiciales, etc.), all the perianth parts are morphologically equivalent and of leaf origin. But in most families the petals have originated as an evolutionary petaloid elaboration of stamens. Thus the corolla has arisen in two ways, and the petals of roses, for example, are not morphologically equivalent to those of magnolias. In several advanced orders the petals are united to form a gamopetalous or sympetalous (as opposed to a polypetalous) corolla. The main function of the petals is to attract pollinators. In many angiosperms the perianth consists of sepals only, which may be scalelike, as in the oaks, or large and showy, as in the anemones, taking the place of the corolla. Such flowers, with the perianth consisting of sepals only, are called apetalous. The absence of a corolla is secondary rather than primitive, at least in the great majority of angiosperm species.

Stamens

The stamens (microsporophylls) are collectively called the androecium. Each stamen is usually differentiated into a stalk (filament) and a distal, spore-bearing portion (anther). The more advanced types of stamens are characterized by a distinct slender filament, whereas in many

primitive angiosperms (*Degeneria, Himantandra,* some Magnoliaceae, and others) the stamens are broad, flat structures with no differentiation of a filament. The anther typically consists of four pollen sacs (microsporangia), two in each lobe or anther half, the tissue separating them disintegrating prior to flowering, and the resulting anther seemingly two-celled (biloculate). The strip of tissue between the pairs of microsporangia is called the connective. In primitive stamens the connective is produced beyond the microsporangia, forming a more or less distal appendage. In primitive groups, the microsporangia are long and narrow.

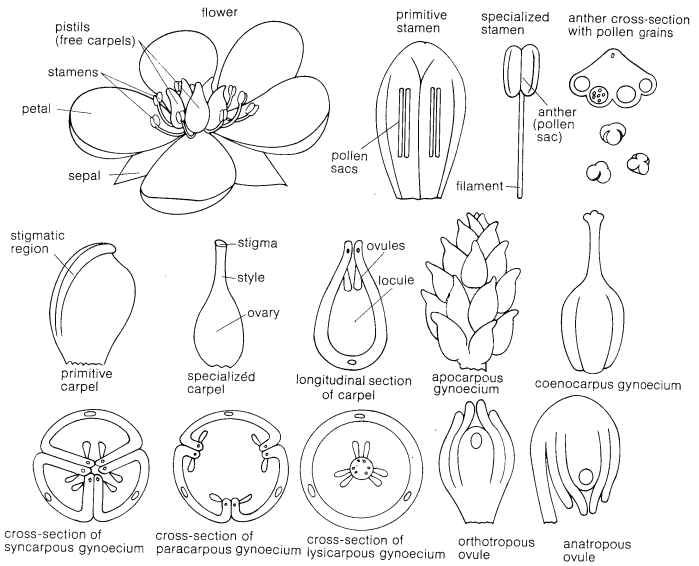


Figure 3: Floral structures characteristic of angiosperms.

The primary spore-forming cells in a microsporangium give rise to a large number of diploid ($2n$) spore mother cells, each of which produces by two cell divisions a group of four haploid (n) microspores. The microspore, after germination, when the cells of the male gametophyte appear within the wall, becomes the pollen grain. Each pollen grain contains a greatly reduced male gametophyte, which consists of two cells only—a small generative cell and a larger tube cell or "vegetative" cell. The microspore (and pollen grain) has a cell wall of two layers: an outer (exine) and an inner (intine). The exine is more or less sculptured in patterns that characterize particular species. The cell wall usually has one or more thin places (apertures) where the exine is not fully formed and through which the pollen tube later may emerge. There are two major types of germinal apertures: furrows (colpi; singular colpa) and pores. The most primitive angiosperm pollen grain is the monocolpate type, with one polar germinal furrow. Such monocolpate pollen grains have a continuous aperture membrane devoid of special openings (ora; singular os) in the exine for the emergence of the pollen tube. Monocolpate pollen is characteristic of the Magnoliales and some other primitive dicotyledons, of the majority of monocotyledons, and of many gymnosperms. In the course of evolution, several other types of pollen grains have been derived, differing in number of furrows or pores.

Carpels

The carpels (megasporophylls) are collectively called the gynoecium. The carpel consists of the ovary with enclosed ovules, and the stigma, a receptive surface for pollen grains, typically borne at the apex of a stalk, called the style (or stylodium, as some authors prefer to name the simple style of an individual carpel). The most primitive carpel is reminiscent of a young leaf folded along the midrib; such a carpel is called conduplicate and is characteristic of certain primitive angiosperms—*e.g.*, the genera *Degeneria, Tasmannia,* and to a lesser degree the family Magnoliaceae. A very important taxonomic characteristic of primitive carpels is the absence of styles, the stigmatic surface being along the margins (sutures) of the carpels. Of the two basic types of ovule arrangement (placentation) within the carpel, laminar (superficial) and sutural (submarginal), the former is probably more primitive.

The gynoecium may consist of one to many carpels. The carpels may be arranged either in a spiral or a whorl on the receptacle. The most primitive types of gynoecia are characterized by a large number of carpels arranged spirally upon a long receptacle, as in *Magnolia.* In the course of gynoecium evolution the number of carpels typically decreases and the whorled arrangement evolves. Gynoecia with individually distinct (free) carpels are called apocarpous; such gynoecia are characteristic for the more primitive orders of angiosperms, as Magnoliales, Ranunculales, Dilleniales, Alismales, and others. During their evolution carpels have gradually united with each other in various forms, generally termed coenocarpous gynoecia. There are technically three types of coenocarpous gynoecia: syncarpous (2—or more—locular, with axile placentation, as in *Lilium*); paracarpous (unilocular, with parietal placentation, as in *Viola*); and lysicarpous (unilocular, with free central placentation, as in *Primula*). In descriptive morphology of plants, however, all gynoecia with united carpels are termed syncarpous. The term pistil has been a source of confusion as it is used both for a single carpel and for coenocarpous gynoecia; some authors suggest abandoning it.

In the less advanced types of coenocarpous gynoecia the individual styles of carpels are more or less free and only the ovaries are fused, as in *Hypericum.* In more advanced gynoecia, the styles and their stigmas are united (compound), as in *Rhododendron.* In more primitive groups, the flower is hypogynous, the gynoecium being situated above the place of attachment of the perianth and stamens, and the ovary is termed superior. In more advanced groups, the flower is frequently epigynous, the ovarian part of the gynoecium being fused with the basal part of the perianth, or in a few cases embedded in the receptacle, and the ovary is termed inferior (the perianth and stamens seem to arise from its summit).

The ovule, destined to become a seed, consists of the nucellus (megasporangium), completely surrounded by one or two coats (integuments), except for a narrow opening (micropyle) at the top of the ovule. Each ovule is borne on a stalk (funiculus) and may be erect (orthotropous) or, more commonly, recurved (anatropous). The basal part of the ovule, where all its parts merge, is called the chalaza. In more primitive angiosperms the ovules are massive and vested with two coats (integuments). In more advanced groups the ovules are small and have only one integument. In some representatives of the order Santalales, especially in the families Loranthaceae and Balanophoraceae, the ovule is reduced, has no integuments, and the sporangium is naked.

The female gametophyte (embryo sac) develops within the megasporangium and most commonly forms by the division of the innermost megaspore of the tetrad of potential megaspores, the other three degenerating. The functional megaspore greatly enlarges, absorbing the abortive megaspores and some of the surrounding nucellar tissue. The nucleus of the functional megaspore divides into two daughter nuclei, which move to opposite ends of the greatly elongated megaspore. Each daughter nucleus then divides, and ordinarily two in turn soon give rise to two groups of four cells (eight in all). One nucleus from each polar group moves to the centre of the young gametophyte. These two nuclei, called polar nuclei, come in contact with each other and may fuse immediately or remain distinct for a time. In the gametophyte there are soon formed three cells at the micropylar end (one is the female gamete, or egg, and the other two are called synergids) and three cells at the chalazal end (called antipodals). Thus, the female gametophyte of angiosperms is extremely simplified and completely devoid of all traces of gametangia.

EVOLUTION AND PALEONTOLOGY

Origin of angiosperms. The angiosperms appear to have arisen from some gymnosperm ancestors in a single

lineage (monophyly). It is quite possible that the angiosperms were connected with the Bennettitales (a group related to living cycads) through a common ancestor as yet unknown but possibly some primitive seed fern.

The first angiosperms

The first angiosperms (and their immediate gymnosperm precursors) may have been inhabitants of the mountains, where they probably formed small, semi-isolated local populations. Such populations are common in mountain plants and provide the combination of genetic drift and selection that is most effective in promoting rapid evolution. If this supposition is correct, then evolution in the immediate ancestors of angiosperms and in the earliest angiosperms themselves must have been very rapid and must have facilitated the production of large adaptive modifications. Under such conditions, one of the ancient branches of gymnosperms might well have initiated a more progressive group of seed plants characterized by high adaptability.

The basic morphological pecularities of angiosperms find their most likely explanation in the hypothesis of their neotenic origin (retention of early developmental stages). Thus, the flower may be regarded as a neotenic form of the strobilus of an ancestral gymnosperm. The carpels of the most primitive angiosperms, especially those reminescent of young folded leaves in *Degeneria* and *Tasmannia*, give the impression of being "infantile" organs corresponding to the juvenile stages of sporophylls of an ancestral gymnosperm. The evolutionary transformation of the open sporophyll into the carpel by folding and gradual closure along its middle vein could easily have occurred in a juvenile stage. The transformation of open megasporophylls into more or less closed carpels may have occurred at a stage in development when they were still in a folded condition. The stamens of angiosperms arose most likely from the juvenile stage of the microsporophylls of an ancestral gymnosperm. Traces of neotenic origin are also visible, though less obvious, in the morphology of angiosperm leaves. The simple and entire leaves of primitive angiosperms, like those of *Magnolia*, could have occurred by arrested development of an ancestral gymnosperm at an early juvenile stage. The primitive structure of the wood of early vessel-less angiosperms is also understandable as a neotenic adaptation: it corresponds to the juvenile wood of gymnosperms.

The flower arose most probably from a bisexual strobilus, a supposition bolstered by the discovery of bisexuality associated with insect pollination among certain fossil gymnosperms. Insects must have played a decisive role in the transformation of a gymnosperm strobilus into the angiosperm flower. Insects visiting the strobili of the ancestors of the angiosperms found shelter therein, attracted by the abundance of pollen grains on which they fed. At the same time, they facilitated cross-pollination. But they also ate the ovules. So this primitive type of pollination may have led to protection of the ovules. In the immediate ancestors of angiosperms this may have resulted in a partial closure of the folded juvenile megasporophyll. Even in the primitive carpels of the early angiosperms the young succulent ovules were protected not only from the primitive unspecialized pollinators but also from the suctorial insects of later epochs. Though they probably arose in the first place only as a defensive adaptation against certain types of insects, the carpels also gave protection from other harmful external influences, particularly atmospheric dryness. Inside the carpels, new environmental conditions were created for the development of the ovule and for the process of fertilization.

Origin of carpels

The origin of carpels made it possible for angiosperms to simplify the structure of their ovules, and this in turn led to smaller ovules, with a reduced gametophyte, capable of much faster development. This was a very important advance, as it enabled the angiosperms to observe the greatest possible economy of material in the construction of the ovules and female gametophyte. The development of carpels also led to the formation of the stigmatic surface, the further evolution of which made possible the perfection of the process of pollination. The acquisition of the stigma was undoubtedly another great event in the evolutionary history of flowering plants. It improved the possibility of pollination, served as a new barrier to "undesirable" pollen and thereby assisted cross-pollination, and still further facilitated the simplification of the ovule. Finally, the evolution of carpels made possible both coenocarpy (union of carpels) and angiocarpy (the enclosure of the carpels by other parts of the flower), which undoubtedly had great adaptive significance, as improving changes of reproductive success.

Fossil evidence. The oldest fully authentic fossil angiosperms are known from the Early Cretaceous Period (130,000,000 years ago), toward the end of which period the frequency and abundance of angiosperms noticeably increased. Fossil angiosperms appear in the company of various gymnosperms (including the Bennettitales) and ferns and predominate only in a few cases. They evidently formed only small populations and were represented by a comparatively small number of individuals. Toward the middle of the Cretaceous Period (100,000,000 years ago), however, one of the most sudden and fundamental transformations of terrestrial plant life occurred, and in the course of a few million years (a very short interval in geological time) the angiosperms came to be widely distributed throughout the world, quickly reaching the Arctic and Antarctic regions. They appeared in great diversity and quickly became dominant, decisively influencing the future of the terrestrial animal world, especially the insects, birds, and mammals, and in the final reckoning made possible the appearance of man.

Opinions concerning the location of angiosperm origin

While the most diverse opinions can be found in the literature concerning the centre of origin and distribution of the angiosperms, an analysis of the present distribution of the more primitive living angiosperms suggests that they arose somewhere in the Old World tropics, in what is now Southeast Asia, where the relatively greatest concentration of primitive angiosperms is found.

CLASSIFICATION

Distinguishing taxonomic features. The classification of angiosperms is based on the synthesis of all of the available data from many fields: comparative morphology, palynology, embryology, cytology, genetics, biochemistry, serology, and biogeography. Although computers and the scanning electron microscope are being increasingly employed in the development of modern methods in systematics, in practice plant taxonomists still rely to a large extent on morphological features to subdivide the angiosperms into orders and families and to determine relationships between them. These features include principally: numbers and arrangement of flower parts; disposition of the flowers on the plant; the type of gynoecium and placentation; the structure of the carpels; the type and structure of the ovules and the number of integuments; the presence or absence of endosperm and perisperm; the structure and development of the androecium, pollen grains, and perianth; leaf structure and arrangement; stomatal patterns; and structure of the wood.

The division Magnoliophyta is divided into two classes, the Magnoliopsida (dicotyledons) and Liliopsida (monocotyledons). Some of the differences between these classes are shown in the Table. Dicots and monocots are distinguished from each other by combinations of characters rather than by any single character. In fact there is no one reliable single character that can serve to separate the monocotyledons from the dicotyledons. Even the number of cotyledons, which formed the basis for the original subdivision of the angiosperms into two groups, may also fluctuate: in *Agapanthus*, a typical monocotyledonous plant of the order Liliales, some embryos have two cotyledons; whereas *Ranunculus ficaria*, some species of *Corydalis*, some Gesneriaceae, and several other dicotyledonous plants have embryos with a single cotyledon.

Evidence indicates that the monocotyledons originated from some long extinct vessel-less herbaceous dicotyledon with apocarpous gynoecia and monocolpate pollen grains (characters much in common with the modern Nymphaeales). The order Nymphaeales is most likely a hydrophilous derivative of some ancient Magnoliales, the vegetative organs of which have more or less degenerated

Comparison of Dicotyledons and Monocotyledons

Magnoliopsida (dicotyledons)	Liliopsida (monocotyledons)
embryo usually with 2 cotyledons	embryo with 1 cotyledon
leaves usually with (1) reticulate (netted) venation (pinnate or palmate), (2) seldom sheathed at the base, (3) petiole usually well developed	leaves usually with (1) parallel venation, (2) often sheathed at the base, (3) petiole seldom developed
bracteoles (when present) usually 2, lateral	bracteole (when present) usually 1, adaxial
vascular bundles usually borne in a ring, with a persisting cambium; cortex and pith are usually well differentiated	vascular bundles generally scattered or seldom in 2 or more rings, usually without a cambium; usually there are no definable cortex and pith
primary root generally persistent and becoming a taproot; the rootcap and piliferous layer have common origin (with the exception of Nymphaeales)	primary root of short duration, soon replaced by adventitious roots; the rootcap and piliferous layer of different origin
woody and herbaceous (primarily woody)	herbaceous and arborescent (primarily herbaceous)
floral parts usually borne in sets of 5 or 4 (seldom 3 or fewer)	floral parts typically borne in sets of 3 (seldom 4 and never 5)
nectaries of various types, frequently transformed stamens, and rarely of the septal type	nectaries chiefly of the septal type (occur in the septa between carpels)
pollen grains usually tricolpate or tricolpate-derived (monocolpate only in a few primitive families)	pollen grains monocolpate or monocolpate-derived type (never tricolpate)

in an aquatic environment. Such ancient herbaceous dicotyledons of nymphaealean type, but more primitive (less reduced) than their present day representatives, quite probably initiated the monocotyledonous line.

In comparison with the majority of dicotyledons, the monocotyledons are characterized by some degree of "infantilism" of their vegetative organs. They usually have no vascular cambium, their primary root is usually ephemeral, the leaves are not differentiated into petiole and blade (or the differentiation is not clear-cut as in the dicotyledons), the leaf venation resembles the venation of the underdeveloped leaf organs of the dicotyledons (like bud scales or bracts), and one of the two cotyledons is suppressed. The earliest monocotyledons were most probably marsh plants with stout rhizomes. Indeed, the majority of Alismales, as well as almost all the Nymphaeales (except *Ceratophyllun*), are typically aquatic geophytes (herbaceous plants with underground storage organs). Apparently the common ancestors of both the Nymphaeales and the monocotyledons were amphibious geophytes that adapted to terrestial conditions under the forest canopy or in the forest margin.

Annotated classification. The classification employed in this article is based upon that of Armen Takhtajan (1969, 1970), with later modifications. A complete listing of all angiosperm orders is given below. Additional articles about the most important orders contain complete family listings and detailed information on their natural history, economic importance, morphology, and evolution.

DIVISION MAGNOLIOPHYTA (Angiospermae)

The entire range of flowering plants, or angiosperms. Perhaps 250,000 species of worldwide distribution.

Class Magnoliopsida (*Dicotyledones*)

Plants typically with two dicotyledons, or seed leaves. The magnolia is the typical form (type genus) for which the class is named.

Subclass Magnoliidae

Mostly woody plants, some of them vessel-less and many of them with oil cells. Stomata mostly with 2 subsidiary cells. Flowers mostly bisexual, frequently spiral or spiro-cyclic. Gynoecium mostly apocarpous. Pollen grains 2-celled or 3-celled, monocolpate or of monocolpate-derived type. Ovules with two integuments. Seeds with copious or scanty endosperm or none at all. Six orders.

Order Magnoliales (*Annonales*). The most primitive living order of angiosperms, comprising 8 families, including Magnoliaceae (magnolia), Degeneriaceae (degeneria), Annonaceae (custard-apple), Myristicaceae (nutmeg), Winteraceae (wintera); see MAGNOLIALES.

Order Laurales. Near to the Magnoliales, but more advanced; evidently derived from some ancient vessel-less representatives of the Magnoliales. Families number 10, including Austrobaileyaceae (austrobaileya), Monimiaceae (monimia), Calycanthaceae (calycanthus), Lauraceae (laurel); see LAURALES.

Order Piperales. Nearest to the Laurales, but more advanced. Two families, Saururaceae (lizard's-tail) and Piperaceae (pepper); see PIPERALES.

Order Aristolochiales. Probably derived from the Magnoliales. Family Aristolochiaceae (birthwort); see ARISTOLOCHIALES.

Order Rafflesiales. Probably derived from the ancestors of the Aristolochiales. Two families, Rafflesiaceae (rafflesia) and Hydnoraceae.

Order Nymphaeales. Probably derived from some ancient vessel-less stock of the Magnoliales. They are all aquatics. Four families, including Nymphaeceae (water lily) and Ceratophyllaceae (hornwort); see NYMPHAEALES.

Subclass Ranunculidae

Near to the Magnoliidae, but more advanced. Mostly herbaceous plants, with vessels and usually without oil cells. Stomata of various types, most frequently without subsidiary cells. Flowers bisexual or unisexual, frequently spiral or spirocyclic. Pollen grains mostly 2-celled, tricolpate or of tricolpate-derived type. Ovules usually with two integuments. Seeds usually with copious endosperm. Five orders.

Order Illiciales. Evidently derived from the Magnoliales, most probably from the ancestors of the Winteraceae. Two families, Illiciaceae (illicium) and Schisandraceae (schisandra).

Order Nelumbonales. Usually placed in or next to the Nymphaeaceae but differs in many important features. A link with the Illiciales–Ranunculales stock is much more probable. One family, Nelumbonaceae (lotus).

Order Ranunculales. Evidently derived from the ancestors of the Illiciales. Families number 9, including Lardizabalaceae (lardizabala), Menispermaceae (moonseed), Ranunculaceae (buttercup), Berberidaceae (barberry); see RANUNCULALES.

Order Papaverales. Very near to the Ranunculales, but gynoecium always paracarpous. Three families, Papaveraceae (poppy), Fumariaceae (fumitory), and Hypecoaceae; see PAPAVERALES.

Order Sarraceniales. A very specialized order, but with some primitive features that place it near to the Ranunculales. One family, Sarraceniaceae (pitcher-plant); see SARRACENIALES.

Subclass Hamamelididae

Mostly woody plants, usually with vessels (except in the order Trochodendrales). Stomata with 2 or more subsidiary cells, or subsidiary cells are absent. Flowers more or less reduced, mostly unisexual and apetalous, seldom naked. Gynoecium usually coenocarpous. Pollen-grains usually 2-celled, tricolpate or of tricolpate-derived type. Ovules usually with two integuments. Seeds with copious or scanty endosperm or none at all. Orders number 15.

Order Trochodendrales. In many respects occupies position between the Magnoliales and Hamamelidales, nearer to the latter. Two families, Trochodendraceae (trochodendron) and Tetracentraceae (tetracentron).

Order Ceridiphyllales. Near Trochodendrales. One family, Cercidiphyllaceae.

Order Eupteleales. Somewhat isolated systematically. One family, Eupteleaceae.

Order Didymelales. One family, Didymelaceae.

Order Hamamelidales. In many ways a connecting link between the Trochodendrales and the apetalous orders Urticales, Casuarinales, Fagales, etc. The order Hamamelidales was evidently derived from immediate ancestors of the Trochodendrales. Four families, including Hamamelidaceae (witch hazel), Altingiaceae (altingia), Platanaceae (plane tree); see HAMAMELIDALES.

Order Eucommiales. One family, Eucommiaceae.

Order Urticales. Probably derived from the Hamamelidales. Four families—Ulmaceae (elm), Moraceae (mulberry), Cannabaceae (hemp), and Urticaceae (nettle); see URTICALES.

Order Barbeyales. One family, Barbeyaceae.

Order Casuarinales. Evidently derived from the Hamamelidales. One family, Casuarinales (casuarina).

Order Fagales. Probably derived from the Hamamelidales. One family, Fagaceae (beech); see FAGALES.

Order Betuales. Probably had a common origin with the Fagales. One family, Betulaceae (birch); see BETULALES.

Order Balanopales. One family, Balanopaceae.

Order Myricales. Probably derived from the Hamamelidales. One family, Myricaceae (sweet gale).

Order Leitneriales. One family, Leitneriaceae.

Order Juglandales. Probably had a common origin with the Myricales. Two families, Juglandaceae (walnut) and Rhoipteleaceae; see JUGLANDALES.

Subclass Caryophyllidae

Usually herbaceous plants, undershrubs or low shrubs, rarely trees. Vessels always present. Stomata with 2 or 3 (rarely 4) subsidiary cells, or subsidiary cells absent. Flowers bisexual or rarely unisexual, mostly apetalous. Gynoecium apocarpous or coenocarpous. Pollen grains mostly 3-celled, tricolpate or of tricolpate-derived type. Ovules usually with 2 integuments. Seeds frequently with perisperm. Four orders.

Order Caryophyllales (*Chenopodiales*). Probably derived from the Ranunculales. The Phytolaccaceae in particular is linked with the Ranunculales, especially with the Menispermaceae and Lardizabalaceae. Families number 16, including Phytolaccaceae (pokeweed), Nyctaginaceae (four-o'clock), Aizoaceae (mesembryanthemum), Cactaceae (cactus), Portulacaceae (purslane), Caryophyllaceae (pink), Amaranthaceae (amaranth), Chenopodiaceae (goosefoot). See CARYOPHYLLALES; CACTALES.

Order Polygonales. Probably derived from the same stock as the Caryophyllales. One family, Polygonaceae (buckwheat).

Order Plumbaginales. Probably derived from the same stock as the Caryophyllales. One family, Plumbaginaceae (leadwort).

Order Theligonales. One family, Theligonaceae (sometimes called Cynocrambaceae).

Subclass Dilleniidae

Woody or herbaceous plants. Vessels always present. Stomata of various types, mostly without subsidiary cells. Flowers bisexual or unisexual, usually with both calyx and corolla; in the more primitive families, perianth is frequently spiral or spirocyclic. Gynoecium apocarpous or more frequently coenocarpous. Pollen grains mostly 2-celled, tricolpate or of tricolpate-derived type. Ovules usually with 2 integuments. Seeds mosty with endosperm. Orders number 17.

Order Dilleniales. A connecting link between the Magnoliales and the Theales-Violales group. Two families, Dilleniaceae and Crossosomataceae.

Order Paeoniales. Near to the Dilleniales. One family, Paeoniaceae (peony).

Order Theales. Near to the Dilleniales, and evidently derived from early Dilleniaceae. Families number 20, including Theaceae (tea family), Clusiaceae (garcinia), Hypericaceae (hypericum); see THEALES.

Order Violales. Very near to the Theales, and shares with it a common origin from the Dilleniales. Families number 8, including Flacourtiaceae (flacourtia), Violaceae (violet), Cistaceae (rockrose); see VIOLALES.

Order Passiflorales. Derived from primitive representatives of the Violales, and very close to them. Families number 5, including Passifloraceae (passionflower), Caricaceae (carica).

Order Cucurbitales. Very closely related to the family Passifloraceae. One family, Cucurbitaceae (gourd); see CUCURBITALES.

Order Begoniales. Probably derived from the Violales. Two families, Begoniaceae (begonia) and Datiscaceae; see BEGONIALES.

Order Capparales. Derived from the primitive representatives of the Violales, most likely from the Flacourtiaceae. Families number 8, including Capparaceae (caper), Brassicaceae or Cruciferae (mustard); see CAPPARALES.

Order Tamaricales. Derived from the Violales. Three families—Tamaricaceae, Fouquieriaceae, and Frankeniaceae.

Order Salicales. In all probability derived from the Flacourtiaceae. One family, Salicaceae (willow); see SALICALES.

Order Ericales. Derived from the Theales. Families number 10, including Actinidiaceae (actinidia), Clethraceae (pepperbush), Ericaceae (heath), and Monotropaceae (Indian pipe); see ERICALES.

Order Diapensiales. One family, Diapensiaceae.

Order Ebenales. Derived from the Theales. Families number 5, including Styracaceae (storax), Ebenaceae (ebony), Sapotaceae (sapote); see EBENALES.

Order Primulales. Stands fairly close to the Ebenales and shares common origin with it from the Theales. Three families, including Myrsinaceae (myrsine), Primulaceae (primrose); see PRIMULALES.

Order Malvales. Evidently derived from early Violales. Nine families, including Tiliaceae (linden), Bombacaceae (bombax), Malvaceae (mallow); see MALVALES.

Order Euphorbiales. Arose from some ancient group intermediate between the Flacourtiaceae and Malvales. Families number 7, including Buxaceae (boxwood), Euphorbiaceae (spurge); see EUPHORBIALES.

Order Thymelaeales. Shares a common origin with the Euphorbiales and Malvales; *i.e.*, arose from ancestors of the Flacourtiaceae-type. One family, Thymelaeaceae (mezereum).

Subclass Rosidae

Woody or herbaceous plants. Vessels always present. Stomata of various types, most frequently without subsidiary cells or with two subsidiary cells. Flowers mostly bisexual, with both calyx and corolla, or apetalous. Gynoecium apocarpous or coenocarpous. Pollen grains mostly 2-celled, tricolpate or of tricolpate-derived type. Ovules usually with 2 integuments. Seeds with or without endosperm. Orders number 19.

Order Saxifragales. Derived from ancestors of the Dilleniales. Families number 28, including Hydrangeaceae (hydrangea), Crassulaceae (orpine), Saxifragaceae (saxifrage); see SAXIFRAGALES.

Order Rosales. Shares a common origin with the Saxifragales. Three families—Rosaceae (rose), Chrysobalanaceae, and Neuradaceae; see ROSALES.

Order Fabales. Close to the Saxifragales and evidently derived from its immediate ancestors. Three families, including Fabaceae or Leguminosae (pea); see FABALES.

Order Connarales. One family, Connaraceae.

Order Nepenthales. Probably derived from the Saxifragales. Two families, Droseraceae (sundew) and Nepenthaceae (nepenthes); see NEPENTHALES.

Order Podestemales. One family, Podostemaceae.

Order Myrtales. Evidently derived from the Saxifragales. Families number 11, including Lythraceae (loosestrife), Rhizophoraceae (mangrove), Myrtaceae (myrtle), Onagraceae (evening primrose); see MYRTALES.

Order Hippuridales (*Haloragales*). Near Myrtales. Three families—Haloragaceae, Gunneraceae, and Hippuridaceae.

Order Rutales. Probably derived from the Saxifragales. Families number 12, including Anacardiaceae (cashew), Rutaceae (rue), Meliaceae (mahogany); see RUTALES.

Order Sapindales. Evidently had a common origin with the Rutales. Nine families, including Aceraceae (maple), Hippocastanaceae (horse chestnut), Sapindaceae (soapberry); see SAPINDALES.

Order Geraniales. Clearly connected with the Rutales, especially with the Rutaceae. Families number 20, including Linaceae (flax), Erythroxylaceae (coca), Zygophyllaceae (caltrop), Geraniaceae (geranium), Balsaminaceae (balsam); see GERANIALES.

Order Polygalales. Very close to family Malpighiaceae of the Geraniales. Five families—Trigoniaceae, Vochysiaceae, Polygalaceae, Krameriaceae, and Tremandraceae.

Order Cornales (*Umbellales*). Probably derived from the Saxifragales. Fourteen families, including Cornaceae (dogwood), Nyssaceae (nyssa), Araliaceae (ginseng), Apiaceae or Umbelliferae (carrot); see CORNALES.

Order Celastrales. Evidently derived from the Saxifragales. Families number 10, including Aquifoliaceae (holly), Celastraceae (staff tree); see CELASTRALES.

Order Rhamnales. Probably had a common origin with the Celastrales. Three families—Rhamnaceae (buckthorn), Vitaceae (grape), and Leeaceae; see RHAMNALES.

Order Oleales. Probably had a common origin with the Celastrales. One family, Oleaceae (olive); see OLEALES.

Order Santalales. Probably shares a common origin with the Celastrales. Families number 13, including Santalaceae (sandalwood), Loranthaceae (loranthus), Viscaceae (mistletoe); see SANTALALES.

Order Elaeagnales. Probably had a common origin with the Celastrales and Santalales from the Saxifragales. One family, Elaeagnaceae (oleaster).

Order Proteales. Probably shares common origin with the Elaeagnales and the Santalales. One family, Proteaceae (protea).

Subclass Asteridae

Mostly herbaceous plants. Vessels always present. Stomata of various types, most frequently without subsidiary cells. Flowers usually bisexual. Gynoecium coenocarpous (probably morphologically always paracarpous). Pollen grains 3- or 2-celled, tricolpate or of tricolpate-derived type. Small ovules with a single integument. Seeds with or without endosperm. Seven orders.

Order Dipsacales. Related to the Cornales and probably had a common origin with it. The Caprifoliaceae exhibits definite links both with the Cornales and with the Hydrangeaceae. Four families—Caprifoliaceae (honeysuckle), Valerianaceae (valerian), Dipsacaceae (teasel), and Adoxaceae; see DIPSACALES.

Order Gentianales. Shares common origin with the Dipsacales. Families number 11, including Loganiaceae (logania), Apocynaceae (dogbane), Asclepiadaceae (milkweed), Gentianaceae (gentian), Rubiaceae (madder); see GENTIANALES.

Order Polemoniales. Very near to the Gentianales and probably derived from the immediate ancestors of the Loganiaceae and allied families. Eight families, including Polemoniaceae (phlox), Convolvulaceae (morning glory), Boraginaceae (borage); see POLEMONIALES.

Order Scrophulariales. Very near to the Polemoniales, with which they had a common origin. Families number 17, including Solanaceae (nightshade), Scrophulariaceae (figwort), Bignoniaceae (bignonia), Gesneriaceae (gesneria), Acanthaceae (acanthus), Plantaginaceae (plantago); see SCROPHULARIALES.

Order Lamiales. Very near to and derived from the Scrophulariales. Four families, including Verbenaceae (verbena), Lamiaceae or Labiatae (mint); see LAMIALES.

Order Campanulales. Probably shares a common origin with the Polemoniales from the Gentianales. Seven families, including Campanulaceae (bellflower), Lobeliaceae (lobelias); see CAMPANULALES.

Order Calycerales. One family, Calyceraceae.

Order Asterales. Shares a common origin with the Campanulales. One family, Asteraceae or Compositae (composite); see ASTERALES.

Class Liliopsida (*Monocotyledones*)

Plants with a single cotyledon in the seed. The lily is the type genus for which the class is named.

Subclass Alismidae

Herbaceous plants, aquatic or semiaquatic. Vessels absent or only in the roots. Stomata (when present) with 2 or rarely 4 subsidiary cells. Gynoecium mostly apocarpous. Pollen grains usually 3-celled, monocolpate, nonaperturate or 2-polyporate. Ovules with 2 integuments. Seeds without endosperm. Three orders.

Order Alismales. Linked with the Nymphaeales among the dicotyledons. Three families—Butomaceae (flowering rush), Alismaceae (water plantain), and Limnochantaceae; see ALISMALES.

Order Hydrocharitales. Near to and derived from the Alismales. One family, Hydrocharitaceae (frog's-bit).

Order Najadales. Near to the Alismales and probably derived from it. Ten families, foremost of which is Potamogetonaceae (pondwood).

Subclass Liliidae

Herbaceous or rarely arborescent plants. Vessels mostly in the roots only, rarely in the stems and leaves. Stomata mostly without subsidiary cells, rarely with 2 (some representatives of the order Liliales) or 2 or more subsidiary cells (Zingiberales). Gynoecium almost always coenocarpous (mostly syncarpous), very rarely almost apocarpous. Pollen grains usually 2-celled, mostly monocolpate. Seeds with or without endosperm. Five orders.

Order Triuridales. One family, Triuridaceae.

Order Liliales. Shares a common origin with the Alismales. Families number 18, including Liliaceae (lily), Amaryllidaceae (amaryllis), Dioscoreaceae (yam), Agavaceae (agaves), and Asparagaceae (asparagus); see LILIALES.

Order Iridales. Probably derived from the family Liliaceae. Four families, including Iridaceae (iris); see IRIDALES.

Order Zingiberales. Probably derived from the family Liliaceae. Eight families, including Musaceae (banana), Zingiberaceae (ginger), Cannaceae (canna), Marantaceae (arrowroot); see ZINGIBERALES.

Order Orchidales. Probably derived from the Liliales. One family, Orchidaceae (orchid); see ORCHIDALES.

Subclass Commelinidae

Mostly herbaceous plants. Vessels in all vegetative organs, or restricted to roots. Stomata with 2, 4, or 6 subsidiary cells, very rarely without subsidiary cells. Flowers mostly bisexual, with more or less reduced perianth. Gynoecium coenocarpous (syncarpous or paracarpous). Pollen grains 2-celled or 3-celled, mostly monocolpate or monoporate. Ovules usually with 2 integuments. Seeds with usually mealy endosperm. Seven orders.

Order Juncales. Probably had a common origin with the Bromeliales and Commelinales from the Liliales. Two families, Juncaceae (rush) and Thurniaceae.

Order Cyperales. Evidently derived from the most primitive Juncaceae. One family, Cyperaceae (sedge).

Order Bromeliales. Probably derived from the Liliales (a common origin with the Juncales and Commelinales). One family, Bromeliaceae (pineapple); see BROMELIALES.

Order Commelinales. Probably shares a common origin with the Bromeliales. Four families, including Commelinaceae (spiderwort), Xyridaceae (xyris), and Rapateaceae.

Order Eriocaulales. One family, Eriocaulaceae.

Order Restionales. Probably shares a common origin with the Commelinales. Four families, including Restionaceae (restio), Flagellariaceae (flagellaria).

Order Poales. Arose from the Restionales, most likely from some early Flagellariaceae. One family, Poaceae or Gramineae (grass); see POALES.

Subclass Arecidae

Herbaceous or arborescent plants mostly with broad, petiolate leaves. Vessels in all vegetative organs or restricted to root (Arales). Stomata with 2, 4 (often) or 6 (rarely) subsidiary cells. Flowers mostly unisexual, usually small and numerous, aggregated into inflorescences generally subtended by more or less prominent spathaceous bracts. Perianth of 6 segments, reduced to bristles or scales, or absent. Gynoecium usually coenocarpous (syncarpous or paracarpous), rarely apocarpous (in some palms). Pollen grains 2-celled, rarely 3-celled, mostly monocolpate. Large ovules usually with 2 integuments. Seeds almost always with endosperm. Five orders.

Order Arecales. Probably derived from the apocarpous ancestors of the Liliales. One family, Arecaceae or Palmae (palm); see ARECALES.

Order Cyclanthales. One family, Cyclanthaceae.

Order Arales. Most likely had a common origin with the Arecales. Two families, Araceae (arum), Lemnaceae (duckweed).

Order Pandanales. Most likely had a common origin with the Arecales and Arales. One family, Pandanaceae (screw pine).

Order Typhales. Probably had a common origin with the Pandanales, but the affinity is not very clear. Two families, Sparganiaceae (bur reed) and Typhaceae (cattail).

Classification systems of the 19th century

Critical appraisal. Since 1900, and until recently, the dominant system of classification was that proposed by the German botanist Adolph Engler and his associates. It is still widely accepted as the basis for arrangement of families in the majority of existing manuals and floras, and for the arrangement of specimens in many larger herbaria, though in British herbaria plants are arranged according to the system devised by George Bentham and Sir Joseph Dalton Hooker and published in three volumes of their *Genera plantarum* (1862–83). The Engler system won wide acceptance because of its thoroughness: it was applied by Engler and his associates to the whole plant world, at all taxonomic levels. It was based largely upon the system proposed in 1883 by August Wilhelm Eichler, with certain features of the systems of the French botanist Adolphe-Théodore Brongniart (1843) and the German botanist Alexander Braun (1864). Engler considered the monocotyledons to be more primitive than the dicotyledons, and the apetalous and catkin-bearing dicotyledons (willows, walnuts, birches, oaks, etc.) as more primitive than the others. The principal fault in the Engler system is that it equates simplicity with primitiveness, completely ignoring the significance of secondary reduction and simplification.

Braun later (1875) put forward some fundamental ideas that anticipated by several decades the basic tenets of phylogenetic angiosperm systematics. According to him, the primitiveness of the Magnoliaceae, Alismaceae, and

similar families is apparent when the flower is considered as a strobilus substantially homologous with the cones of the cycads, and from this it follows that the petal-less forms (Apetalae), characterized by small, simply constructed unisexual flowers, must be derivative, and that the simplicity of their flowers is not primitive. To Braun belongs the well-known aphorism: "In nature, as in art, the simplest may be the most perfected." Braun, however, died only two years later, without having undertaken the reform of the system in accordance with the new principles he had enunciated. The German botanist Karl Wilhelm von Nägeli, in 1884, and the French palaeobotanist Gaston de Saporta, in 1885, came to hold similar views independently. But the credit for reforming angiosperm classification on an evolutionary basis belongs to the American botanist Charles E. Bessey and the German botanist Hans Hallier. Their first papers appeared in 1893 (Bessey) and 1903 (Hallier), but the most complete account of Hallier's system appeared in 1912, and of Bessey's in 1915. In 1916 the Russian botanist Christopher Gobi published the system that he had already used in the late 1890s in lectures to his students at St. Petersburg University. These three workers came to the conclusion that the "Apetalae" were not only derivative and simplified, but also heterogenous, the different members thereof having arisen along different evolutionary lines within the dicotyledons. Later, the British botanist John Hutchinson proposed a new system of classification in his *Families of Flowering Plants* (1926–34; 2nd edition, 1959) based on many of the principles adopted earlier by Bessey and Hallier. One of the main differences is the division of the dicotyledons into two main groups, Lignosae and Herbaceae, the former embracing those families "fundamentally woody," and the later those that are "fundamentally herbaceous." This division leads to the wide separation of some obviously closely related families, such as the Araliaceae and Apiaceae or the Verbenaceae and Lamiaceae. The Hutchinson system rests mainly upon gross morphological data. In spite of deficiencies it has noticeably stimulated the development of the phylogenetic systematics of angiosperms.

The Takhtajan and Cronquist systems

The most recent systems of angiosperm classification are those of the Soviet Armenian botanist Armen Takhtajan and of the American botanist Arthur Cronquist. Takhtajan's system has gone through a series of modifications since its first appearance in 1942, its last version having been published in 1969 and 1970 in Russian and in English. Cronquist's system was published in part in 1957 (the dicotyledons), and appeared in its latest, complete version in 1968. Both systems are synthetic in the sense that their authors have synthesized data from different fields of botany into one taxonomic whole. Both authors recognize, however, that knowledge of very many angiosperm groups is scanty, insufficient, and in many cases conflicting, and that many details of classification are still uncertain. These two systems have much in common. They evolved independently in parallel fashion from a similar ancestry (that of Cronquist more nearly from Bessey, that of Tahktajan more nearly from Hallier), and have influenced each other, leading toward an increasing convergence of the two systems. The most recent published versions of these two systems agree in treating the angiosperms as a division Magnoliophyta, with two classes, the Magnoliopsida (dicotyledons) and Liliopsida (monocotyledons). These new names are the result of the extension of the type concept in nomenclature up to the rank of division. They also agree in the subdivision of the Magnoliopsida and Liliopsida in subclasses, and in many other points. There have been also proposed two other modern systems of classification, by the Hungarian botanist R. Soó (1967) and by the American botanist Robert F. Thorne (1968); these systems have much in common with those of Cronquist and Takhtajan.

BIBLIOGRAPHY. D.I. AXELROD, "The Evolution of Flowering Plants," in *Evolution after Darwin* ed. by S. TAX, 1:227–305 (1960), a palaeobotanical approach to the origin and evolution of the angiosperms; A. CRONQUIST, *The Evolution and Classification of Flowering Plants* (1968), an excellent book on the evolutionary classification of the angiosperms containing useful keys and many references; P.H. DAVIS and V.H. HEYWOOD, *Principles of Angiosperm Taxonomy* (1963), a most useful and comprehensive single volume, but now a little dated; A.J. EAMES, *Morphology of the Angiosperms* (1961), a high-level classic book containing many useful references; A. ENGLER, *Syllabus der Pflanzenfamilien*, 12th ed. rev. by H. MELCHIOR, vol. 2 (1964), a valuable guide, in German, to the current version of the Engler system of classification, including family descriptions, useful keys, and many references; K. ESAU, *Plant Anatomy*, 2nd ed. (1965), probably the best textbook of plant anatomy, containing a comprehensive bibliography; V.H. HEYWOOD (ed.), *Modern Methods in Plant Taxonomy* (1968), a comprehensive review of the application of newer techniques by well-known experts in the field, with many references; J. HUTCHINSON, *The Families of Flowering Plants*, 2nd. ed., 2 vol. (1959), contains useful keys, descriptions, and illustrations of the families recognized by the author; C. JEFFREY, *An Introduction to Plant Taxonomy* (1968), an excellent, clearly written semipopular introduction, with a list of additional readings and reference works; G.H.M. LAWRENCE, *Taxonomy of Vascular Plants* (1951), a well-illustrated textbook of college level difficulty, containing excellent descriptions of the main angiosperm families, and further references, but now somewhat dated; A.C. SMITH, "The Pacific as a Key to Flowering Plant History," University of Hawaii Harold L. Lyon Arboretum Lecture No. 1 (1970), geographical aspects of angiosperm origin; A.L. TAKHTADZHIAN (TAKHTAJAN), *Flowering Plants: Origin and Dispersal* (1969; Eng. trans. from the 2nd Russian ed., 1961) a semipopular treatment of the angiosperms, with an extensive bibliography (including many important Russian works), and an outline of the Takhtajan system of classification; J.C. WILLIS, *A Dictionary of the Flowering Plants and Ferns*, 7th ed. rev. by H.K. AIRY-SHAW (1966), an extremely useful reference book.

(A.T.)

Angkor

For more than half a millennium the city of Angkor, which is located in northwestern Cambodia just to the north of the modern town of Siem Reap, served as the royal centre from which a dynasty of Khmer (Cambodian) kings ruled one of the largest, most prosperous, and most sophisticated kingdoms in the history of Southeast Asia. From the last decade of the 9th century AD, when King Yaśovarman I moved his capital to Angkor, until the early years of the 13th century (excluding one short period when the capital was moved to another site, and occasional crises created by foreign invasions), the kings of Angkor ruled over a territory that extended from the tip of the Indochinese peninsula northward to Yunnan and from Vietnam westward to the Bay of Bengal. During this entire period these rulers utilized the vast resources of labour and wealth at their disposal to carry through a series of prodigious construction projects designed to glorify both themselves and their capital city. After the reign of King Jayavarman VII (1181–*c.* 1215), the power and vitality of the kingdom gradually waned until finally, after Thai armies captured and sacked Angkor in 1431, the city was abandoned.

During the period of great construction and building at Angkor, which lasted more than 300 years, many changes in architecture and artistic style can be discerned, and there was a religious movement from the Hindu cult of the God Śiva (Shiva) to that of Viṣṇu (Vishnu) to a Mahāyāna Buddhist cult devoted to the *bodhisattva* Avalokiteśvara.

Symbolic significance

Like many other premodern cities in Southeast Asia, Angkor was pre-eminently a centre for administration and for the worship of a divine monarch; it was planned, constructed, and often reconstructed on the basis of religious and political conceptions imported from India and adapted to local traditions. From the time of Yaśovarman I, who gave the city its original name of Yaśodharapura, Angkor was conceived as a symbolic universe structured according to the model provided by the traditional Indian cosmology. Thus, the city was oriented around a central mountain or pyramid temple, which was identified with Mt. Meru, the central mountain of the Indian cosmic system, and was, at the same time, believed to concentrate the potency of the soil upon which the prosperity of

the kingdom depended. In the case of the original Yaśodharapura, the central mountain temple was an architectural adaptation and completion of the one natural hill in the area, the Phnom Bakheng; in the later history of the city the central temples were completely architectural creations (*i.e.*, pyramid temples), such as the Phimeanakas of Jayavarman V (reigned 968–1001); the Baphuon of Udayadityavarman II (reigned 1050–66); and the Buddhist temple of Bayon, which was the central temple built by Jayavarman VII when he gave the city, which by then had become known as Angkor Thom, its more or less final form. Moreover, the vast system of reservoirs, canals, and moats that constituted one of the primary characteristics of Angkor served not only as a means of water control and irrigation but also as symbols of the great ocean that, according to the Indian conception, surrounded the central cosmic mountain.

The relationships between the cosmological concepitons and the city itself had their counterparts in the relationships between its leading inhabitants and the gods. In the central mountain, or pyramid, temples a cult of the *deva-rāja*, or god-king, was maintained, through which the king became identified with one of the great sovereign deities, an identification that was finalized when, at the king's death, the central temple became his personal funerary temple or mausoleum. Many of the other great temples at Angkor, all of which gave expression to Indian cosmological and mythic themes, were built in order to provide a locus for cults through which the kings who did not construct a new central temple, members of the royal family who were not actual rulers, and, in some cases, certain members of the aristocracy could be assured of immortality through becoming identified with Śiva or one of the other pre-eminent gods of the realm.

Angkor Wat

For example, Angkor Wat, which is perhaps the greatest and certainly the most famous of all the temples in the Angkor complex, was built by King Suryavarman II (reigned 1113–*c.* 1150) as a vast architectural microcosm within which his remains were to be deposited and his permanent identity with Viṣṇu symbolically and cultically confirmed.

Guy Nafilyan



Angkor Wat, seen from the west side.

In the late 13th century, according to the vivid account of the Chinese commercial envoy Chou Ta-kuan, Angkor was still a large and thriving metropolis and one of the most magnificent capitals in all of Asia. Nevertheless, by this time the great building frenzy that had culminated during the reign of Jayavarman VII had clearly come to an end, the new and more restrained religious orientation represented by Theravāda Buddhism was on the rise, and the armies of the Thai kingdoms established in the western sections of the empire were already beginning to encroach on the Khmer heartland. By the 16th century, when the next available firsthand description was written, these trends had long since culminated in the abandonment of the city; and all that remained were the jungle-covered remnants of the ancient temples and the ruins of the once-magnificent system of reservoirs and waterways.

During the more than four centuries between the demise of the ancient city and the beginning of the modern period (*i.e.*, from the early 15th century to the late 19th century), interest in Angkor was largely focused on Angkor Wat, which, having been taken over and kept largely intact by Theravāda monks, became one of the most important pilgrimage sites in Southeast Asia. Even during this period, however, a number of early European visitors to Cambodia evidenced a strong curiosity concerning the "lost city," and when the French colonial regime was established (1863) the entire site became the focus of an intense scholarly interest and concern. Working at first independently and then under the aegis of the government-sponsored École Française d'Extrême-Orient, a talented and dedicated group of French archaeologists and philologists initiated a comprehensive program of research, which has gradually yielded the knowledge now possessed about the history of the city and the fascinating religious and political system that informed and guided its life. Archaeologists have also carried through an arduous and painstaking program of reconstruction, through which the ancient complex of temples, reservoirs, and canals has been restored to something of its original grandeur. As a result of these efforts and of the continuing support of the project by the postcolonial Cambodian government, the monuments of Angkor stand today as a striking testimony not only to the awesome power and wealth of the ancient Khmer kings but also to the technical proficiency and artistic genius of their subjects.

BIBLIOGRAPHY. The classic work on Angkor, including both the city and the "Angkor region," remains GEORGE COEDES, *Pour mieux comprendre Angkor*, 2nd rev. ed. (1947), which is available in slightly abbreviated form under the title *Angkor: An Introduction*, ed. and trans. by EMILY F. GARDINER (1963). For an excellent visual presentation, see BERNARD GROSLIER and JACQUES ARTHAUD, *Angkor: hommes et pierres*, rev. ed. (1968; Eng. trans. by ERIC E. SHAW, *Ankor: Art and Civilization*, rev. ed., 1966). For a colourful 13th-century description, see PAUL PELLIOT's translation of the famous memoirs of Chou Ta-kuan, published in the *Bulletin de l'École Française d'Extrême-Orient devarāja*, 2:123–177 (1902). An English rendering of Pelliot's translation has been published in Bangkok by J. GILMAN D'ARCY PAUL, entitled *Notes on the Customs of Cambodia* (1967).

(F.E.R.)

Anglican Communion

The Anglican Communion, descendant of the Church of England, one of the three major religious bodies that separated from the jurisdictional and doctrinal control of the papacy in the 16th-century Reformation (the others were the Lutheran and the Reformed), is a family of churches with few written rules but many unspoken loyalties and commitments. It is united by a common loyalty to the archbishop of Canterbury in England as its titular leader and by an agreement with the doctrines and practices defined in *The Book of Common Prayer*, a 16th-century liturgical book whose roots lie in the early history of the Christian Church in England.

Nature and significance. *Authority and structure.* Having no central authority and no one person, such as the pope of the Roman Catholic Church, to whom its various parts can unfailingly appeal and from whom it can expect to receive final authority or directives, the Anglican Communion consists of national, independent, and autonomous churches that are bound together by intangible links probably best described as ties of loyalty between the ancient see of Canterbury and in consequence with each other. The archbishop of Canterbury, though respected and revered throughout the whole Communion and whose words carry great moral authority, exercises no jurisdiction over any part of the Communion other than over the diocese of Canterbury, of which he is the chief pastor, and over the Church of England as a whole through the authority vested in synods and convocations. Like a family, the Anglican Communion changes its form and shape, increasing when new provinces (areas of jurisdiction) are formed and decreasing when schemes of union with non-Anglican churches are consummated.

Role of the archbishop of Canterbury

Containing oddities in its structure, which is itself sometimes considered too formal a word for the Anglican Communion, the basic unit is a diocese, of which, in the early 1970s, there were some 350 spread throughout all the continents of the world. For the most part these dioceses form part of a larger unit known as a province, but even these are far from uniform in configuration. A province, for example, may be part of an autonomous church: the Anglican churches in Australia and Canada have four provinces; the churches of England and Ireland have two each; and the Protestant Episcopal Church in the United States of America has nine. Some provinces, however, include whole countries, such as Japan, South Africa, West Africa, Kenya, Uganda, Tanzania; other provinces cover a number of countries, such as the provinces of Central Africa and the West Indies. On occasion, one diocese covers a whole country or even several countries, such as Zambia and Polynesia.

Variations in titles of Anglican ecclesiastical leaders

Variations occur in the titles of the heads of the various provinces or national churches. England has two archbishops (Canterbury and York), known as metropolitans, as does Ireland. Canada has a primate (who has no province) and four metropolitans; Australia has four archbishops, one of whom—while having jurisdiction over a province—is known as the primate. The Church of Japan has a presiding bishop who also has a diocese, and the United States has a presiding bishop who does not. To complicate matters of structural or organizational concerns still further, the Church of Scotland has a primus (primate); the Protestant Episcopal Church in the United States has presidents (who are elected only for three-year periods) of its nine provinces; and there are two bodies that for geographical reasons can hardly be called provinces—the council of the Church of South East Asia and the South Pacific Anglican Council. Further compounding the structural situation, there is a North American regional council that itself does not have provincial status, though its members belong to various provinces. To the despair of those who lean toward strictly defined structural or organizational entities, the Anglican Communion also includes a loose association of dioceses in Latin America, some belonging to the province of Brazil, some to one of the nine provinces of the church in the United States, and some even having extra-provincial relations with the Church of England, which along with a few other dioceses scattered around the world, look to the archbishop of Canterbury as their metropolitan.

In addition to these dioceses, provinces, churches, and councils, there also is the Holy Catholic Church of China (Chung Hua Sheng Kung Hui), of which little is known but which is believed to have a chairman of the house of bishops and 15 dioceses. The only diocese of the Holy Catholic Church of China of which accurate knowledge is available is that of Hong Kong and Macau, represented in the council of the Church of South East Asia.

The total worldwide membership of the Anglican Communion is approximately 45,000,000.

Comprehensiveness in doctrine and practice. Often said by many of its ministry to be the middle way between Roman Catholic and Protestant churches, the Anglican Communion is comprehensive in matters of doctrine and practice. While asserting the importance of the apostolic succession of bishops and *The Book of Common Prayer*, it nevertheless allows a considerable degree of flexibility in most doctrinal and liturgical matters. Thus, within the communion there are several schools of thought and practice, including High Church, Anglo-Catholic, Low Church, evangelical, and others. The various churches of the Anglican Communion, though autonomous, are bound together by a common heritage and common doctrinal and liturgical concerns, and there has always been a considerable amount of interchange of ecclesiastical personnel.

History. *Developments in England.* The Church of England, the mother church of the Anglican Communion, has had a long history. When the Christian Church in England began is not known for certain, but it probably was not later than the early 3rd century, according to evidence given by Tertullian and Origen, two theologians from northern Africa who flourished in the first half of the 3rd century. The church was well enough established by the 4th century to send three British bishops—of Londinium (London), Eboracum (York), and Colonia Linum (Lincoln)—to the Council of Arles (in modern France) in 314. In the 5th century, after the Romans had withdrawn from and the Anglo-Saxons had invaded Britain, Illtud performed missionary work in Wales and Patrick in Ireland. Though somewhat isolated from continental Christianity in the 5th and 6th centuries, Christianity remained active in the British Isles, especially in the area of monasticism. About 563 Columba founded a monastic missionary colony on the island of Iona off Scotland. In 597 the Roman monk Augustine, at the request of Pope Gregory the Great, went to England—landing at Kent—and soon established the archbishopric of Canterbury. By 634 Aidan was carrying on missionary work in northern England. As a result, the early history of the Catholic Church in England was represented by three strains—Romano-British, Celtic (especially in Wales and Ireland), and Roman—that by the time of the 16th century evolved into what could properly be called the Church of England. During the Reformation, under the initial leadership of King Henry VIII, the Church of England broke with the pope but, it has insisted, not with the Catholic faith.

Missionary efforts of the English Church

Three centuries passed before the phrase the Anglican Communion could be used with meaning. From the time of the Reformation to the 19th century, the Church of England had been involved in various internal disputes and developments. After the death of Henry VIII, in 1547, reform and reversion to Roman Catholicism took place; Edward VI, his son, encouraged reform through the advice of his supporters, and—after Edward died six years later—Mary, Edward's half sister, restored relations with the papacy. Upon Mary's death in 1558, Elizabeth I restored independence on the Church of England from Rome in two acts of 1550 later known as the Elizabethan settlement. Though the Elizabethan settlement satisfied those who desired a "reformed Catholicism"—*i.e.*, a church that was reformed in worship and customs but Catholic in doctrines and practice—two other groups in the Church of England did not accept the settlement. Some wanted a return to Rome; others wanted greater reform. The 17th century was the period of influence of the latter, the Puritans; the 18th century witnessed a modification of the reform minded in the Evangelical movement (led by John Wesley); and the 19th century was a time of various reforms: an Anglo-Catholic movement (led by John Henry Newman, who later became a Roman Catholic cardinal) and reforms in both church and state relationships, especially in the areas of education and social concerns.

Developments in worldwide Anglicanism. The Church of England, from the time of the Reformation, began to spread throughout the then-known world; following explorers, traders, colonists, missionaries, and ministers, it served both the natives of the areas that were colonized and the expatriates. All of them considered themselves to belong to the Church of England and looked to the ancient diocese of Canterbury as their source. Most significant in this enterprise were the great missionary societies, such as the Society for the Promotion of Christian Knowledge, founded in 1699, the Society for the Propagation of the Gospel, founded in 1701, and the Church Missionary Society in 1799. Though not officially sponsored by the established Church of England, missionaries sponsored by such societies began work among the indigenous people of English colonies and thus were probably the unconscious founders of the Anglican Communion as it is known in the 20th century.

The Anglican Church in the United States, because of the American Revolution, completed its own separate organization by 1789. Political circumstances had made it necessary, however, for the former colonies to secure their first bishop (Samuel Seabury, in 1784) from Scottish sources. The Anglican Church of Canada had its own separate organization by 1893, though it was known as the Church of England in Canada until 1959, just as the

Anglican Australian Church continues to be so designated. Thus, the Anglican Communion grew in place after place in various parts of the world over the last four centuries, so that there are in the 20th century 23 autonomous churches and two regional councils.

First meeting of the Lambeth Conference

The various churches and councils of the Anglican Communion probably began to be conscious of themselves as a single family of churches with the first meeting of the Lambeth Conference held in Lambeth (the residence of the archbishop of Canterbury in London, England) in 1867. The immediate cause of the first conference was a controversy that arose in one of the colonial churches. The archbishop of Cape Town (South Africa), Robert Gray (who was High Church, or traditionalist), wanted the bishop of Natal, John Colenso (who was Low Church, or evangelical), to be arraigned on charges of heresy for holding what were then regarded as advanced views of the creation stories in the opening chapters of Genesis. The controversy centring on Bishop Colenso aroused intense feelings and anxieties over a wide range of issues—doctrinal, personal, and organizational—among all the Anglican churches throughout the world. Bishop Colenso was convicted and deposed in the church courts but upon appeal to the civil courts of England won his case and retained his church properties. What started out as a personal and jurisdictional dispute in South Africa soon was a matter of concern for all the colonial churches of England. The law of the land and the constitution of a church were revealed to be in opposition to each other. Bishop Gray held that the deposition and excommunication that he as metropolitan had pronounced against one of his suffragans (the Bishop of Natal) could not be appealed. Challenging this claim, some of Gray's own clergy protested to A.C. Tait, bishop of London and the future archbishop of Canterbury:

> If the claim put forward by the Bishop of Cape Town to have his decision as metropolitan regarded as final, be allowed, if, in other words, as he affirms, there is no appeal to any court on earth from a judgement which he may pronounce as metropolitan, it is evident that the Suffragan Bishops of the provinces are in a far worse position than the humblest priest in pre-Reformation times; he at least had an appeal to the Roman pontiff, whilst they are subject without appeal to the sentence, however arbitrary, of the metropolitan.

This says as much about Anglicanism's concern for responsible freedom as it does about the emergence of the Anglican Communion's consciousness of itself.

The archbishop of Canterbury, Charles Thomas Longley, did not follow such a course but instead responded to a request from the United Church of England and Ireland in Canada, meeting in Montreal in September 1865, for a meeting of Anglican bishops from all over the world for purposes of mutual discussion and consultation. Archbishop Longley made it clear that the resulting Lambeth Conference (first held in 1867) was not and still is not a synod or a general council but rather a purely informal gathering of bishops meeting at the invitation of the archbishop of Canterbury, with no power of itself to constrain attendance or to make binding decisions. Despite the apparent limitations, the Lambeth Conference commands considerable moral authority and has from time to time made significant pronouncements, such as an "Appeal to All Christian People" at the 1920 Lambeth Conference.

The Lambeth Conference has met at ten-year intervals, save for the exigencies of war, from 1867, the most recent and tenth being held in 1968 (see also REFORMATION; PURITANISM; PROTESTANTISM, HISTORY OF; BRITAIN AND IRELAND, HISTORY OF).

Teachings and practices: the essence of Anglicanism. *Doctrinal views.* The Lambeth Conference of 1930 described the Anglican Communion as:

> a fellowship within the One Holy Catholic and Apostolic Church, of those duly constituted Dioceses, Provinces or Regional Churches in communion with the See of Canterbury, which uphold and propagate the . . . faith and order as they are generally set forth in the Book of Common Prayer . . .; promote within each of their territories a national expression of Christian faith, life and worship; and are bound together not by a central legislative and executive authority, but by mutual loyalty sustained through the common counsel of the Bishops in conference.

The Anglican Communion thus holds to the Catholic faith as expounded by the Holy Scriptures and by the early Church Fathers. It respects the authority of the state but does not submit to it; and it equally respects the freedom of the individual. In its relationship to the world, the Anglican Communion does not seek to evade the challenges of the world nor live a life separate from it. It views itself not so much as having a mission but as being one. Basing its doctrines on the Bible, the Anglican Communion allows a remarkable latitude of interpretation. Though the Church of England holds close to the spirit of the Thirty-nine Articles (a 16th-century doctrinal document that allows for broad interpretations), subscription to them is not required of laymen and only in a general way of clergy. Other churches or councils of the Anglican Communion take different views of the Articles, but none regard them as having, for example, the status of the historic statements of belief as set forth in the Apostles' or Nicene creed, nor do they accord them the status given to other 16th-century doctrinal statements, such as the Augsburg Confession of the Lutheran churches or the Westminster Confession of the Reformed and Presbyterian churches.

Bishops, priests, and deacons

The ministry. The Anglican Communion holds to a threefold order of ministry: bishops, priests, and deacons. Though holding to the view of succession from the Apostles, it is not committed to any particular theory as to the conveyance of that ministry. Over the centuries, the theological pendulum has swung frequently; sometimes the more sacerdotal (priestly) point of view has been in the ascendancy and at other times the more presbyteral, in which authority is shared by both priests and laymen.

Anglican worship. Anglican public worship has always been liturgical and formal in style—*i.e.*, based on prayer books—but there has always been flexibility and very little uniformity. Even in the Church of England it was not until 1549 that a prayer book for the whole church was made mandatory—and that by act of Parliament. The Prayer Book of 1662, which was the culmination of a number of revisions, is still in general use, though the Church of England in the latter part of the 20th century has been given authority from Parliament to devise revisions of different parts of the Prayer Book.

Some of the overseas Anglican churches still use the Prayer Book of 1662, but many now have their revisions, all holding to the essential elements of the Prayer Book but reflecting local conditions and cultures. The sacraments of Baptism and Holy Communion are held to be most important in the Anglican Communion. Various experimental rites in different parts of the Anglican Communion have been very common in the latter part of the 20th century. Forms of public worship are characteristically changing in the Anglican Communion, and there is a constant tendency toward freedom of expression, though at the same time there is a marked interest in holding to the essential elements of the liturgy.

Organization. *Internal developments.* The mother church of the Anglican Communion, the Church of England, has maintained close connections with the state; it has representative bishops in the House of Lords and can properly be called the established church, even though, contrary to much popular opinion, it is in no sense supported financially by the state. The Church of England itself is without question the church of the English people, even though many of the country's citizens do not so regard it. Only in England do Anglicans comprise a majority, accounting for about 65 percent of the world Anglican population.

Apart from its assured position in the life of England, the Anglican Communion has never had much of a structure on a worldwide basis. Apart from missionaries on leave, Anglican individuals moving about the world, and similar personal relationships, the Anglican Communion has been characterized by its lack of structured cohesion. Even meetings of Anglican Church leaders have been restricted, except in very recent times, to the

meetings of the Lambeth conferences that are held only once every ten years and to three congresses that involved not only bishops but also clergy and lay people. Only three such congresses have been held: London in 1908; Minneapolis, Minnesota, in 1954; and Toronto in 1963. The fact that only three such meetings have been held in the 20th century indicates that the Anglican Communion is not given to frequent large meetings. At two-year intervals between Lambeth conferences, however, meetings of what is called the Lambeth Consultative Body are held. Like other such bodies, the Lambeth Consultative Body has no real authority. It has tended to give some cohesion to the work of the Anglican Communion as a whole and, for the most part, has concerned itself with reflecting on the results of previous Lambeth conferences. Members of the Lambeth Consultative Body were the primates or presiding bishops of the various autonomous national churches with additional representation from the ancient churches, such as England and Ireland, and from the larger ones, such as the Anglican churches in Australia, Canada, and the U.S.

The Lambeth Consultative Body

In 1948 the initiation of a significant change in the Lambeth Conference took place. Because of the outbreak of World War II, no Lambeth Conference had been held since 1930—a period of 18 years. Thus, because only a few of the bishops in 1948 knew their fellow bishops as a result of a long intermediate period, it became apparent that if the various Anglican churches were to work together, some kind of continuing flexible structure was necessary. The Advisory Council on Missionary Strategy came into being, of which the key word was advisory, which is almost synonymous with the word "consultative" in the title Lambeth Consultative Body. The use of the word advisory further emphasized an essential aspect of Anglicanism—*i.e.*, that it has no central authority, no powers binding on all Anglican churches, and nothing approaching a curia (an ecclesiastical advisory and administrative organ, as in the Roman Catholic Church). Though this emphasis has sometimes been frustrating to many Anglican churchmen, it is of the essence of the Anglican notion of freedom and as such has been jealously guarded.

The Advisory Council on Missionary Strategy met only at the call of the archbishop of Canterbury, and its membership (varying from 40 to 45) consisted of the members of the Lambeth Consultative Body and archbishops, or their representatives, who are metropolitans in their own right but are not heads of national churches. Though not infringing on the rights of individual churches, the Advisory Council recognized the need for consultation in the prosecution of the missionary activities of the Anglican Communion.

The Anglican Communion, a self-admittedly slow-moving body, recognized after a period of ten years that someone was needed to develop and suggest methods of common strategy, even if it could only be given by way of suggestion and inference and certainly not by authority. Thus, at the Lambeth Conference of 1958 an Anglican executive officer (Stephen F. Bayne, Jr., who resigned as bishop of Olympia, Washington) was appointed. The post of Anglican executive officer, as it was so envisioned, was one that had no authority and was responsible both to the archbishop of Canterbury and to all the heads of the other churches that make up the Anglican Communion (in 1958 there were 19 such churches). The position of the Anglican executive officer has been likened to that of the secretary of the United Nations, having no authority but acquiring in the process considerable power because of a knowledge of developing situations and worldwide strategy. After five years as the Anglican executive officer, Bishop Bayne believed that a change was necessary, and consequently upon his resignation Ralph S. Dean, bishop of Cariboo in western Canada, was appointed to the position in 1964. Appointed for a term of five years—unlike Bishop Bayne's term, which had no terminal stipulations—and retaining his diocese on a leave-of-absence basis, Bishop Dean brought to the position the views of a missionary bishop. Significant changes took place. At previous Anglican bodies, such as the Lambeth Conference and the Lambeth Consultative Body, persons, though worthy, were in attendance because of privilege. The Lambeth Conference of 1968, in a change of direction, recommended the formation of the Anglican Consultative Council, which would represent all the areas of the Anglican world, all of which would appoint or elect their own representatives, who would include not only archbishops and bishops but clergy and lay people (including women). This new body was to have a total membership of a little over 50 and was to meet every other year, with the addition of a standing committee of nine that would meet annually.

The Anglican executive officer

The first meeting of the Anglican Consultative Council met in Limuru, Kenya, in March 1971, and the standing committee in London in the same year. This body, like its predecessors, is still consultative and has the archbishop of Canterbury as its president. Though the archbishop is not necessarily its chairman, he always presides at the opening session. The council has clear terms of reference that are set out in the 1968 *Lambeth Conference Report*. Considered an important step forward in Anglican cooperation, the Anglican Consultative Council has no more authority than that of previous Lambeth conferences, possibly less, but it reflects the trends of Anglicanism in the latter part of the 20th century. In 1969 J.W.A. Howe, formerly bishop of St. Andrews, Dunblane, and Dunkeld, succeeded Bishop Dean as executive officer and was appointed secretary general, thus becoming responsible to the council itself instead of to a vast number of Anglican prelates.

Relations with other churches. Because the Anglican Communion consists of a cluster of related churches, it does not, as a worldwide Communion, have membership in the World Council of Churches; each of the Anglican churches, however, does hold such membership. This type of ecumenical relationship is in keeping with one of the stated objects of the Anglican Consultative Council:

> To encourage and guide Anglican participation in the Ecumenical Movement and the ecumenical organizations; to cooperate with the World Council of Churches and the world confessional bodies on behalf of the Anglican Communion; and to make arrangements for the conduct of pan-Anglican conversations with the Roman Catholic Church, the Orthodox Churches, and other Churches. (*Lambeth Conference 1968: Resolutions and Reports;* SPCK and Seabury Press, p. 47.)

Ecumenical processes involving the Anglican Communion have been occurring on both international and national levels. After a visit in March 1966 of the Archbishop of Canterbury with Pope Paul VI in Rome, the Anglican/Roman Catholic International Theological Commission was set up to explore mutual doctrinal concerns. Similar conversations have been taking place with other churches (*e.g.*, Lutheran and Reformed) at a variety of levels all over the world. In such contexts, the Anglican Communion, wrestling in the latter half of the 20th century with its own liturgical forms and reforms, has entered into conversations with less liturgically oriented churches. For many Anglicans, such conversations are viewed as being the truest path to Christian unity—providing a forum for debate as to whether organic union, federalism, or grassroots cooperation is the most fruitful way to advance the cause of the mission of the church (see also ECUMENISM).

The quest for Christian unity

World Anglicanism. In the latter part of the 20th century, 80 percent of the membership of the Anglican Communion was white. There has been, however, obvious growth in Africa and Asia, most notably in Africa. Though at the first Lambeth Conference of 1867 there were no non-white bishops, in the early 1970s there were at least 100. In some areas Anglicanism has grown as a result of the increasing expansion of the gospel; in others it enters into fuller life in united churches, such as in South India, North India, Pakistan, and Sri Lanka (formerly Ceylon); and in still others the Anglican Communion has special interchurch relations, such as with the Lusitanian Church in Portugal, the Mar Thoma Syrian Church in India, the Old Catholic churches in Europe and the United States, the Philippine Independent

Church, and the Spanish Reformed Episcopal Church. Some Anglican provinces have varying degrees of intercommunion with Lutheran churches in different parts of Europe (*e.g.*, the Church of Sweden).

Conclusion. Though deeply rooted in tradition, the Anglican Communion is not static. It has been flexible, susceptible to an ever-changing world, sometimes conservative, and sometimes radical.

The expression of the mood of many members of the Anglican Communion in the latter part of the 20th century was reflected in the concluding paragraphs of a document entitled "Mutual Responsibility and Interdependence in the Body of Christ," which was endorsed by the 1963 Toronto Anglican congress:

> The Church is not a club or an association of like-minded and congenial people. Nor is our Communion, named for its historic roots, a federation commissioned to propagate an English-speaking culture across the world. . . . The Church exists to witness, to obey and to serve. . . . In substance, what we are really asking is the rebirth of the Anglican Communion, which means the death of many old things but—infinitely more—the birth of entirely new relationships. We regard this as the essential task before the churches of the Anglican Communion now.

The Anglican Communion has tried to establish itself as the middle way in Christianity, attempting to bridge the gulfs between Protestant, Roman Catholic, and Eastern Orthodox churches. In 1947 Anglican dioceses were included in the new Church of South India, a communion that also included mission churches of the Methodists and Congregationalists. In 1960 in the U.S. the Consultation on Church Union was initiated, with Anglicans taking an active part, the intention being to unite about 30,000,000 members of 10 denominations, but in the early 1970s one of the participating churches withdrew. By such actions the Anglican Communion tried to demonstrate its ultimate goal, a reunited church.

BIBLIOGRAPHY. STEPHEN C. NEILL, *Anglicanism*, 3rd ed. (1965), a comprehensive survey of the Anglican Church from its beginnings to the formation of the national indigenous and interdependent churches that now make up the Anglican Communion; THE LAMBETH CONFERENCE 1968, *Resolutions and Reports* (1968), useful for the study of the Anglican Communion in matters of faith, ministry, and unity and includes 69 resolutions covering a variety of other concerns; STEPHEN F. BAYNE, JR. (ed.), *Mutual Responsibility and Interdependence in the Body of Christ, with Related Background Documents* (1963), the report of the Advisory Council on Missionary Strategy that is especially useful for information on Anglican policies in education and special ministries, such as ministries to merchant seamen, travellers, and the role of Anglican laymen abroad; JOHN W.C. WAND, *Anglicanism in History and Today* (1961), discusses the organization, doctrines, and practices of the churches of the Anglican Communion.

(R.S.De.)

Angola

The largest state of Portugal, Angola is situated on the southwest coast of Africa. With an area of 481,351 square miles (1,246,700 square kilometres), it is 14 times the size of metropolitan Portugal and, in 1970, had a population of almost 5,700,000. The former overseas province, which Portugal declared in 1972 to be an autonomous state, is bounded on the north and northeast by Zaire, on the southeast by Zambia, and on the south by South West Africa. To the west it fronts on the Atlantic Ocean. The small enclave Cabinda, with an area of 2,807 square miles (7,270 square kilometres), lies 20 miles to the north; for administrative purposes it is considered a district of Angola. Cabinda is bounded by the Congo (Brazzaville) on the north, Zaire on the east and south, and the Atlantic Ocean on the west. The coastline of Angola proper stretches for 1,025 miles, from the mouth of the Congo River in the north to the mouth of the Rio Cunene in the south. The capital is the port city of Luanda.

Angola's rich resources

Originally an agricultural country, Angola has undergone considerable industrial development in recent years. Rich oil strikes, the exploitation of large iron-ore deposits, and diamond mining have placed the state economically in the forefront of Portugal's overseas territories. One of the world's leading producers of coffee, Angola enjoys a favourable balance of trade.

Angola is the largest surviving unit of the white-dominated colonial empires of the past, even though it now has the status of a Portuguese state and no longer that of an overseas province or colony. Together with the Portuguese state of Mozambique, Rhodesia, South West Africa, and South Africa, it is part of the cluster of the white-dominated governments of southern Africa. Continuing Portuguese control of the country is contested by the guerrilla forces of the Angolan national movements —most notably of the Popular Movement for the Liberation of Angola (MPLA), whose president, Augustinho Neto, obtained recognition of the movement by the Organization of African Unity in 1964. In the early 1970s it was claimed that MPLA forces were operating on several fronts and had established schools and medical services in the areas they controlled; these claims were, however, denied by the Portuguese authorities. It was not difficult for the guerrilla forces to win support in most areas—a history of poor social conditions and of police repression had combined to produce a desire for change. (For associated physical features, see CONGO RIVER; ZAMBEZI RIVER; for historical aspects, see SOUTHERN AFRICA, HISTORY OF.)

THE LANDSCAPE

Natural features. *Relief.* The shape of Angola is roughly square. The land rises from the low coastal zone to an interior plateau that is tilted to the east, north, and south. The Lunda Divide forms a watershed that separates the north- and south-flowing rivers.

The land may be divided into four zones—the flat coastal zone, and three plateau regions (the *planalto*, the higher tablelands, and the uplands). The principal one, the *planalto*, lies from 3,500 to 4,500 feet above sea level, forms part of the great southwest African plateau, and covers three-fifths of the total area of Angola. The higher tablelands include the Rand Plateau in the north; the Benguela, Bié, Huíla, and Lunda Divide plateaus in the central region; and the Humpata Highlands and Serra da Chela (Chela Mountains) in the south. These areas lie at an altitude of up to 7,500 feet. The uplands slope down to the coastal plain in a series of escarpments joined by level areas. In the east the plateaus dip gradually toward the drainage systems of the Congo and Zambezi rivers. In the south the Moçâmedes Desert stretches inland from the port of Moçâmedes to the Serra da Chela. It is crossed by seasonal rivers that are dry most of the year and is sparsely inhabited by herding and hunting tribes; it is renowned for its game, particularly its zebra.

Soils. The coastal belt of Angola belongs to the partly Tertiary formation (from 2,500,000 to 65,000,000 years old) that follows the coast from the Gulf of Guinea to the Cape of Good Hope. The southern zone generally consists of recent superficial deposits of dunes and alluvia. The west area of the central plateau has a crystalline bedrock, and mineral deposits lie close to the surface. The east, except for the Cassange Depression to the north, is mainly buried under relatively recent sand formations. In the coastal area there are large concentrations of granite, crystalline limestones, and gneiss (a laminated or foliated rock, formed under heat and pressure), as well as stretches of quartzite. The Zambezi and Congo areas display sandstone, schist (a rock with a foliated structure, which splits easily and which, unlike gneiss, contains no feldspar), and quartzite and a wide variety of volcanic rocks. There are diamonds in the gravel and laterite (leached, red, iron-bearing soil) in the Lunda area. The soil is sandy and dry in the coastal zone, merging to desert in the south. Petroliferous structures are found in the deposits, and there are limestone and sandstone cliffs, as well as sandbars and spits. Alluvial soil along the riverbeds is fertile. The river valleys north of the Lunda Divide are rich in tropical forest, the ridges between them being covered with savanna grasslands.

The rivers

Drainage. The principal rivers flow either from the plateau westward to the Atlantic or, in the eastern part of Angola, to the north and south. The Rio Cuanza, which

flows into the sea some 30 miles south of Luanda, drains much of the central area. Other important rivers are the Cuando and the Cubango, which flow generally through the southeast and empty into the Okavango Swamps of Botswana; the Zambezi (*q.v.*) in the east; and the Cunene, which rises in the Bié Plateau and flows south and west for 587 miles inside Angola before reaching the Atlantic Ocean. Ambitious irrigation schemes have made possible the reclamation of land in the river basins, while hydroelectric plants are harnessing the rivers' waters to produce power. The 350-foot high Duque de Bragança Falls on the Rio Lucala and the Catarata Ruacaná (Ruacana Falls) on the Rio Cunene are scenic wonders.

Climate. Angola has a tropical climate with two distinct seasons. In general, the average annual temperatures decrease as one travels away from the Equator, as well as with altitude; temperatures increase as one approaches the sea. To the south, the cool Benguela oceanic current creates atmospheric conditions not conducive to rainfall. The town of Santo António do Zaire, in the north, has an average temperature of 79° F (26° C). In Nova Lisboa, in the central uplands, the average temperature is 67° F (19° C). Seasonal fluctuations are only between 2° and 5° F (1° and 3° C). The rainy season lasts about seven months, from October to mid-May. The greatest amount of rainfall—about 70 inches—occurs annually in the Maiombe Forest located in Cabinda, while the smallest amount—as little as two inches—falls in coastal areas. About 13 inches falls on the northern coast at Luanda, and 11 inches at Lobito. Average wind speed is never less than about three miles an hour and never more than about ten miles an hour.

Rainfall and temperature

Vegetation. Tropical rain forests and dry and wet savanna (grassland with scattered trees) lie inland and to the north of Malanje. The northern forest zone is not only rich in timber but also has proved suitable for the establishment of coffee plantations, as well as for growing cocoa, tobacco, cotton, maize (corn), rice, and sisal. The grasslands produce maize (corn), wheat, sisal, citrus fruits, strawberries, figs, melons, and European flowers. Sugarcane, olives, and grapes are cultivated in the coastal zone. Cereals and sweet potatoes are grown inland, and the high plains are covered with grasslands and shrubs. A unique Angola growth is the *Welwitschia mirabilis*, a primitive, low-growing plant of great botanical interest. Described as a living fossil, it is peculiar to the Moçâmedes Desert.

Animal life. The larger plant-eating animals of Angola include giant black antelope, elephants and hippopotamuses, white and black rhinoceroses, giraffes, zebras, gazelles, and gorillas. Among the carnivores are lions, cheetahs, and leopards. The land is rich in bird species and in a wide variety of reptiles, including crocodiles. Butterflies, poisonous spiders, white ants, and locusts are indigenous. Marine life includes the conger eel, the black whale, crustaceans such as the lobster and giant prawn, and tunny and many European coastal fish such as mackerel, sardine, catfish, and mullet. To conserve the state's flora and fauna, nine national parks and nature reserves have been created, the largest of which are Iôna National Park (about 6,000 square miles) and Milando Reserve (4,600 square miles).

Traditional regions. Originally, the coastal plain contained a relatively sparse African population that lived largely in fishing communities and on small subsistence farms. The first European settlements were made in this area, and the urban centres they created attracted increasing internal migration. The great focal point of human settlement and farming is the savanna plateau in the northeast. The northern tropical rain forest and tropical grasslands are thinly populated because of poor soil.

The landscape under human settlement. *Rural settlement.* There is a sharp division in Angola between the way of life in the traditional African villages and that in the Portuguese settlements. The white settlement at Cela, on the Amboim Plateau, about 140 miles northeast of Lobito, is typical of the latter—the farmers live in houses built in the Portuguese style and do not use African labour of any kind, while the Africans live in household compounds or small villages. In dry areas the villages may be moved every few years, after farming has exhausted the soil. The houses are made of rushes and bamboo and are roofed with palm leaves. They are round or rectangular in shape, according to tribal tradition, and in some areas are surrounded by palisades.

African and Portuguese life-styles

Urban settlement. The cities resemble those of Portugal. Luanda is reputed to be one of the most European cities in Africa, with fine villas on the outskirts and office buildings, blocks of flats, and shops along its modern thoroughfares. The pattern of life in Luanda, as in the other Angolan cities, such as Silva Porto and Nova Lisboa, and in ports, such as Lobito, Benguela, and Moçâmedes, is that of a busy multiracial community, but most of the urban African population lives in shantytowns called *muçeques* on the outskirts of these cities.

PEOPLE AND POPULATION

Population groups. *Ethnic identities.* The population consists of Africans, Europeans born in Africa, European immigrants, and mestizos (persons of mixed African and European descent). At the time of the 1960 census, the total population was about 4,800,000, of whom about 170,000 were white, 53,000 mestizo, and the remainder African. In 1970 the official census figure was about 5,670,000. Most of the Africans are Bantu (*i.e.*, blacks who belong to the Bantu-speaking group). The largest Bantu group is the Ovimbundu (numbering between 1,500,000 and 1,700,000 people), who live in the central highlands; they are farmers, astute traders, and skilled craftsmen. Next in size are the Kimbundu (or Ambundu), numbering over 1,000,000, who speak the Kimbundu language; they consist of three main tribes—the Mbaka, Ndongo, and Mbondo. Other tribal groups include the Kongo (Bakongo; about 500,000), the Ambo (62,000), the Herero (25,000), and the Chokwe-Lunda (360,000). The Hottentot–Bushmen group, which is non-Bantu, number about 7,000. Other small separate groups are the Vatua (Koroca; 6,000) and the Xindonga.

Linguistic groups. Most Africans speak one of the Bantu languages as their first language, and there are more than 90 Bantu-speaking tribes altogether. A small minority consisting of Hottentots, Bushmen, and Cazamas, whose numbers are estimated at 4,000, speaks a Khoisan language, characterized by recurrent clicking noises. Only 5 percent of the people speak a European tongue as their first language, even though Portuguese is the official language and is understood by about one-quarter of the Africans, particularly those living in urban areas; but efforts are being made to extend the use of Portuguese. In the countryside tribal dialects are usually spoken. It is estimated that about 1,200,000 people speak Kimbundu; about 1,700,000 speak Umbundu (the language of the Ovimbundu people); 600,000 Lunda-Quioco (spoken by the Lunda and some other peoples); 500,000 Kikongo (the language of the Kongo [Bakongo]); 110,000 Nhaneka; 85,000 Humbe; 60,000 Herero; and 25,000 are Ambo-speakers.

Religions. The majority of the Europeans and a segment of the Africans are Roman Catholics; altogether about 2,200,000 Angolans adhere to Catholicism. The state is divided into the archdiocese of Luanda and six dioceses. There are almost 200 Roman Catholic missions, and more than 600 resident priests, of whom about 45 percent are African, including a bishop who was ordained in 1970. There are about 700,000 Protestants, including Baptists, Methodists, members of the Council of Evangelical Churches, Plymouth Brethren, and Seventh-day Adventists. Ever since guerrilla warfare broke out in the north in 1961 against Portuguese rule, many Protestant missionaries have been expelled from Angola on suspicion of having participated in activities inimical to the Portuguese government. There are now fewer than 100, as compared with nearly 300 before the hostilities.

Catholics and Protestant groups

About half the population still practice tribal religions that are basically similar but diverge on minor differences of ritual or belief. The sorcerer of the tribe, who is regarded as being in communication with spirits, practices medicine, interprets dreams, and advises on personal

Angola, Area and Population

	area		population	
	sq mi	sq km	1960 census	1970 census
Districts (*distritos*)				
Benguela	14,598	37,808	488,000	475,000
Bié	27,749	71,870	453,000	650,000
Cabinda	2,807	7,270	59,000	81,000
Cuando Cubango	74,162	192,079	113,000	112,000
Cuanza Norte	10,466	27,106	263,000	298,000
Cuanza Sul	22,884	59,269	405,000	459,000
Cunene	*	*	*	*
Huambo	11,841	30,667	597,000	838,000
Huíla	64,227*	166,348*	595,000*	645,000*
Luanda	13,046	33,789	347,000	561,000
Lunda	64,782	167,786	247,000	303,000
Malanje	39,007	101,028	452,000	559,000
Moçâmedes	21,601	55,946	43,000	53,000
Moxico	77,137	199,786	266,000	213,000
Uíge	21,551	55,818	399,000	386,000
Zaire	15,494	40,130	104,000	42,000
Total Angola	481,351†	1,246,700	4,830,000‡	5,673,000‡

*Figures for the recently organized Cunene District are included in Huíla District. †Converted area figures do not add to total given because of rounding. ‡Figures do not add to total given because of rounding.
Source: Official government figures.

problems. Organized religious ceremonies accompanied by appropriate dances take place occasionally, and worthy burial is important lest the soul of the departed return to cause harm to the living.

Demography. As mentioned, the population increased from about 4,800,000 in 1960 to 5,673,000 in 1970. In the late 1960s the birthrate was 50 per 1,000 inhabitants and the death rate was 30 per 1,000. In 1969 approximately 20,000 Portuguese entered Angola to take up or resume residence. On the other hand, Angolan Africans have long been leaving the state to work in other African countries, particularly South West Africa, often as miners. Further, government settlement schemes stimulate both white immigration from Portugal to Angola and African migration to government settlements within the state; both white and African settlers are provided with houses, tools, seed, and cattle. The main white settlement scheme is at Cela, near Gabela, in Cuanza Sul District. The Matala settlement, which is located in the Huíla District, accepts both white immigrants and local Africans. The Caconda settlement, also in Huíla District, is for Africans only. The former African settlements in the Loge and Damba valleys in the Uíge District have been abandoned by their original inhabitants and repopulated by immigrants from Portugal and the Cape Verde Islands.

The settlement schemes

The population of Angola is essentially rural. More than 70 percent live in the central coastal and tropical upland areas of Benguela, Huambo, Bié, Malanje, Luanda, Cuanza Norte, Cuanza Sul, and Uíge, where the climate is favourable. The highest density occurs in the Huambo, Benguela, and Luanda regions. The African population, particularly, is found along the banks of rivers in the Huambo, Cunene, Huíla, Benguela, Malanje, and Cuanza Sul areas. Coastal fishing towns also have attracted a high percentage of population.

The Portuguese government is attempting to direct the stream of emigrant labour from metropolitan Portugal toward the African states, and its soldiers are being encouraged to settle in Angola after serving there. Local migration within Angola will soon be diverted toward the huge settlement schemes now being established under government sponsorship in the Cunene Basin in the south, where extensive arable land and cattle-raising areas will be available primarily for African settlers. A steady rise in the population has accompanied industrial expansion in other areas.

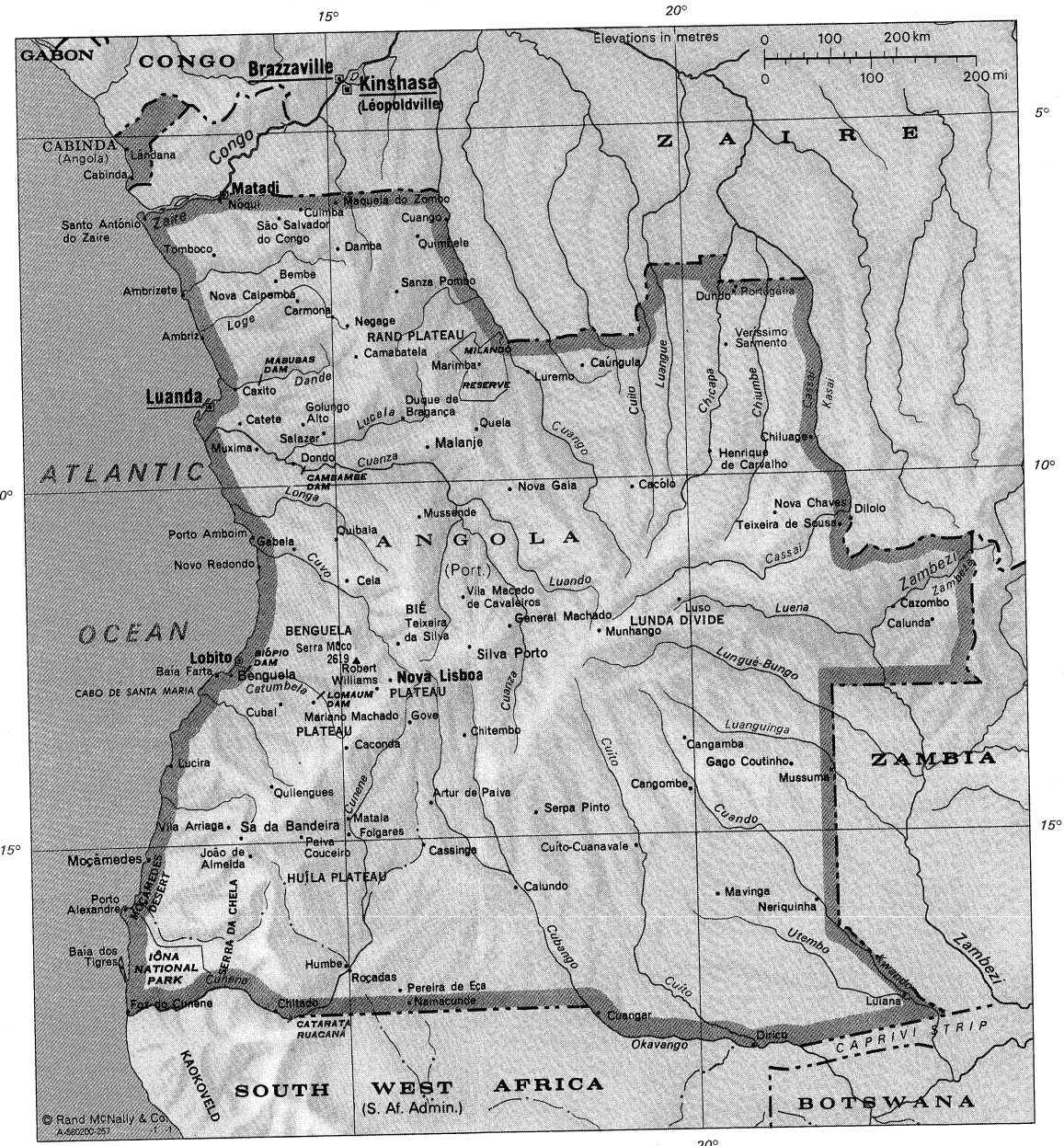
THE NATIONAL ECONOMY

A comparative view. Angola's economic significance to Europe and the world in general has increased in proportion to the exploitation of its rich natural resources. Portugal is the state's chief trading partner, followed by the United States, Germany, The Netherlands, Belgium, the United Kingdom, and France. Japan is an important customer for iron ore. The state is bounded on the north and east by politically hostile countries—Zaire, Congo (Brazzaville), and Zambia—but continues to trade with them on a small scale. Pending completion of the Tanzam railway to Dar es Salaam, Tanzania, Zambia uses Angola's Benguela Railway to deliver copper from its mines to the Angolan port of Lobito. The rate of investment of foreign capital increased in the mid-1970s, and since Mediterranean and Middle Eastern oil has become more difficult for Western countries to obtain, Angola's oil deposits have grown in importance, thus helping to transform the economy.

MAP INDEX

Cities and towns
Ambriz..... 7·50s 13·06e
Ambrizete..... 7·14s 12·52e
Artur de Paiva...14·28s 16·20e
Baía dos Tigres..16·36s 11·43e
Baía Farta.....12·40s 13·11e
Bembe..... 7·02s 14·18e
Benguela.....12·35s 13·25e
Cacólo.....10·07s 19·17e
Caconda.....13·43s 15·06e
Caiundo.....15·46s 17·28e
Calunda.....12·06s 23·23e
Camabatela..... 8·11s 15·22e
Cangamba.....13·40s 19·54e
Cangombe.....14·24s 19·59e
Carmona..... 7·37s 15·03e
Cassinga.....15·08s 16·05e
Catete..... 9·06s 13·43e
Caúngula..... 8·25s 18·40e
Caxito..... 8·33s 13·36e
Cazombo.....11·54s 22·52e
Cela.....11·25s 15·07e
Chiluage..... 9·30s 21·47e
Chitado.....17·20s 13·54e
Chitembo.....13·34s 16·40e
Cuangar.....17·36s 18·39e
Cuango..... 6·17s 16·41e
Cubal.....13·02s 14·19e
Cuimba..... 6·07s 14·38e
Cuíto-Cuanavale.15·10s 19·10e
Damba..... 6·41s 15·08e
Dirico.....17·58s 20·47e
Dondo..... 9·38s 14·25e
Dundo..... 7·21s 20·50e
Duque de Bragança..... 9·06s 15·57e
Folgares.....14·54s 15·08e
Foz do Cunene..17·16s 11·50e
Gabela.....10·48s 14·20e
Gago Coutinho...14·08s 21·25e
General Machado.....12·03s 17·30e
Golungo Alto.... 9·08s 14·46e
Gove.....13·27s 15·54e
Henrique de Carvalho..... 9·39s 20·24e
Humbe.....16·40s 14·55e
João de Almeida.....15·11s 13·41e
Lândana..... 5·13s 12·08e
Lobito.....12·20s 13·34e
Luanda..... 8·48s 13·14e
Lucira.....13·51s 12·31e
Luiana.....17·23s 23·03e
Luremo..... 8·31s 17·50e
Luso.....11·47s 19·52e
Malanje..... 9·32s 16·20e
Maquela do Zombo..... 6·03s 15·07e
Mariano Machado.....13·02s 14·40e
Matala.....14·46s 15·04e
Mavinga.....15·50s 20·21e
Moçâmedes.....15·10s 12·09e
Munhango.....12·12s 18·42e
Mussende.....10·32s 16·05e
Mussuma.....14·14s 21·59e
Muxima..... 9·31s 13·56e
Namacunde.....17·18s 15·50e
Negage..... 7·45s 15·16e
Neriquinha.....15·58s 21·42e
Nóqui..... 5·51s 13·25e
Nova Caipemba.. 7·26s 14·38e
Nova Chaves....10·34s 21·17e
Nova Gaia.....10·09s 17·31e
Nova Lisboa.....12·44s 15·47e
Novo Redondo...11·13s 13·50e
Paiva Couceiro...14·49s 14·34e
Pereira de Eça...17·03s 15·47e
Porto Alexandre .15·49s 11·53e
Porto Amboim...10·44s 13·44e
Portugália..... 7·20s 20·47e
Quela..... 9·16s 17·02e
Quibala.....10·46s 14·59e
Quilengues.....14·05s 14·04e
Quimbele..... 6·28s 16·13e
Robert Williams..12·51s 15·33e
Roçadas.....16·43s 15·01e
Sá da Bandeira..14·55s 13·30e
Salazar..... 9·18s 14·54e
Santo António do Zaire..... 6·07s 12·18e
Sanza Pombo.... 7·19s 15·59e
São Salvador do Congo..... 6·16s 14·15e
Serpa Pinto.....14·36s 17·48e
Silva Porto.....12·22s 16·56e
Teixeira da Silva.12·12s 15·52e
Teixeira de Sousa.....10·42s 22·12e
Tomboco..... 6·48s 13·18e
Veríssimo Sarmento..... 8·10s 20·39e
Vila Arriaga.....14·46s 13·21e
Vila Artur de Paiva, see Artur de Paiva
Vila Macedo de Cavaleiros.....11·30s 16·45e
Vila Teixeira de Sousa, see Teixeira de Sousa

Physical features and points of interest
Atlantic Ocean...11·00s 12·00e
Benguela Plateau.....12·30s 15·00e
Bié Plateau.....12·30s 16·30e
Biópio Dam.....12·30s 13·50e
Cambambe Dam. 9·43s 14·30e
Cassai, *river*..... 7·20s 21·47e
Catumbela, *river*.12·27s 13·29e
Chela, Serra da, *mountains*.....16·00s 13·10e
Chicapa, *river*.... 6·55s 20·38e
Chiumbe, *river*... 7·20s 21·10e
Congo (Zaire), *river*..... 6·04s 12·24e
Cuando, *river*....17·37s 23·22e
Cuango, *river*.... 5·52s 16·35e
Cuanza, *river*.... 9·19s 13·08e
Cubango, *river*...18·01s 21·25e
Cuilo, *river*..... 7·36s 19·25e
Cuíto, *river*.....18·01s 20·48e
Cunene, *river*....17·20s 11·50e
Cuvo, *river*.....10·50s 13·47e
Dande, *river*..... 8·28s 13·21e
Huíla Plateau....15·30s 15·00e
Iôna National Park.....16·30s 12·00e
Loge, *river*..... 7·49s 13·06e
Lomaum Dam....12·55s 14·50e
Longa, *river*.....10·15s 13·30e
Luando, *river*....10·19s 16·40e
Luangue, *river*... 7·00s 19·27e
Luanguinga, *river*.....14·13s 22·00e
Lucala, *river*..... 9·10s 15·30e
Luena, *river*.....12·31s 22·34e
Lunda Divide....12·30s 20·00e
Lungué-Bungo, *river*.....13·28s 22·00e
Mabubas Dam... 8·28s 13·44e
Milando Reserve. 8·30s 17·15e
Moçâmedes Desert.....15·30s 12·10e
Môco, Serra, *mountain*.....12·28s 15·10e
Rand Plateau.... 8·00s 16·00e
Ruacaná Catarata, *falls*...17·22s 14·12e
Santa Maria, Cabo de, *cape*...13·25s 12·32e
Utembo, *river*....17·06s 22·01e
Zaire, see Congo
Zambezi (Zambeze), *river*.....13.00 22.41e



ANGOLA

Diamonds, iron ore, and petroleum

The extent and distribution of resources. *Mineral resources.* Angola had still not been fully surveyed by the early 1970s. The most important mineral resources known then were diamonds, iron ore, and petroleum. Diamonds, mined around Dundo in the Lunda area of the northeast, have long been one of the most valuable exports. The most important iron-ore deposits are at Cassinga (Huíla) in the southwest. The discovery of oil, in the Luanda area and in the northern enclave of Cabinda, has encouraged further exploration, particularly off the Atlantic coast. Production supplies the needs of the state and also allows a margin for export. Other minerals include manganese deposits in the districts of Cuanza Norte, Cabinda, and Malanje; copper in several districts, including the Uíge area; and mica at Dande. Platinum has been located in southern Angola, and there are also deposits of lead, tin, silver, and gold. The subsoil provides bitumens (various mixtures of hydrocarbons), limestone, ocher (an earthy, impure iron ore), marble, gypsum, asphalt, salt, and mineral waters.

Biological resources. Sugar, sisal, cotton, manioc, palm oil, olives, and vines are grown in the northern half of the coastal belt. The southern strip grows progressively more arid toward the south and ends in desert. The main coffee-growing region is the hot and humid north, where cotton, tobacco, maize (corn), rice, and sisal are also cultivated. The chief agricultural production area, which lies along the Benguela Railway, is favourable to corn (maize) and other cereals, as well as to subsistence crops such as beans, peas, and peanuts (groundnuts). The chief tropical rain forests are in the north, and 84 percent of the Cabinda enclave is forested with valuable trees such as mahogany. Cabinda also produces cocoa. An abundance of fruit grows in Angola, including the banana, pawpaw, tamarind, guava, avocado pear, pineapple, mango, cashew nut, and custard apple. Bananas are now a valuable export, whereas pineapples and cashew nuts support important processing industries.

Animal husbandry

The European farmers raise cattle, sheep, and hogs, and the Africans herd cattle and goats. Karakul sheep are raised in the Moçâmedes area for the export of skins. The Huíla District, in the southwest, contains more than half of all the cattle, more than one-third of all goats, one-fifth of all hogs, and over one-third of sheep. The Benguela, Huambo, and Cuanza Sul districts in the west form an important cattle-raising region. The industry is

being encouraged, in view of the growing local demand for meat and the opportunities for exporting frozen meat to Europe. Dairy farming is also developing in answer to the needs of the expanding urban population. The formerly flourishing industry of obtaining beeswax for export was on the decline in the 1970s.

Power resources. Angola's rivers have proved rewarding sources of hydroelectric power. The waters of the Rio Cuanza are harnessed for the Cambambe plant, and those of the Dande for the Mabubas complex. The Salazar plant at Matala is on the Rio Cunene; less important rivers have been used to power smaller plants.

Sources of national income. *Agriculture, forestry, and fishing.* About 2 percent of Angola's arable land is under cultivation. Agricultural establishments range from large European plantations to African and European small holdings and subsistence plots. Europeans operate the large-scale cattle ranches and dairy farms. The most important agricultural product—coffee—was grown over more than 1,200,000 acres in 1964. Other outstanding exports include cotton, maize (corn), sisal, and bananas. There are about 129,000 acres of forest in Angola. The forestry industry depends, however, mainly on the rain forests in the northern enclave of Cabinda, and in the Ganguelas and Zambezi forests, as well as the Miyombo forest along the Benguela Railway; eucalyptus is planted beside the railway near Nova Lisboa to supply fuel. Most timber production from the forests is for export, although there is a growing demand in Angola for timber for manufactured goods, such as furniture.

Fishing is a traditional occupation, and the coast is dotted with fishing settlements. Fish canning has become an important activity, as also are fish processing and the extraction of fish oil. Apart from the ports of Moçâmedes, Benguela, and Luanda, other important fishing centres are Baía dos Tigres, Baía Farta, Lucira, and Porto Alexandre. The fishermen are for the most part European; there is a tendency to use factory ships that possess their own freezing and processing plants, but the main ports are also equipped to process the catch.

Mining and quarrying. Gross mining production in 1970 was more than 11,000,000 tons. The chief exports were diamonds, iron ore, oil, and copper. Diamonds are mined in the Lunda District in the northeast under a concession held by the Angola Diamond Company; other companies have also obtained concessions to prospect. Iron ore is found at Salazar, and most abundantly at Cassinga Mines, which are state-owned and produce about 6,000,000 tons a year. Manganese is found in the Cassala–Quitungo area between Luanda and Salazar. Copper deposits are located near Mocâmedes, Novo Redondo, and in the Cambambe Dam area. There is a little gold at M'popo near Cassinga, and bituminous asphalt and petroleum deposits near Luanda.

Manufactures. The main manufacturing industries are those concerned with food processing, the production of beer and soft drinks, the processing of cotton, tobacco, and the by-products of the fishing industry, the manufacture of cement and textiles, oil refining, and the milling of flour. Other prosperous industries are tire manufacture, plastics, paper, canning, soapmaking and the dairy trade. A local steel industry was in its initial stages in the early 1970s. The processing of fish; the production of beer, tobacco, and cement; flour-milling; and sugar refining were of particular importance.

Energy. The generation of power for industry and irrigation has been an important factor in economic development. Urban living has been improved, industry supplied, and the availability of water for farming and cattle raising considerably extended. About half of the power is supplied by the Sociedade Nacional de Estudo e Financiamento de Empreendimentos Ultramarinos (Sonefe).

The five most important hydroelectric plants in Angola are Cambambe, on the Rio Cuanza near Dondo; Mabubas, northeast of Luanda on the Rio Dande; Lomaum, on the upper Catumbela; Biópio, on the Rio Catumbela; and the Salazar Dam at Matala on the middle Cunene. Cambambe supplies the capital city of Luanda, aided by Mabubas as an emergency source of power. Lomaum provides power for the cities of Lobito, Benguela, and Nova Lisboa in conjunction with the Biópio Dam; and Matala is the chief source of supply for Sá da Bandeira, Moçâmedes, Porto Alexandre, and the Cunene agricultural and irrigation schemes. This latter hydroelectric centre is the first step in a larger plan to be undertaken in conjunction with the South African government under which hundreds of thousands of acres will be irrigated on both the Angolan and South West African borders of the river. The key point in this scheme will be the dam to be built at Gove on the lower Rio Cunene and the major power station at Catarata Ruacaná (Ruacana Falls). Smaller privately owned hydroelectric plants are scattered throughout Angola, as are thermoelectric centres. Total power production in the province in 1970 totalled 644,000,000 kilowatt-hours.

Financial services. Angola belongs to the escudo monetary zone. A widespread increase in the number of banking centres has helped to stimulate internal production and the export trade. The chief banking organizations are the Bank of Portugal in Lisbon, the Bank of Angola—the official bank of issue controlled by the Bank of Portugal—the Commercial Bank of Angola, the Industrial and Commercial Bank, and the Totta-Standard Bank. There is also a Postal Savings Bank and an operative Exchange Fund (Fundos Cambiais). There is no stock exchange, but both Lisbon and local insurance companies maintain branches. Banking

Foreign trade. Portugal is Angola's chief trading partner, supplying the state with most of its imports and absorbing many of its exports. The deficit thus created is balanced, however, by trade with the United States and western European countries. Generally, exports have grown faster than imports since World War II, further strengthening foreign-exchange reserves. More than half of the exports are agricultural products; coffee is the most important commodity, followed by corn (maize) and sisal. Mineral exports have increased since the 1950s, and are comprised of diamonds, iron ore, and petroleum and petroleum products. Processed foods and manufactured goods, such as cement, are also of growing importance as exports. Imports reflect Angola's industrial development. The amount of machinery, worked iron and steel, and transport equipment have increased since the 1960s, while the importing of wine, textiles, dairy products, and wheat have decreased slightly.

Management of the economy. *The public sector.* Development is aided by government allocations within the framework of a series of six-year development plans. The plan for 1968 to 1973 provided for investments of 25,000,000,000 Portuguese escudos. The basic unit of currency is the Angola escudo. A multilateral clearing system operates between Portugal and her overseas territories with an escudo-zone fund that stabilizes the exchange position. Revenue is derived from taxation, profits from government participation in private industry, income on investment, and stipulated sums from big industry.

The private sector. Although Angola may draft its own budget, private finance is regulated by the central government. Monetary exchanges are also subject to control by Lisbon. The Angolan authorities may authorize some local firms to operate, but industries considered of national importance—dealing with foodstuffs, chemicals, engineering, and machinery—remain subject to Portuguese metropolitan control.

Taxation. Most revenue is derived from indirect taxation that includes customs duties, a stamp tax, a wine-consumption tax, and small taxes on such items as petroleum by-products, tobacco, and beer. Direct taxation includes an income tax levied on all workers, an industrial tax on business and industrial firms, and a building tax on property owners. A settlement stamp tax (levied to help finance provincial settlement projects) and one on consumer goods also aid the budget, and a contractual arrangement with the major diamond, petroleum, and railway firms increases government revenue. All Portuguese and foreign companies, the earn- The tax structure

ings of which exceed $17,000 (488,750 Angola escudos) a year, pay a defense-aid tax that ranges from 4 to 30 percent of their earnings.

Trade unions and employer associations. The trade unions (*sindicatos*), like those in European Portugal, are associations of employees formed with government authorization and under government supervision. Trade unions have representatives in the Angolan Legislative Assembly; and they are represented on district boards and municipal councils, provincial settlement boards, and other administrative bodies. The trade unions provide social and employment services and medical care. Their capacity to negotiate labour conditions is restricted because strikes are illegal and because the membership includes managerial personnel. Most of the syndicates' members are skilled European workers, and the large African labour force is essentially unrepresented.

Business associations serve as consulting bodies within the government. They represent the employers and often lobby against labour legislation that is considered unfavourable. The largest organizations are the Commercial Association and the Industrial Association. Agricultural and fishing interests are also organized.

Policies and problems. Angola is an autonomous economic entity within the framework of the general economic organization of Portugal as a whole. The Bank of Portugal controls the Bank of Angola. The reduction of tariffs was begun in 1962 and was directed toward making Portugal and her overseas territories a single domestic market by 1972. The Portuguese treasury controls any external loans contracted by Angola.

TRANSPORTATION AND COMMUNICATIONS

Transport patterns. Local and transcontinental traffic flows through the railway system to Angola's Atlantic ports. The main railway lines are for public use, but there are many small branches belonging to private industrial firms.

The best roads are those that connect the coastal cities and link them with some of the major towns in the highlands. The others are largely single-track dirt roads that are often impassable in rainy weather or are interrupted by unbridged rivers. An internal airline serves the main centres of population, both inland and on the coast; international airlines use the Luanda airport. Rivers are generally not navigable. The Cuanza, however, can carry small craft for about 140 miles upstream; the Congo (*q.v.*) is navigable by oceangoing ships up to Nóqui.

The major roads

Component systems. *Roads.* Fewer than 3,500 miles of Angola's 43,500 miles of roads are paved. The coastal road runs from Santo António do Zaire, at the mouth of the Congo River, south to Luanda, where it splits into two sections. The major branch continues southeastward to Nova Lisboa, turns southwestward to Sá da Bandeira, and then swings southeastward to Namacunde on the border of South West Africa. The smaller branch cuts through southern Luanda Province and then continues along the coast through Lobito and Benguela to Fóz do Cunene at the mouth of the Rio Cunene. Another north–south road runs from Maquela do Zombo on the Zaire border to Negage, where it branches westward to Luanda and southward to Dondo, connecting with the road to Namacunde. Three main roads run from west to east—one from Luanda and one from Lobito to Vila Teixeira de Sousa on the Zaire border, and one from Moçâmedes to Luiana near the Zambian border.

Motor-coach and private haulage services link both agricultural trading centres and towns, and municipal transport services operate in the main urban centres.

Railways. There are some 2,000 miles of publicly used railway lines. The longest private railway is the 840-mile Benguela Railway, which runs eastward from Benguela on the coast, through Nova Lisboa, to the eastern Zaire frontier. It connects with the Katanga Railway and forms part of the trans-African railway line, which terminates at the port of Beira, Mozambique.

While the Benguela Railway is a private concession, the Moçâmedes Railway is owned by the government. It is about 500 miles long and runs eastward from Moçâmedes port through Sá da Bandeira, Matala, and Vila da Ponte (Vila Artur de Paiva) to Serpa Pinto. It has short branch lines running southward from Sá da Bandeira to Chianje and from Dongo to Cassinga. The 378-mile Luanda railway is the oldest in Angola. Its main track (323 miles) runs from Luanda to Malanje, and there are two branch lines—one to Dondo, and one to the Golungo Alto coffee area. One of the shortest railways is the privately owned Amboim Railway, which is less than 100 miles long and runs from the coast at Porto Amboim through the coffee plantation area to Gabela.

Port facilities. The chief ports are Lobito, Luanda, and Moçâmedes. Next in importance are Lândana, Porto Amboim, Cabinda (which is growing rapidly since the discovery of oil offshore), Santo António do Zaire, Ambrizete, Novo Redondo, Benguela, Porto Alexandre, and Baía dos Tigres. There are also several small shipping harbours, some of which can be used by oceangoing vessels. Lobito has two L-shaped wharves, totalling some 6,500 feet in length, as well as the only floating dock in Angola. The port is fully equipped with storehouses, railway lines and marshalling yards, liquid-fuel depots, and silos for loading and storing grain. As a terminus of the Benguela Railway, it handles Katangan copper, agricultural products, livestock, industrial goods, and iron ore from inland Angola.

Luanda Harbour

Luanda Harbour is deep and easy of access, with about four square miles suitable for berthing ships, and five first class quays. It is linked by rail with the main Luanda railway station and is equipped with up-to-date loading and unloading equipment. It has a special port for loading ore and is an outlet for coffee, cotton, timber, and cereals. Moçâmedes, at the railhead of the Moçâmedes Railway, is the outlet for iron ore from the Cassinga mines in Huíla District and also serves an inland region potentially rich in industrial and agricultural products. It has refrigeration plants to store meat from the food industries of Sá da Bandeira. North of Lobito, Porto Amboim serves the coffee region around Gabela in the Cuanza Sul District.

Air transport. The main international airport at Luanda is a point of call for airlines operating between Europe and southern Africa. The inland city of Nova Lisboa, to the southwest, is equipped for use as an alternative landing site. The Air Transport Directorate (Direcçao de Exploração dos Transportes Aéreas) links the main towns and cities. Its chain of 27 airports includes those at Benguela, Cabinda, Carmona, Lobito, Luso, Malanje, Moçâmedes, Sá da Bandeira, Silva Porto, Santo António do Zaire, Portugália, Porto Amboim, and Novo Redondo. There are more than 300 runways and airstrips for small planes, many of them used by the ever-increasing number of aero clubs.

Telecommunications. All internal telephone and telegraph services are government-operated through the Correios e Telecommunicações de Portugal. The Companhia Portuguêsa Rádio Marconi operates international telegraph and telephone services connecting Luanda and Lisbon. The main cities are linked by telephone, and there is also a telex service. There was no television in the early 1970s, but the province was covered by a radio network that included the Emissora Oficial de Angola (Radio Angola) and the Radio Eclesia, owned by the Roman Catholic Church. Radio clubs serve as a useful link in emergencies, such as warning when guerrilla attacks occur or when medical help is needed. There is a medium-wave central transmitter at Mulenvos near Luanda, and there are three short-wave transmitters that cover the state. Large companies have their own broadcasting stations. Programs are beamed to Angola from Portugal, as are those of revolutionary movements in exile, and transmissions from Brazil and the Soviet Union can be received.

ADMINISTRATION AND SOCIAL CONDITIONS

The Organic Law

Government structure. *Constitutional framework.* Angola is governed within the framework of the Organic Law covering all Portuguese overseas territories. It guarantees a degree of administrative and financial au-

tonomy. The Political and Administrative Statute of the Province of Angola, which regulates the application of the Organic Law to the state, became operative in 1964. The state is governed as an integral part of Portuguese territory, as if it were a part of Europe. Angola is represented in the National Assembly in Lisbon by 12 elected candidates (see below) and in the Assembly's Corporative Chamber by chosen delegates.

The overseas minister in Lisbon is directly responsible to the prime minister and the Council of Ministers for the general administration of Angola. He usually acts in consultation with the Overseas Council, the supreme consultative body within the Overseas Ministry, which has two Angolan members. The Council of Ministers appoints the governor general, who serves a four-year term in the state. He administers the state, with advice from the Consultative Council. The Angolan Legislative Assembly consists of 53 members, among whom are 32 members chosen by direct vote of the registered electorate, and the remainder nominated by designated corporate groups (these include industrial, economic, and cultural groups). Provincial secretaries are appointed by the Overseas Ministry.

Local government. Angola is divided into 16 districts: Cabinda (capital, Cabinda); Zaire (São Salvador do Congo); Uíge (Carmona); Luanda (Luanda); Cuanza Norte (Salazar); Cuanza Sul (Novo Redondo); Malanje (Malanje); Lunda (Henrique de Carvalho); Cuando Cubango (Serpa Pinto); Benguela (Benguela); Huambo (Nova Lisboa); Bié (Silva Porto); Moxico (Luso); Moçâmedes (Moçâmedes); Cunene (Pereira de Eça); and Huíla (Sá da Bandeira). There are almost 100 townships, or *concelhos*, and 20 circumscriptions (small administrative units known as *circunscrições*) that are slightly less developed economically and socially. The townships and circumscriptions are further divided into parishes and administrative posts, the former constituting both an administrative and a religious unit. The districts are controlled by governors, and administrators hold authority in the township and circumscription. Local officers are in charge of parishes. Advisory bodies include district councils (*juntas distritais*), municipal councils (*câmaras municipais*), municipal committees (*commissões municipais*), and parish and local boards (*juntas de Freguesias* and *juntas locais*). The governor general appoints the presidents of the municipal councils. Areas inhabited by Africans living in their traditional manner under customary law are designated as *regedorias*. There are more than 1,300 *regedorias;* these units consist of villages or groups of villages, and they contain most of Angola's population.

The political process. The citizens of Angola participate every seven years in the election of the president of the Portuguese Republic through local bodies that form part of the electoral college. Angola's 12 representatives in the National Assembly in Lisbon are elected by direct vote every four years.

Outlawed African nationalist groups

As in European Portugal, the only legal political party is the Acção Nacional Popular (National Popular Action). Breakaway opposition consists of several outlawed African nationalist groups, the most important of which are the Movimento Popular de Libertação de Angola (MPLA), which is based in Brazzaville, and the Frente Nacional de Libertação de Angola (FNLA), which is based in Kinshasa.

The right to vote is extended only to those who meet specified requirements. Only men and women who are over 21 years of age, literate in Portuguese, and included in a census are granted the franchise. The vote is also extended to those persons who are the heads of legitimately constituted families and who have paid a property tax.

These requirements virtually eliminate the African population from the electorate. To be the head of a "legitimate" family a citizen must be married under civil rather than customary law, and the property tax implies individual ownership; most Africans, however, live under a system of collective ownership. Electoral status is determined by civil registration of those who live under civil law, but most Africans live under customary law in the *regedorias* and are not eligible for registration. In general, only those Africans who have adopted the Portuguese language, law, and customs are able to vote.

Debate upon government-proposed alterations to the Portuguese Constitution in the National Assembly at Lisbon has resulted in more administrative autonomy being given to the overseas territories.

Justice. The Portuguese Criminal Code applies to Angola. The Overseas Judicial District of Luanda includes 13 judicial districts, or *comarcas*, in Angola and also that of the Atlantic island province of São Tomé. The highest judicial appeal is to the Supreme Court of Justice in Lisbon. There are special courts for administrative and labour disputes, as well as juvenile courts. African customary law is operative in the *regedorias* and is applied by Portuguese officials.

Civil law applies to those citizens with one or two European parents or to those Africans and their descendants who have elected to be governed by it. Civil law is applied when a case involves individuals of different legal status or when a case involves Portuguese sovereignty. Customary law has been weakened for lack of codification and because appointed African officials often have no traditional authority.

The armed forces. Angola is one of the seven military regions of Portugal. Since the outbreak of hostilities in March 1961, the army in Angola has been reinforced to an estimated 60,000 troops. Their activities cover not only defense of the frontiers and internal security but also psychological warfare through medical care and educational work among the local inhabitants. They also engage in the building of roads and bridges and other permanent structures. The army is supplemented by a small navy and an air force. There is also a paramilitary state organization composed of volunteers and of civil defense forces.

School attendance

Services. *Educational services.* School attendance in 1970 was officially cited as about 450,000 pupils in more than 4,200 establishments. The University of General Studies, in Luanda, which is similar to universities in Portugal, has over 1,000 students. The educational system includes government-operated free schools, officially recognized private schools, officially subsidized Catholic free schools, and other unsubsidized Catholic and Protestant mission schools. Many large companies also run their own schools. Further, there are commercial, industrial, and technical schools, agricultural colleges, and training centres for nurses and health officers, as well as teachers' training colleges.

Both European and African students attend the same classes. School attendance is compulsory for all children between six and 12 years of age who live within access to a free school. Tuition is free in all government and most Catholic primary schools and teacher-training institutions below the university level. Government scholarships are available for post-primary schools. The literacy rate has improved from 1 percent in 1950 to 30 percent in 1970 because of a major expansion of educational facilities in the rural areas.

Health and welfare services. There were four state, 14 regional, and 70 rural hospitals in the early 1970s. These were supplemented by health centres, dispensaries, maternity homes, tuberculosis clinics, one mental hospital, and leper colonies. Medical outposts serve the distant rural areas to which mobile brigades make regular visits. Catholic and Protestant missions operate their own hospitals and the larger companies extend their private-health services to the surrounding districts, but traditional African medicine continues to be important in the rural areas. In 1964 revised health and welfare services provided for the instruction of the population in health and sanitation and the protection of the public against endemic diseases. The Escola Nacional de Saùde Pública e de Medicina Tropical in Lisbon aids in fighting such diseases as malaria, sleeping sickness (trypanosomiasis), leprosy, yaws, hookworm, and schistosomiasis (a severe disease marked by blood loss and tissue damage).

Housing. Urban Europeans live in blocks of apartments or Portuguese-style villas equipped with water, light, and sanitation facilities. In contrast, rural Africans live in compounds composed of wattle-and-clay huts that vary in style according to tribal tradition. The native quarters, *muçeques*, are shantytowns on the outskirts of the cities, consisting of wood and corrugated-iron huts. The towns are gradually being replaced with low-priced blocks of concrete dwellings. In the organized agricultural settlements (see above *Demography*), there are government-built houses with plaster walls and tiled roofs.

Police. The Public Security Police (PSP) and the Criminal Investigation Police (PJ) are responsible for maintaining public order and for suppressing general criminal activities, while the Lisbon Department of Security (DGS) controls political activities. The Customs Guard has been incorporated in the PSP. A corps of local militia and the army aid in policing the state.

The Rural Labour Code

Social conditions. *Wages and cost of living.* The Rural Labour Code of 1962 was aimed at correcting abuses in the labour system by establishing minimum wages and better working conditions for unskilled workers. In general, Europeans occupy the higher wage brackets, and their salaries are better than those in Portugal. Africans are employed primarily as unskilled labourers. Increasing educational facilities, coupled with the growing demand for labour, however, are developing the Africans' wage economy. During 1970 the Luanda consumer price index (whose base of 100 was established in 1955) was 136, having risen 11 points during the year, but the cost of living continues to rise, affecting rural Africans and unskilled labourers primarily.

Health conditions. Widespread malnutrition, climatic conditions favouring tropical disease, and the widespread lack of sanitation are conducive to disease. Hence, about one-third of the rural children die in infancy. Tuberculosis, malaria, and yaws are major problems, and in urban areas disease is encouraged by unhealthy conditions in the African quarters. Deficiency diseases are prevalent.

Social and economic divisions. Although the Portuguese government officially practices a policy of social integration, there is a gulf between the Europeans and the African tribal groups, and there is but little mutual integration between tribes. A small number of educated Africans, however, have managed to attain European social status and there is no colour bar as there is in South Africa. The social divisions are reflected in differences in economic status.

CULTURAL LIFE AND INSTITUTIONS

European culture in Angola depends upon European Portugal. The state, nevertheless, has its own cultural institutions, newspapers, and reviews in the Portuguese language, as well as broadcasting programs and other entertainments. There are museums, libraries, and cultural associations. Visiting theatrical and other artists from Portugal arrive regularly, and international sports events are programmed.

Sculpture and rock painting

The arts. Sculpture is one of the arts of Angola and the artists of Lunda, Bailundo, Cuango, and Bié are outstanding. Pottery with incised designs and carved wooden articles, such as furniture, combs, sticks, and staffs, are characteristic. Stylized rock painting is found, particularly among the Bushmen of the south, and rock paintings have been discovered in a cave at the base of Mt. Chitundulo in the Moçâmedes Desert. The crafts also include the weaving of baskets and mats and the making of bracelets and necklaces of shells. Both men and women tattoo their faces and bodies, and the women have evolved elaborate hair styles typical of their region and tribe. African music has evolved for use in religious ceremonies, to accompany work, and for recreation. Percussion instruments set the pace, and the art of drumming is popular. In the Congo Basin a kind of portable, or finger, piano is played. Ritual dancing exists for such occasions as the coming of the rains, harvesting, birth, marriage, death, hunting, and fishing, while literature is largely confined to oral poetry, songs, myths, and folk tales, although a small school of modern Angolan poets and other writers has arisen in the 20th century.

Cultural institutions. There are cultural associations and clubs, both for Europeans and assimilated Africans. Some of these are run by the larger companies for their workers. The museums include the Dundo (ethnological) Museum, sponsored by the Diamond Company of Angola; the Angola Museum in Luanda; and the Angola Scientific Research Institute. The Institute of Angola, in Luanda, promotes prizes for authors, artists, and scientists, and encourages scholars, and there are government-sponsored youth and cultural organizations in most towns. But because of the official fear that they may be used for political purposes, the organizations, particularly those for assimilated Africans, may be curbed. On such grounds the Regional Association of Natives of Angola (Anangola) was disbanded in 1966 and replaced by the Beneficent and Cultural Association, whose charter stipulates that religious, ethnic, or political discussion are to be avoided; it also proclaims a commitment to the defense and preservation of the integrity of the Portuguese nation. In all, there are about a dozen libraries in Angola, including a National Library in Luanda. All large towns have cinemas, and there are more than 100 sports clubs in the state.

Press and broadcasting. Angola has more than 50 newspapers, reviews, and bulletins. As in European Portugal, the press is censored; and most newspapers are European-owned, are limited to the urban areas, and are published in Portuguese. There are some 20 radio stations and a large chain of amateur radio operators as well. Broadcasts are in the Portuguese and African languages. Since 1961, in an effort to counter liberation broadcasts from other countries, radio has been extended to the rural areas.

Prospects for the future. Angola's future is dependent upon Portugal's retention of her African possessions. As a member of the European Free Trade Association (EFTA), the initiation of talks between Portugal and the Common Market, with a view to some form of associative cooperation, is bound to affect Angola. The state's trade with the EFTA and the European Economic Community (EEC) countries is already an important economic asset. In the United Nations, Portugal is confronted by a combination of the Communist and Afro-Asian blocs that disputes her right to continue administering her former colonies, and that calls for urgent preparation of Angola's peoples for independence.

It is probable that if Portugal continues to curb warfare in the north and east of Angola and obtains enough investments from abroad, the state will remain stable and economically viable. Nevertheless, since Angola forms an intrinsic part of the pattern of African states, its future would be vitally affected by any mass onset against Portuguese control elsewhere in Africa. But the continuing Portuguese presence also depends upon bridging the gulf between the majority of the African inhabitants of Angola and the economically privileged European inhabitants of the state.

BIBLIOGRAPHY. NORBERTO GONZAGA, *Angola: A Brief Survey* (1967), an intelligent and very well-written compendium sponsored by the Portuguese government (in English and French); DAVID M. ABSHIRE and MICHAEL A. SAMUELS (eds.), *Portuguese Africa: A Handbook* (1969), an up-to-date survey of all sectors of activity in Angola; FOREIGN AREA STUDIES DIVISION OF THE AMERICAN UNIVERSITY, *Area Handbook for Angola* (1967), an intensively detailed survey prepared for military or other personnel; HUGH KAY, *Salazar and Modern Portugal* (1970), a well-researched book by an expert on the Iberian Peninsula and the Portuguese overseas territories; PORTUGAL, MINISTERIO DO ULTRAMAR, AGENCIA GERAL DO ULTRAMAR, *Províncias Ultramarinas Portuguesas*, vol. 1 (1961), a basic description of the physical features, together with other facts; JOSE REDINHA, *Distribução Étnica*, 5th ed. (1969), an authoritative study of the ethnology of Angola; FERNANDO DE CASTRO PIRES DE LIMA (ed.), *A Arte Popular em Portugal*, 3 vol. (1963); PAUL FORDHAM, *The Geography of African Affairs* (1965), a general survey covering geography, ethnology, resources, politics, etc.

(J.Sh.)

Anguilliformes

True eels are elongated or even wormlike bony fishes of the order Anguilliformes and include the common freshwater eels as well as the voracious marine morays. Only the freshwater eels (family Anguillidae), which are in places abundant and greatly valued as food, are of major economic importance. All eels probably pass through an extended larval phase (leptocephalus) in the open ocean and undergo metamorphosis to a juvenile stage that is a smaller version of the adult. At maturity, eels range from ten centimetres (in the deep sea *Cyema atrum*) to 3½ metres (in the moray *Thyrsoidea macrura*). They occur to considerable depths in most oceans and are greatly diverse in tropical seas. They range in colour from drab gray or black (in deep sea species) to colourful and patterned (in tropical reef species).

NATURAL HISTORY

Life cycle

Eels have a remarkable life cycle. Broadly, it consists of development and early growth in the open ocean: the planktonic (free-floating) dispersal of eggs and larvae, metamorphosis, juvenile and adult growth, and the migration of maturing adults to an oceanic spawning area. Eels share the leptocephalus phase with several other orders (Elopiformes, Saccopharyngiformes, and Notacanthiformes). A prolarva, hatching from a relatively large egg (up to 2.5 millimetres diameter), rapidly becomes a leaflike leptocephalus, which floats in the surface layers of the open ocean for as long as 2½ years before metamorphosing.

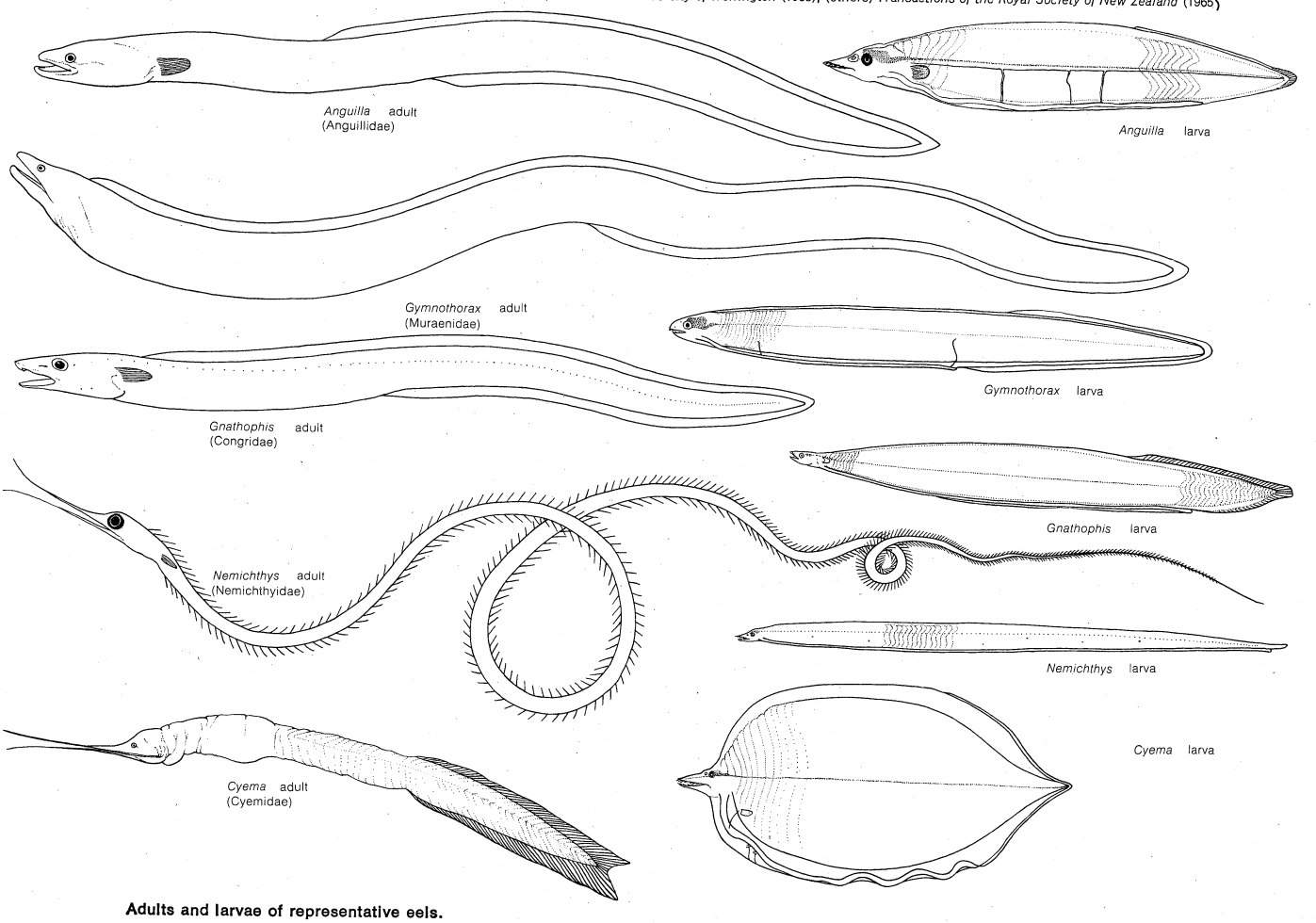
Although the leptocephali were once thought to have been fishes of a distinct group, their relationship with the Anguilliformes was soon recognized from transitional specimens that showed larval and adult characters. They proved so difficult to identify, however, that new larval types were named as species of the genus *Leptocephalus* (though they cannot actually be considered different species from the adults that produced them), accounting for the several hundred forms known.

Leptocephali are not uncommon in the upper 500 metres of the ocean, a distribution that may be associated with the availability of food (diatoms and minute crustaceans). Their predators include various pelagic fishes. In tropical eels larval life is possibly four to six months, but temperate species may spend upwards of a year as larvae. During this time leptocephali, in the presence of suitable currents, may disperse widely from the adult spawning area. Working on massive collections of larvae from 1905 to 1930, a Danish biologist, Johannes Schmidt, established the early life history of the European and American freshwater eels. Although parts of his work have been questioned, his description of a western Atlantic spawning and a trans-Atlantic dispersal of leptocephali of these eels still stands.

After reaching full growth, the larva begins a rapid metamorphosis in which the following progressive changes are typically involved: the body becomes cylindrical and greatly reduced in bulk, perhaps by as much as 90 percent by weight; the anal vent advances from its subterminal position to about the midpoint; larval teeth are lost; the snout becomes rounded; the dorsal fin originates farther forward; and the larval melanophores (black pigment cells) disappear. Other changes, such as the loss of the pectoral fins or a reduction of body length, may also occur.

Leptocephali are markedly unlike their adults and the metamorphic changes are so great that a fundamental problem arises in the correlation of the great variety of known leptocephali with their adults. Metamorphosis has

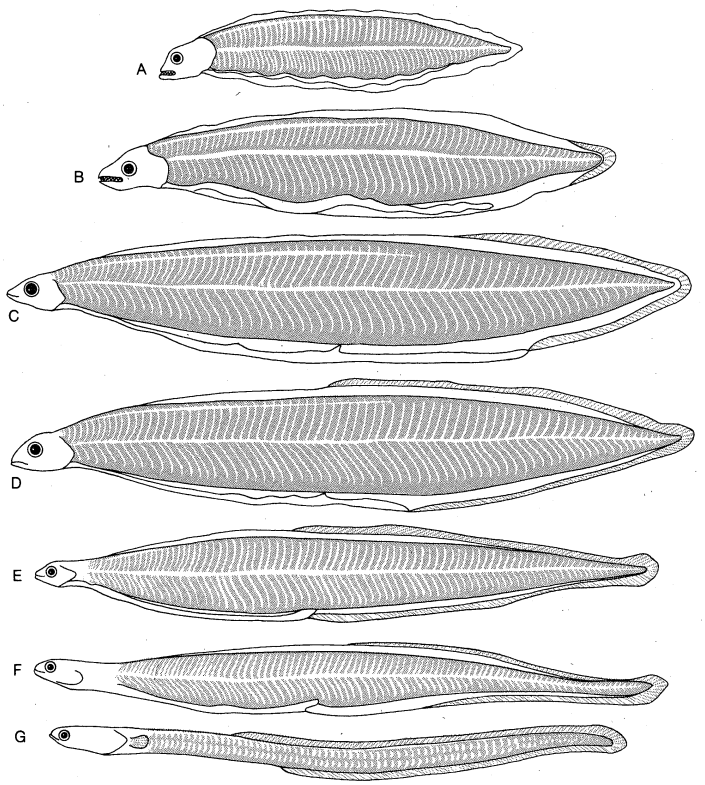
Drawing by J. Helmer based on (*Cyema* adult) by courtesy of Carl L. Hubbs; from (*Gymnothorax* larva) E. Hearld, *Living Fishes of the World*, Doubleday & Co., Inc., (*Anguilla* adult, *Nemichthys* adult) P.H.J. Castle, "World of Eels" (1968) in *Tuatara* vol. 16, no. 2; (*Anguilla* larva, *Gnathophis* adult, *Gnathophis* larva) *Zoology Publications from Victoria University of Wellington* (1963), (others) *Transactions of the Royal Society of New Zealand* (1965)



Adults and larvae of representative eels.

been followed through in aquariums and deduced from progressive growth series in plankton samples. Certain characters survive metamorphosis and are important in the recognition of eel species. These include the number of muscle segments (myomeres); the development of dorsal, anal, and caudal fin rays; and the relative positions of the renal vessels and gallbladder. In many leptocephali, the larval melanophores also remain into the juvenile (or elver).

By courtesy of Direction generale des Peches maritimes, Ministere de l'Industrie et du Commerce, Quebec



Metamorphosis of the American eel. (A–C) Larvae or leptocephali of various sizes. (D–F) Larvae in the process of metamorphosis. (G) Glass eel, mature stage.

The process of metamorphosis

Metamorphosis involves physiological and behavioural as well as structural changes, particularly those related to the assumption of a deep-sea, shallow-water or freshwater mode of life. Metamorphosis is the mechanism by which the leptocephalus, after a period of growing, feeding, and competing with other similarly organized planktonic animals, can enter a markedly different habitat, where body shape, differentiated feeding mechanisms, sense organs, and body coloration, play an important role in survival. Metamorphosis in all eels is probably completed in the open ocean. The annual invasion of freshwaters by *Anguilla* elvers is a locally well-known process; it occurs during October–March in Europe and in spring in other temperate regions.

During several years' growth to maturity, eels are essentially carnivores, feeding diversely on planktonic or benthic (bottom-living) animals. Maturity is reached after about ten years in the European freshwater eel but possibly much earlier in tropical marine species. The process of growth and maturation has been most closely studied in the European freshwater eel. In this species both sexes pass through successive phases of neutrality, precocious feminization, and juvenile hermaphroditism prior to becoming definitively male or female, the sex being mainly determined by environmental factors.

All eels apparently undergo a short or long distance migration at maturity to a spawning area within the area of adult distribution (in most tropical marine eels) or some distance from it (temperate Anguillidae and Congridae). These areas are generally located over the continental slope or in ocean basins some distance offshore.

During their juvenile and adult life, most eels are solitary fishes, swimming slowly by means of sinuous lateral movements of the body and median fins. Some species burrow rapidly, using a pointed tail and backward body movements. Morays and congers inhabit rock crevices, while certain congrids (Heterocongrinae) form vast colonies of several hundred individuals in tropical reef areas.

FORM AND FUNCTION

An eel is distinguished externally from most other fishes by its elongated body, which is seldom laterally compressed. A continuous dorsal, anal, and caudal fin runs around the tail tip; pelvic fins are always absent; and gill openings are usually reduced. The body covering is usually scaleless. Minor departures from this overall body plan occur in the various eel families and are correlated well with different modes of life.

Typically, a leptocephalus is elongate, laterally compressed, transparent, and gelatinous, with prominent, W-shaped myomeres and sharp, forwardly directed larval teeth. At full growth, eel larvae are five to ten centimetres but may be much larger (45 centimetres in the case of the snipe eel *Nemichthys scolopaceus*). Leptocephali are at least as diverse morphologically as their adults: some are filamentous, while others are deep-bodied, even resembling a small dinner plate in shape and size.

Pectoral and median fins are present in most leptocephali but may disappear during metamorphosis. Eyes are usually normal in shape but are occasionally telescopic, and the rostrum (an anterior beaklike projection) may be greatly extended forward from the snout. The attenuate viscera are located along the ventral aspect, below the myomeres. There is usually a straight, unmodified intestine, and the anal vent is often forward of the tail. In some families the larval gut is swollen or festooned at various points along its length, a modification of unknown significance. A long liver, with a well-defined gallbladder, occurs anteriorly on either side of the intestine. The developing adult kidney lies at about midlength. The organs are supplied and drained by many vertical blood vessels, of which the last one or two are the largest. The position of the last of these vessels equates well with the division between precaudal and caudal vertebrae and indicates the approximate position of the adult kidney.

In most leptocephali melanophores occur almost anywhere on or in the body. These vary from minute and compact to large and irregularly shaped, often distributed along the gut and laterally in a variety of patterns.

In adults there is a strong tendency towards a reduction or loss of fins and a streamlining of the profile, sometimes also with an attenuation of the body. The gill region is variously elongated by a posterior displacement of the gill arches, accompanied by separation of the pectoral girdle from the cranium. There are well-developed pharyngeal tooth plates, and the pharynx has assumed an important function in the movement of food into the esophagus. It also serves as an effective pump in the virtual absence of suction by the gill covers (opercula) for the passage of the respiratory current. The opercular series is much reduced, although numerous curved branchiostegal rays (internal gill supports) are strongly developed to support the long throat wall. Respiration through the skin is important in *Anguilla* and probably also in other eels.

EVOLUTION, PALEONTOLOGY AND CLASSIFICATION

Studies of the few known fossil eels and the comparative anatomy of adults and larvae suggest that eels arose in the Cretaceous Period from two or more types that had at least some characters of the Elopiformes (tarpons, bonefishes, etc.).

Distinguishing taxonomic features. The important characters in determining the taxonomic ranking of eels include: the number and arrangement of myomeres and fin rays; the relative positions of mouth, nostrils, gill openings, renal vessels, and gallbladder; and the number and arrangement of teeth, skull bones, and vertebrae.

Also important in identification are larval structures and pigmentation patterns.

Annotated classifiication. The classification used in this article is a synthesis of the work of many authorities. It is certain to be further modified as more is learned about some of the little-known families.

ORDER ANGUILLIFORMES

Elongated, bony fishes of streamlined profile; continuous median fin with up to 650 soft rays; pelvic fins absent; reduced cycloid scales (*i.e.*, with skin-covered bone overlying a layer of fibrous connective tissue) sometimes present; swimbladder with duct; cranial bones reduced, particularly the palatoquadrate and opercular series; pectoral girdle detached from cranium; vertebrae up to 700, with fused centra and arches; a leptocephalus larva. There are 19 families, about 140 genera, and more than 500 species. One family in freshwater; the remainder marine, in all oceans, mainly tropical Atlantic and Indo-Pacific, to considerable depths.

Suborder Anguilloidei

Frontal bones of skull paired, supraoccipital present.

Family Anguillidae (freshwater eels)

Worldwide, but not on Pacific coast of America and south Atlantic coasts. Scales present, gill slits ventrolateral. Important as food.

Family Heterenchelyidae

Tropical Atlantic. No fins, mouth large.

Family Moringuidae (worm eels)

Tropical Indo-Pacific and western Atlantic. Anus in posterior half of body, degenerate, burrowing.

Family Xenocongridae (false morays)

Pantropical. Burrowing.

Family Muraenidae (morays)

Pantropical to subtropical. No pectorals, large mouth, often brightly coloured, voracious, sedentary.

Family Myrocongridae

South Atlantic. Laterally compressed, poorly known.

Suborder Nemichthyoidei

Frontals paired or fused, supraoccipital present or absent, paired nostrils close in front of eye.

Family Nemichthyidae (snipe eels)

Bathypelagic (deepwater), worldwide. Jaws greatly extended, minute teeth.

Family Serrivomeridae (sawtooth snipe eels)

Bathypelagic, worldwide. Jaws moderately extended; blade-like teeth on vomer bones.

Family Cyemidae (bobtail snipe eels)

Bathypelagic, worldwide. Jaws extended, body truncated.

Suborder Congroidei

Frontals fused, supraoccipital present.

Family Congridae (congers)

All oceans to considerable depths.

Family Muraenesocidae (conger pikes)

Pantropical. Large teeth, voracious.

Family Nettastomatidae (witch eels)

Deepwater. No pectoral fins.

Family Nessorhamphidae (duckbilled eels)

Bathypelagic. Duckbill snout.

Family Derichthyidae (neck eels)

Bathypelagic. Short snout.

Family Ophichthidae (snake eels)

All oceans, many branchiostegals, caudal reduced or absent.

Family Macrocephenchelyidae

Deepwater, Pacific. Rare.

Suborder Synaphobranchoidei

Frontals fused, supraoccipital present, 3rd hypobranchial directed backward, larva telescopic-eyed.

Family Synaphobranchidae

Deepwater, worldwide. Gill slits ventrolateral to ventral, united. Scales present.

Family Simenchelyidae (parasitic eels)

Deepwater, worldwide. Semiparasitic, mouth transverse, scales present.

Family Dysommidae

Deepwater, worldwide. Anus near pectoral region. No scales.

Critical appraisal. The eels were early split into the Colocephali (morays) and Enchelycephali (others). Later the suborder Carenchelyi was added (for the Derichthyidae). Two fundamental but undesignated groups, separated on the poorly evaluated character of fused versus separate frontal bones, have been considered the main evolutionary lines, forming the suborders Anguilloidei and Congroidei. In the absence of a workable alternative this division must stand. In 1940 L.S. Berg added the Nemichthyoidei, the subordinal position of which requires further investigation. Various osteological and larval characters suggest that the Synaphobranchoidei, as recently recognized, may truly represent a further lineage. Some would include still another suborder, the Saccopharyngoidei. These strange deep-sea fishes, while closely related to true eels, appear distinctive enough to merit separate ordinal rank as the Saccopharyngiformes.

The limits of the 19 living eel families are now broadly defined. However, their interrelationships and the existence of subordinal groupings will only be firmly established by an overall survey of a range of characters and not only of specific structures throughout the order. Rare forms need investigations and larval characters require consideration. Closer study is needed of environmental effects on segmental features, and fundamental questions of species nomenclature have yet to be settled.

BIBLIOGRAPHY. L. BERTIN, *Les anguilles*, 2nd ed. (1951; Eng. trans., *Eels: A Biological Study*, 1956), a comprehensive, readable account of the biology of freshwater eels; J. SCHMIDT, "The Breeding Places of the Eel," *Rep. Smithson. Instn. (1924)*, pp. 279–316 (1925), a summary by the original researcher of the classical biological study of eels; A.F. BRUUN, "The Breeding of the North Atlantic Freshwater-Eels," *Adv. Mar. Biol.*, 1:137–170 (1963), an important review of current controversy regarding the eel life-cycle; L. BERTIN and C. ARAMBOURG, "Anguilliformes," in P.P. GRASSE (ed.), *Traité de Zoologie*, 13:2314–2327 (1958), a classical text in French with a good coverage of eels; P.H.J. CASTLE, "The World of Eels," *Tuatara*, 16:85–97 (1968), a general article of college level difficulty, and "An Index and Bibliography of Eel Larvae," *Spec. Publs. Instn. Ichthyol. Rhodes Univ.*, 7:1–125 (1969), a comprehensive bibliographic index to leptocephali.

(P.H.J.C.)

Anhwei

Anhwei (An-hui in Pin-yin romanization), one of the smallest of the 21 provinces of the People's Republic of China, covers an area of 54,000 square miles (139,900 square kilometres) and stretches for 400 miles from north to south. The population numbers about 35,000,000 (1970). Anhwei, which is landlocked, is bounded by the provinces of Kiangsu to the northeast, Chekiang to the southeast, Kiangsi to the south, and Hupeh and Honan to the west. Its northern extremity barely touches the southern extremity of Shantung. Its name means Beautiful Peace and is derived from the names of two cities—An-ch'ing and Hweichow (now She-hsien). The capital, Ho-fei, is located in the heart of the province.

Anhwei was long one of China's poorest and most undeveloped areas. Since the establishment of the People's Republic in 1949, however, successful attempts have been made to utilize the province's economic and human resources. Vast irrigation schemes on the major rivers have alleviated severe periodic flooding and have also provided increased agricultural land and electric power. Soybeans, tea, and rice are intensively cultivated, and local coal and iron deposits are the basis of a growing iron and steel industry. (For an associated physical feature, see YANGTZE RIVER.)

History. During the Warring States period of the Chou dynasty, between 481 and 221 BC, Anhwei formed part of the large southern state of Ch'u. Between 221 and 206 BC, the Ch'in dynasty unified the states, and a great southward migration along the natural highway of the North China Plain and the Huai Ho (Huai River) Basin began. Anhwei became the first part of southern China to be settled by the Han-jen, or Chinese people. Unrest following the fall of the Han Dynasty in AD 220 led to further immigration into the area.

The Yangtze River Basin subsequently became the granary of the empire, and an improved transport and canal system was developed across northern Anhwei to

The early canal system

carry tribute grain to the capital. The Old Pien Canal was built along the upper Sze Ho from Süchow to K'ai-feng in Honan and from there to Lo-yang in Honan. It was later superseded by the New Pien Canal, built during the Sui dynasty between AD 581 and 618. The new canal ran along the Kuei Ho and then cut across the region to K'ai-feng, forming the main line of communication to the capital. During the 12th century, Anhwei was the scene of bitter battles between the Southern Sung emperors and the invading Juchen. After the establishment of the Mongol capital at Peking, the Yüan dynasty (AD 1279 to 1368) constructed the Grand Canal to the east, connecting Hanchow in Chekiang Province with Peking, and the previously built waterways fell into disuse.

In the early 1850s the Huang Ho, or Yellow River, made one of its great changes in course, flowing into the Gulf of Chihli, or Po Hai, north of the Shantung Peninsula instead of south into the Yellow Sea, or Huang Hai. The loss of water for the Huai Ho Basin resulted in great distress for the farmers of northern Anhwei. Subsequent peasant risings—together with the Taiping Rebellion of 1850 to 1864—led to prolonged unrest in the province, resulting in widespread devastation.

The Yangtze was opened to foreign shipping in 1860, but it was not until 1877 that the walled city of Wu-hu in Anhwei was opened to international trade. Although it was the province's only treaty port, the city never figured prominently in overseas commerce.

In 1938 the Huang Ho was temporarily diverted south of Shantung by the Nationalist government, which blew up the river's dikes in Honan in an attempt to stem the advance of Japanese invaders. The river waters then surged south to Hung-tse Hu (Hung-tse Lake) on the Anhwei border, flooding an enormous area at the cost of about 1,000,000 lives.

During World War II, most of Anhwei was occupied by Japanese forces, but, due to the resistance of the Chinese inhabitants, Japanese control could be effectively enforced only during daylight hours. Between 1946 and 1949 the province was controlled by the Kuomintang forces. It formed part of the East China Administrative Area from 1949 to 1954, when it became a province of the People's Republic.

The landscape. *Geology.* Anhwei lies in the path of the Neo-Cathaysian Geosyncline (a great downward flexure of the earth's crust), which stretches across the entire length of eastern China from Heilungkiang on the Soviet border to Kwangsi on the Gulf of Tonkin. The floor of the geosyncline is steadily sinking under the weight of the silt carried by the Huang Ho and the Huai Ho. The sediment is estimated to be more than 2,000 feet deep.

Plains and mountains. The northern portion of Anhwei Province is occupied by the North China Plain—an immense level surface that has periodically been flooded by its dominant rivers. The southern section of the province, the Yangtze Valley, is separated from the northern plain by a series of mountains that stretch roughly from west to east. The Ta-pieh Shan (Ta-pieh Mountains)—an extension of the Tsinling–Fu-niu Shan, a range lying to the north of the Yangtze—form a convex curve of steep slopes facing east and northeast on the southwestern Hupeh–Anhwei border. The Paichi Shan lie south and east of the Yangtze and form the southeastern border between Anhwei and Chekiang. Composed mainly of granite, metamorphic rock (rock formed in the solid state by heat and pressure), and sandstone, they include the Huang Shan, a range that rises to a height of 5,900 feet (1,800 metres) and is beloved by poets and artists for its massive shape and lush vegetation.

The Huai and Yangtze river basins

Drainage. The northern plain is drained by the Huai Ho. Its many left-bank tributaries—including the Ko, Fei, Ying, Hung, and Ju—rise in the western mountains and flow eastward across Honan Province into Anhwei. The shorter right-bank tributaries rise in the Ta-pieh Shan and include the Shih Ho and the P'i Ho. The Huai Ho flows across the level plain and drains into the Hung-tse Hu, a lake that lies just across the eastern border with Kiangsu Province. The river basin is subject to widespread and disastrous floods.

The watershed between the Huai Ho and the Huang Ho to the north is barely perceptible. At least three times the Huang Ho has changed its course to flow south of the Shantung Peninsula, joining its waters with the Huai to flow into the Yellow Sea.

The low line of hills (Hua-yang Shan) that extends northeast from the Ta-pieh Shan to the Hung-tse Hu marks the divide between the Huai and Yangtze river basins. Within Anhwei, the Yangtze maintains a fairly constant width of about three-quarters of a mile. The river plain is studded with lakes that, in time of flood, join the river and increase its width in places to five miles. Although the water level changes only slightly between summer and winter, winter navigation is difficult. The Yangtze plain is criss-crossed by canals that are used for irrigation, drainage, and transport.

Climate. Anhwei shares with the rest of China the seasonal monsoon climate characterized by hot, wet summers and cooler, dry winters. As it extends for about 400 miles from north to south, however, Anhwei experiences appreciable variations in climate. Mean January temperatures in the north of the Huai Basin range between 32° F (0° C) and 36° F (2° C). Mean July temperatures in both regions are 82° F (28° C). The north has between 200 and 230 frost-free days, while the south experiences between 230 and 250. The relative humidity is considerably higher in summer in the Yangtze Valley than in the Huai Basin. Northern Anhwei experiences some discomfort when winter dust storms, brought by bleak winter winds, sweep down from the Shansi Plateau.

Rainfall distribution

There is a marked difference between the rainfall of the Yangtze Basin and that of the Huai plain. In the south the precipitation amounts to between 47 and 71 inches and is relatively evenly distributed. In the north, the mean annual rainfall is between 24 and 31 inches, over half of which falls between June and September. Rainfall becomes progressively unreliable as one progresses northward until the famine lands of the northwest are reached. Summer evaporation in northern Anhwei is intense.

Soils. The province's complex soil structure can be divided broadly into three categories. The uplands to the north and south of the Yangtze are mainly podzolized (leached) old and young red earths that are susceptible to erosion but are valuable for the cultivation of tea. The Yangtze flood plains are composed chiefly of alluvium and rice paddy soils that have a slight acid character. The alluvial soil of the Huai Basin—particularly the area drained by the left-bank tributaries and extending into the North China Plain—is, however, calcareous (chalky). It includes curious mineral masses that are known as *shachiangs*, or "sand ginger," because they resemble ginger roots. They form in low-lying places where the ground is waterlogged, rarely occur on the surface, and sometimes form a hard pan, or basin, some feet below ground level.

Vegetation and animal life. The original natural vegetation was composed of forest and woodland, but centuries of intensive settlement have led to deforestation everywhere except in the western and southern uplands. This process, accompanied by soil erosion, has been very rapid since the mid-18th century. Cultivated crops and grass have replaced the trees, but fuel gatherers annually strip the hillsides of grass. The only remains of the forests on the plains are the bamboo groves and woods surrounding temples and villages. Since the 1950s, resolute efforts at afforestation have been made; conifers and poplars have been planted on hillsides and along canals and roads, and fuel gathering has been restricted.

One of the world's rare animals, the Yangtze alligator, lives in the environs of Wu-hu. It is less than six feet long and feeds mainly on fish and small animals. Apart from rodents and reptiles, very few wild mammals remain in the densely settled and cultivated plains. Deer and foxes are found in the highlands to the west of the province. Pheasants, bluejays, and golden orioles are outstanding members of the bird population.

Water conservation. Because of its susceptibility to disastrous floods, the Huai Basin in 1949 was chosen to be the site of the first large-scale water-conservation project to be undertaken by the People's Republic. The scheme began in 1951 and entailed the building of no fewer than ten dams on the upper reaches of the Huai and its tributaries in order to control the flow of floodwaters. The largest of these dams within Anhwei is located at Fo-tzu-ling on the P'i Ho. It has a hydroelectric capacity of 11,000 kilowatts a year and supplies electricity to Ho-fei. Hundreds of miles of dikes and irrigation and drainage canals along the main rivers have also been rebuilt or newly constructed. The entry to and the exit from the Hung-tse Hu have been cleared, and a new Peh Kiangsu Canal has been cut from the eastern edge of the lake near Hung-tse, Kiangsu, to the Yellow Sea, about 20 miles south of the mouth of the Hsin-huai Ho. Since 1956 there have been no serious inundations, the communes have been able to control local flooding, and irrigation has successfully compensated for threatened drought.

Population distribution

The people. *Demography.* Anhwei had an estimated population of about 35,000,000 in the late 1960s. The regions of densest population are the tributaries and banks of the Huai Ho above Pang-fou and the diked areas along the right bank of the Yangtze. In both these areas the average density exceeds 1,000 persons per square mile. Generally, the villages in the Huai Basin are somewhat larger than those in the south; they average from 50 to 100 households and are located less than one mile apart. In the Yangtze Valley, villages are farther apart, and there are more scattered homesteads. Population density is considerably less in the hilly region that separates the two river basins and averages about 250 to 500 persons per square mile. In the mountains to the west and south, the density is 125 persons per square mile, and villages are two to three miles apart.

As in the rest of China, there is no precise demarcation between rural and urban population. Places of 2,000 people and fewer that are essentially engaged in agriculture are classified as rural. On this basis, less than 9 percent of Anhwei's population was urban in 1958. There are four large towns of more than 200,000 inhabitants—Ho-fei, the capital; Huai-nan; Pang-fou; and Wu-hu. Towns of lesser importance are An-ch'ing, the former provincial capital; Ma-an-shan; and Ta-t'ung. Since 1954, many of these towns have grown rapidly because of industrialization, and new towns have also appeared.

Ethnic composition. The population is almost totally Chinese, and there are no pockets of aboriginal peoples as there are in the provinces south of the Yangtze. Until the 7th century AD there were large numbers of Hakka (Guest or Stranger People), but they were later driven south into Fukien and Kwangtung. Those of their descendants who remain have been integrated within the main body of the Chinese population.

Administration. From 1949 to 1954 Anhwei was included in the East China Administrative Area, which embraced all the east-coast provinces from Shantung to Fukien. In 1954 provincial (*sheng*) government was restored.

The pattern of local government again underwent considerable change with the formation, in 1958 and 1959, of communes, which took over all administrative functions below the county level. There were, however, considerable differences between individual communes, and the outbreak of the Cultural Revolution in 1966 brought confusion to local administration. In 1968 a 15-member Revolutionary Committee was appointed provincially and approved by the central government. Eight of its committeemen are members of the People's Liberation Army, and the remaining members are political or administrative personnel. The administration was divided into six municipalities (*shih*) and nine special districts (*chuan-ch'ü*) on the sub-province level, and into two municipalities and 70 counties (*hsien*) on the county level.

Social conditions. *Education.* Western learning was introduced in the early decades of the 20th century through the teaching of Christian missions. The national government after 1928 attempted to expand education throughout the province, but the mass of the population remained illiterate. After 1949 the problem of illiteracy was attacked with vigour. Ten years later the communes were made responsible for primary and middle-school education, and a program combining work and study was introduced. On the slopes of the Ta-pieh Shan in Chien-chai *hsien*, six schools of sericulture (the production of raw silk by the raising of silkworms) began after 1965. On the tea-growing slopes of the Ta-pien Shan there are two special schools that combine farming and study and are run by a research institute that specializes in the production of Keemun black and Tunki green tea. The students attend six months of classes during the slack season and then engage in five months of farming on the plantations.

Role of the communes in education

Health services. Medical and health services that were developed after 1949 were at first concerned primarily with public hygiene and preventive medicine. The training period for doctors was reduced to three years, and medical teams were sent into the countryside to inoculate the population and to teach and advise on matters of public health. Since 1960 a longer period of medical training is again required, and hospitals have been built in the communes and all the major towns. The emphasis on public hygiene has, however, been maintained. Rural teams are given short, intensive courses in diagnosis and treatment of such debilitating diseases as kala-azar (a protozoan infection of the bloodstream) and schistosomiasis (a parasitic infestation of the bladder or intestines) and are then set to work. Both Western and traditional medical practices are employed.

The economy. Until 1949 Anhwei was regarded as the most backward province of eastern China. Over 90 percent of its population was rural, and the standard of its agriculture was low because of poor use of water resources. Mineral resources were little developed. Since the 1950s there have been great advances in agriculture and industry.

Agriculture. Wheat is the predominant crop in the Huai Basin to the north, and—more importantly—rice is grown in the Yangtze Basin to the south. In the relatively wetter Yangtze Valley well over half of the cultivable land is devoted to rice, while in the drier Huai Basin more than 30 percent of the land is under wheat. Most of the land produces two crops a year. Summer crops include rice, sweet potatoes, sorghum, soybeans, peanuts (groundnuts), and sesame. Winter crops are wheat, barley, rape (an herb of the mustard family), peas, and green manure crops which are plowed in as fertilizer. Improvements have been made in irrigation, so that more land in the south is devoted to double cropping rice. Anhwei is also one of China's most important soybean producers; the beans are grown mainly in the north in rotation with wheat or barley. An increasing amount of kaoliang (a grain sorghum) and millet is also grown in the north.

The main industrial crops are vegetable oilseeds, cotton, tea, fibres, and tobacco. Among the vegetable oils, the most important are rapeseed, peanut, and sesameseed. Cotton is grown mainly on the northern Huai Ho plain. Hemp, jute, and ramie (an Asian nettle that yields a fibre used for making textiles) are grown. Since 1949 the acreage devoted to tobacco has been increased to more than 60,000 acres.

Tea and silk cultivation

Anhwei has been famous for its tea since the 7th century, when the "Six Rare Teas from Anhwei" were exported to the rest of China, as well as abroad. This trade became depressed in the late 19th and early 20th centuries but has now revived; Keemun black tea is especially famous. The main areas of cultivation are on the slopes of the Ta-pieh Shan, north of the Yangtze, and on the Paichi Shan along the Anhwei–Chekiang border.

Sericulture has also been revived since the 1950s. During the years of warfare from 1937 to 1949, many of the mulberry trees were felled to deprive guerrilla forces of cover. These have now been replaced, and both the mulberry-feeding moth (*Bombyx mori*) and the tussah silkworm are reared, providing the raw material for Nanking brocades and Wuhan silk fabrics.

Livestock and fish. There are more than 2,750,000 large domesticated animals that are used almost exclusively for draft purposes. About 80 percent are oxen, of which one-third are water buffalo; their numbers, however, are still inadequate for farming needs. Pigs are the main source of meat, and sheep are raised in increasing numbers in northern Anhwei. There are more than 2,750,000 acres of rivers and lakes that abound in fish, mainly carp and white bream. Fish are bred all along the Yangtze.

Industry. Thick seams of coal are widely distributed throughout the province. The main coalfield is located near Huai-nan and is the largest in southern China. There are also coal deposits in the north and near Ta-t'ung on the Yangtze.

Coal and iron deposits

Iron-ore reserves are estimated at 360,000,000 tons and have an average iron content of more than 50 percent. Anhwei is an important producer of iron and steel. The main centre is at Ma-an-shan on the Yangtze, where there is a large, integrated iron and steel plant and a dressing plant (where ore is treated to remove some of the impurities before smelting). Second in importance is T'ung-ling; Ho-fei, An-ch'ing, and Huai-nan produce smaller quantities. Ho-fei is also important for its textile manufactures and its four chemical fertilizer plants. The coal-mining centre at Ta-t'ung also produces steam and diesel locomotives. Pang-fou is the site of China's first machine-accessories plant, which supplies spare parts for Lo-yang's tractor plant and for the province's farms. Local handicrafts include the making of wrought-iron silhouettes for wall decoration.

Power. The use of both thermal and hydroelectric power has increased enormously since 1956, particularly because of the needs of the thousands of irrigation and drainage pumps in use throughout the province. An interesting experiment, which has had some success, is the collection of methane gas from the mud at the bottom of innumerable ponds; the gas is gathered in bags and used for domestic cooking and lighting.

Communications. *Water transport.* Throughout the centuries waterways have been the main means of communication. Water-conservation schemes have increased the navigability of rivers and canals, and traffic on them is heavy. Traffic on the Yangtze to the ports of Ma-an-shan, Wu-hu, T'ung-ling, Ta-t'ung, and An-ch'ing increases yearly. These ports can be reached by ocean-going vessels of 15,000 tons during the highwater summer months.

Railways. Anhwei is served by the Tientsin–P'u-k'ou railway, which enters the province from Süchow across the northern Kiangsu border. It runs south to Pang-fou, where it divides—the main line runs southeast to Nanking in Kiangsu Province, and a branch line runs south to Ho-fei, from where it runs southeastward to the Yangtze opposite Wu-hu. The value of the main line was greatly enhanced between 1960 and 1970, when it was double tracked and the Nanking bridge was built over the Yangtze. Another rail line connects Nanking and Wu-hu, from where it runs through southern Anhwei before entering Kiangsi.

Roads. Road building since 1949 has been undertaken by local labour, especially since the formation of the communes. Ho-fei stands at the centre of the province's highway system, with one main road running to Nanking in the east, another to Pang-fou in the north, and two—one north and one south of the Ta-pieh Shan—to Wu-han. A highway also connects the large towns along the right bank of the Yangtze.

Cultural life. The cultural traditions and institutions of Anhwei are the same as those found throughout China proper. The population is almost exclusively of Han, or Chinese, origin and its folklore and culture are therefore part of the greater national unit.

BIBLIOGRAPHY. SUN CHING-CHIH (ed.), *Economic Geography of East China Region* (1961; orig. pub. in Chinese, 1959), one of many volumes by Chinese geographers covering the whole country; P.M. ROXBY (ed.), *China Proper*, 3 vol. (1944), a survey made during World War II; A.G. DONNITHORNE, *China's Economic System* (1967), a standard work drawing extensively on Chinese sources; T.R. TREGEAR, *A Geography of China* (1966), a general geography to serve university undergraduates, and *Economic Geography of China* (1970), an advanced geography that emphasizes the economy and social conditions; R. HSIA, "Changes in the Location of China's Steel Industry," *China Quarterly*, 17:125–133 (1964); WU YUAN-LI, *The Steel Industry in Communist China* (1965).

(T.R.T.)

Animal Breeding

Animal breeding is the controlled propagation of domestic animals. Its aim is the improvement of qualities considered desirable by man. Breeding procedures involve the application of several basic sciences, chiefly reproductive physiology, genetics, and statistics. This article deals with the practical application of scientific principles to the selection of superior animals and the planning of mating combinations. The fundamental biological principles underlying animal breeding are discussed in the articles REPRODUCTION; HEREDITY; and GENETICS.

Uses of highly bred animals

Animals are bred for utility, sport, pleasure, and research. Dogs, for example, have been bred to serve as watchdogs, police dogs, hunters, sheep dogs, and pets. Many species of small mammals, especially rats, mice, rabbits, and guinea pigs, are bred for research, chiefly in genetics, physiology, and medicine. The basic principles of breeding are the same no matter what the animal species or the purpose in breeding may be, but the practical approach to the problem may differ in several respects, depending on such considerations as the mode and rate of reproduction and the relative effects of genetic and environmental factors on the traits of greatest interest.

The term population is used in this article to denote a group of interbreeding individuals; *i.e.*, a breed, or strain within a breed, which in some respects is genetically different from other breeds, or strains, of the same species. The word purebred is used here in its sense of referring to animals registered in the herdbook maintained for a certain breed, or to animals eligible for such registration, and the mating of purebred animals is called pure breeding. It is to be understood that genetically pure breeds (homozygous for all traits) do not exist.

History. It is generally agreed by authorities in the field that the dog was the first animal to be domesticated, probably more than 12,000 years ago. The domestication of horses, cattle, sheep, goats, and pigs took place sometime later, during the period 6000 to 2000 BC in India, the Middle East, and northeast Africa. Somewhat later cattle and pigs apparently were domesticated also in Europe. Rabbits were domesticated in southwestern Europe in the early Middle Ages. The fowl seems to have been domesticated for the first time in India about 2000 BC; ducks and geese apparently were domesticated in southern Europe at about the same time, and turkeys in Central and North America about 1500 BC. Foxes, mink, and some other species are still undergoing domestication at the present time (see DOMESTICATION, PLANT AND ANIMAL).

As long as 5,000 years ago, man developed breeds of dogs, probably by selecting animals showing desirable traits for the purpose at hand. The domestic pigeon also exhibits great diversity of breeds, some of which have existed for several thousands of years. Thus, for a very long time man has been interested in the breeding of animals not only for utility but also for the preservation of highly deviant types that could not remain in existence without man's protection from the hardships of the environment. There are, in fact, several indications that breeding for utility started considerably later than breeding for pleasure and fancy.

Introduction of planned matings

The formation and improvement of utility breeds of the larger mammals, based on selection and planned matings, was initiated in Great Britain in the middle of the 18th century. Before that time local breeds arose and became adapted to their areas of distribution mainly by natural selection. In 18th-century England, however, it became the practice to judge animals on the basis of their physical structure (conformation) and various other qualities, and to use inbreeding to consolidate desired

traits in strains and breeds. In many cases, however, inbreeding was carried too far and led to a reduction in fertility and viability.

A new era in animal breeding started about 1930 chiefly as a result of three factors: (1) progress in population genetics and application of its principles to animal improvement; (2) development and use of artificial insemination, especially in dairy cattle; and (3) general economic trends, which necessitated efficient conversion of feeds into animal products of high quality. Especially with regard to economic factors, fancy points used in judging and strict standardization of breed type became rather unimportant with regard to utility breeds.

Goals. The objectives of animal breeding vary with regard to species, local conditions, and time. Early in history horses were bred mainly for riding or loading purposes; later they were bred for traction; and nowadays, to a large extent, for sport (racing and hunting). In North America and western Europe, cattle populations are specialized for beef or milk production, or bred for a combination of both. In southern Europe and in many parts of Africa and Asia, oxen are still produced for pulling plows or carts. Some breeds of sheep are specialized for wool production, some for meat, and one breed, the Karakul sheep, is bred for fur production (Persian lamb). Pigs always are bred for meat production, but they may be specialized to produce a certain type of meat, either pork or bacon. At one time, chickens were bred for the combined production of eggs and meat, but in the Western world there is now a pronounced specialization of breeds and crosses to produce either eggs or meat.

EVALUATION OF ANIMALS

In breeding farm animals for utility, selection must be based as far as possible on objective measurements of traits that are decisive for the economy of production.

Judging animals on appearance

Judging animals on the basis of appearance alone has not become obsolete but its importance has been reduced. Certain traits, of course, are difficult or impossible to measure objectively. In all kinds of horses, for example, the legs should be well proportioned in relation to the body and free from faults and weaknesses; this judgment is difficult to render quantitatively. Similarly, the strength of the legs is important in cattle and pigs, but difficult to measure. In dairy cows, the attachment of the udder to the body is difficult to measure; but it is significant, and it can be judged and scored. In mink breeding, a judging of the animals with regard to fur quality and shade of colour is necessary and cannot, so far, be substituted by laboratory tests. In the breeding of pet animals, such as dogs and cats, judging the animals for conformity with breed standards is usually decisive in determining their market value.

Actual measurement of an individual animal's performance is a fairly recent innovation in animal breeding, except with regard to Thoroughbred horses, the selection of which for centuries has been based on speed at standard racing distances. Systematic recording of milk production in dairy cows started in Denmark in 1895, and the movement spread rapidly in northwestern Europe and North America. In several countries, young beef and dairy bulls are tested at special stations for rate of growth and muscle development. In sheep breeding, the number of lambs borne by each ewe is recorded, as is the weight of the lambs at five months of age. The recording of growth rate, feed consumption, and carcass quality of pigs began in 1907. Progeny testing and performance testing (see below *Methods for estimating breeding values*) has subsequently been added to judge young males (bulls, rams, and boars) intended for breeding. The productivity of sows is measured by recording the number of pigs in each litter at birth and their weight at three weeks of age. An ultrasonic technique measures the thickness of backfat on live pigs. Special nests with hinged doors (trapnests) were employed to record the egg production of individual hens from the end of the 19th century until the introduction of individual laying cages made trapnests unnecessary. The recording of individual performance in breeding populations of farm animals developed with great rapidity after World War II, in terms of both frequency of use and refinement of methods. It may be considered to be the very foundation of progress in breeding programs.

Judging animals on performance

Genetic and environmental variation in animal traits. The traits of animals in any population may be classified roughly into two groups; viz., qualitative and quantitative traits. Qualitative traits show discontinuous variation; for example, coat colour, presence or absence of horns, certain blood characteristics (*e.g.*, blood types), presence or absence of particular enzymes, and existence of metabolic defects and congenital malformations (*e.g.*, bleeding disease [hemophilia] in dogs). In general, the inheritance of qualitative traits is relatively simple, in accordance with the laws of heredity, and environment plays only a minor role in their variation.

The quantitative traits show continuous variation between the extreme variants, the mean type being, as a rule, most frequent. Growth rate, live weight, body measurements at mature age, milk yield and composition in cows; body length and backfat thickness in pigs; wool yield and quality in sheep; and egg production in fowl are typical examples of quantitative traits. As a rule, such traits are influenced by many genes (polygenes), each gene exerting relatively small effect, and environmental factors are responsible for a considerable part of the variation.

There is, in fact, no distinct borderline between qualitative and quantitative traits. A congenital malformation, for example, may be determined by a major gene, as well as a number of polygenes, and the latter may cause a considerable variation in the expression of the trait. Similarly, variation in the expression of a trait regulated by a single gene may be caused by environmental factors. Nevertheless, this classification is useful because typical qualitative traits can be analyzed with regard to a single gene and its behaviour; whereas quantitative traits are best studied by statistical methods that permit one to proceed without knowledge of the number of genes involved or their interrelations. By using appropriate statistical methods, for example, it is possible to estimate the fraction of the total variation in the population that is caused by the additive effect of the genes; this fraction is termed the heritability. A heritability of 1.00 indicates that all the variation observed in the population of the trait in question is genetically determined; a heritability of 0.00 indicates that the variation is wholly environmental in cause; and when the heritability is 0.50 the genetic and environmental factors are equally responsible for the observed variation of the trait.

Heritability of traits

A few estimates of heritability of traits are given to illustrate typical values. In horses: pulling power, 0.25. In dairy cows: height at withers at three years of age, 0.45; milk yield at first lactation, 0.30; fat and protein content of milk, 0.55; resistance to mastitis, 0.30. In beef cattle; daily weight gain (from weaning to slaughter), 0.40. In sheep (Merinos): yield of clean wool, 0.45. In pigs: daily weight gain (from weaning to slaughter) in group feeding, 0.30, in individual feeding, 0.60; backfat thickness, 0.50; resistance to atrophic rhinitis, 0.25. In chickens: egg production (in first laying year), 0.30; egg size, 0.50; resistance to leucosis, 0.10. In general, heritability is relatively low for such traits as fertility and resistance to infectious diseases, and it is high for growth rate, body size at mature age, and composition of cows' milk.

The magnitude of environmental effects, and to some extent also the additive genetic effects, may vary considerably between different populations of the same species; therefore, heritability estimates also vary and cannot be considered as constants for the various traits. For example, as noted above, when the heritability of growth rate of pigs is estimated from individual-feeding data, it is twice as high as when it is estimated from group-feeding data. In spite of these limitations, heritability estimates for quantitative traits have proved to be of great value in planning animal improvement programs.

Methods for estimating breeding values. The breeding value of an animal depends on the genes it passes on to

its offspring. Each individual offspring receives a random sample of one-half of the genes from each of the parents.

Concerning qualitative traits with known inheritance patterns, the situation is relatively simple. A dominant gene manifests its presence in the physical character (phenotype) of the individual even when the gene it is paired with is different—that is, in the heterozygous condition. A recessive gene also can be recognized but only when it is present in duplicate (the homozygous state). In both cases, of course, there must be no complications caused by modifying polygenes or environmental factors. The breeder can then select for or against the trait according to his wishes. To recognize heterozygous carriers of recessive genes by visual inspection of the animals is impossible in most cases, but it can be done by mating the suspected carrier to a certain number of known heterozygotes or to its own offspring.

Breeding values

The breeding value of a quantitative trait may be assessed on the basis of the merits of: (1) the individual himself, (2) his ancestors, (3) his collateral relatives, (4) his progeny, or (5) a combination of any two or more of the above. The relative value of these different approaches depends on the heritability of the trait and the rate of reproduction. With regard to quantitative traits that are sensitive to environmental influence, it is often advisable to use the deviation of the individual's record from the mean record of the herd, rather than the individual's record itself. This procedure eliminates, or at least reduces, the effect of nongenetic differences between herds. Each of the above methods of assessing quantitative traits has its own uses (as discussed below).

Individual merit. If it is assumed, for example, that a cow in her first lactation has produced 4,500 kilograms of milk, and that her contemporary herd mates in the corresponding lactation have produced 4,000 kilograms on average, then the phenotypic merit of this cow is 500 kilograms above that of her contemporaries. By this procedure, environmental effect is minimized. When cows of different age (or undergoing different lactation periods) are compared, it is necessary to correct the yield of each individual to a standard age, because the yield increases until the fourth or fifth lactation. The fat content of the milk is much less influenced by environment and by individual age; therefore, no age corrections are needed, and individuals can be compared directly on their actual records. The same principles apply to other traits and other species of animals.

Ancestor merit. The merit of ancestors is usually the first available information on the breeding value of an individual, and such pedigree information, therefore, is valuable as a rule. With each earlier generation in the pedigree, however, the value of this information is halved. Furthermore, since a grandparent can pass his genes to a grandson or a granddaughter only through one of its parents, the more information known about this parent, the less valuable is the information known about the grandparent. For example, if a reliable progeny test (see below) shows that the sire of an animal has a high breeding value, there is no need to consider the parents of that sire. In many cases, especially in horse and dog breeding, the importance of long pedigrees has been greatly exaggerated.

Merit of collateral relatives. The genetic similarity between an individual and a randomly chosen full sib (brother or sister) is the same, on average, as that between the individual and one of his parents; and an individual's genetic similarity with a half-sib is the same as that with one of his grandparents. An individual, however, can have many more full sibs and half-sibs than parents or grandparents, and the sibs, therefore, may be of much greater value than parents or grandparents for estimating the breeding value of the individual. In pigs, rabbits, dogs, and fowl, the number of full sibs can be fairly large, and in artificial insemination of cattle the number of half-sibs likewise can be large.

Judging by sibs

Progeny merit. Progeny tests yield the final information on an animal's breeding value. The relative importance of such tests increases with decreasing heritability of the trait, and it is especially valuable for sex-limited traits, as, for example, in testing a bull's breeding value with respect to the milk yield of his daughters. The lower the heritability of a given trait, the larger a progeny group is needed for a reliable test of the individual.

Combined methods. Information from two or more of these methods can be combined into a single estimate of the individual's breeding value. In such circumstances, the different criteria should be weighted according to their expected contribution to the accuracy of the final estimate.

THE BREEDING PROGRAM

The genetic improvement of a herd or a breed requires careful planning with regard to the choice of animals for breeding and the mating combinations that are carried out.

Methods of selection. Selection of breeding animals can be carried out in different ways. Among the more important are mass selection, pedigree selection, family selection, and progeny selection.

1. Mass selection is based solely on individual merit. Applied to traits with high heritability and about equal manifestation in both sexes, mass selection can be expected to give good results. With decreasing heritability the efficiency decreases, and for sex-limited traits (and heritability considerably below 0.5) it is always inefficient.

2. Pedigree selection depends on the merits of the ancestors. It is valuable in the first selection among young animals, especially when the heritability of the traits is high. Relying solely on pedigree selection, however, results in very slow progress.

3. Family selection is based on the merits of collateral relatives, such as full sibs or half-sibs, and it is used mainly as an aid to individual selection. It is especially valuable with regard to sex-limited traits, for traits with low heritability, or when some animals have to be slaughtered, as for determining the carcass quality. In the selection of young males for breeding, for example, no data may be available on their individual performance; *e.g.*, egg production in the fowl. When sib groups of pullets start laying early in the autumn, the cockerels may be selected for breeding on the laying records of their full sibs and half-sibs. Similarly, young bulls may be selected mainly on the milk records of their paternal half-sibs; that is, on the progeny tests of their sires. Individual pigs may be selected on the basis of carcass tests made on their sib groups. With regard to traits that can be recorded for all the animals alike, males and females, selection of individuals can be based simply on their family average or on an index that combines the individual's own performance and the average for the rest of the family. A simple procedure, applicable in dog and pig breeding, would be to select the best individual from the best litter.

One difficulty in the application of family selection is that systematic environmental differences may occur, especially between full-sib groups, and these tend to mask the genetic differences. When inherited defects appear in sib groups, there is a certain risk that some of the healthy animals carry a hidden gene for the defect.

4. Progeny selection has been applied with great success in dairy cattle breeding, and in general it is valuable in all types of livestock when applied to sex-limited traits and traits with low heritability. Early progeny testing of males on a sufficient number of offspring and an effective selection among those tested are very important. The disadvantage of selection of sires on progeny testing is that it means increased length of the generation interval and thereby tends to slow down the rate of genetic improvement.

In general, it is not necessary to include all the traits used for selection in an overall index; for example, all animals used for breeding should possess normal fertility, and those that do not should be excluded from the breeding program. Also, as a rule, any animal known to be a carrier of a gene for a serious metabolic or morphological defect should be eliminated even if the merits for some other traits are fairly high.

Stepwise selection

Usually selection is made in a stepwise fashion. With regard to dairy bulls, for example, selection on the basis of pedigree can be made soon after birth; a second selection can be made later based on growth rate during the first year of life and fertility in the first series of inseminations; and finally a third selection can depend on the results of progeny testing, when offspring are old enough to be judged.

Mating systems. *Random mating.* Random mating implies that each possible mating in a population has the same probability of occurrence. Only artificial selection by the breeder can really be eliminated; however, a certain amount of natural selection always takes place. Random mating often is used in breeding experiments to minimize genetic changes in a control population with which selected populations are compared.

Inbreeding. Inbreeding may be defined as mating of individuals more closely related than the average of the population. It increases the homozygosity and decreases the heterozygosity of the inbred animals. The so-called inbreeding coefficient is a measure of the loss of heterozygosity due to inbreeding, and it is expressed as a fraction, or percentage, of the amount of heterozygosity present when the inbreeding started. After one generation of mating between full sibs or mating of sire with daughter or dam with son, the heterozygosity of the offspring is reduced by 25% (or the inbreeding coefficient is 25%). In the mating of half-sibs or double first cousins the inbreeding coefficient is 12.5%. Mating of single first cousins gives an inbreeding coefficient of 6.25%, and that of half first cousins 3.12%. Mating between full sibs in two successive generations decreases the heterozygosity by 37.5%, and in three generations by 50%. So-called inbred lines are produced by continuous consanguineous matings in several generations; the genetic variation decreases within each line and increases between separate lines. In experiments with mice, rats, and guinea pigs, full-sib matings have been continued through large numbers of generations. Farm animals and birds are much more sensitive to inbreeding, and usually full-sib matings can be continued for a few generations only because of a marked decrease in viability and fertility.

Breeding within small populations, such as a herd or flock, without infusion of new animals from outside, leads automatically to a certain amount of inbreeding. In farm mammals, each male is invariably used to serve a large number of females, and under such conditions the increase of the inbreeding coefficient per generation in a closed breeding unit can be estimated simply by dividing 100 by eight times the number of males used. Thus, when only one homebred sire is used, the decrease in heterozygosity in each generation is 12.5%. When artificial insemination is used, a small number of sires again are used to serve a large number of females. Nevertheless, the reduction in heterozygosity may be relatively small in such cases, because many young sires are used for progeny testing, and several tested sires enter into service each year. The modern trend in artificial insemination is to use the tested sires for only relatively short periods of time, after which they are replaced with younger sires, which presumably have made even better records. Such rapid turnover in sires serves also to reduce the length of the generation interval in breeding programs.

Inbreeding increases the homozygosity of unfavourable as well as favourable genes. As a result there is a segregation of various kinds of congenital defects and, more important, a general decline in fertility and viability of the inbred animals. The latter finding has been demonstrated in numerous experiments with farm animals, especially pigs and poultry. In experiments with laboratory animals it has been shown also that the sensitivity to unfavourable environmental influences increases with inbreeding. Although the decrease in fitness resulting from inbreeding is a general phenomenon, the amount of decrease is dependent on the genetic constitution of the animals used for inbreeding.

Linebreeding is a form of mild inbreeding designed to concentrate the genes of a certain ancestor in a strain of animals. The most intensive form of linebreeding is repeated backcrossing to a particular parent, but usually a more distant relationship is preferred; for example, a female may be mated to her grandsire or uncle.

Outbreeding. Outbreeding is defined as mating individuals less closely related than the average of the population. The degree of outbreeding can vary, just as that of inbreeding can. In some cases, it is even possible to make crosses between species, as in the crossing of the horse and ass for the production of mules. The term crossbreeding, however, usually refers to crosses between breeds within the same species. Crosses often are made between more or less inbred strains or lines of the same breed (called, respectively, strain and line crosses).

Crossbreeding

Crossbreeding has been practiced for a long time, and it also has been subjected to experimental research in the United States and Europe. Various methods have been developed and tested, with generally favourable results. The main function of crossbreeding is to increase the heterozygosity of the offspring. One of the breeds selected may be superior in certain traits and the other breed may excel in other traits. It can be expected, then, that the first-generation crossbred animals will be about intermediate to the parental breeds with respect to both traits. In some cases, however, the first-generation animals are somewhat superior to the better parental breed with regard to total merit. When the average quality of the first generation exceeds the average of the two parental breeds, the phenomenon is called heterosis, or hybrid vigour. Heterosis is displayed mainly in the so-called fitness traits, fertility and viability. It is thought to result from interaction of different forms (alleles) of a given gene (a phenomenon called overdominance) or from interactions of quite separate genes (epistasis). Crossbreeding thus is a way of utilizing nonadditive gene effects that cannot be exploited by selection within the separate breeds. When the heterozygosity is increased, the number of different genes is increased, a result that probably makes the animals better able to adapt to environmental stress.

Backcrosses (crosses of crossbred offspring to one of the parental breeds) and successive three-breed crosses seem to have an advantage when maternal influence on the offspring is important; crossbred sows, for example, generally take better care of their young than purebred sows do. Rotational crossbreeding, usually involving three breeds, is a favoured method in commercial pig breeding since it necessitates purchase of males only. The first cross is made between breeds A and B, for instance, and the female offspring are mated to a boar of breed C; the females resulting from this cross are then mated to a boar of breed A, and in the next generation a boar of breed B is used, and so on.

Rotational crossbreeding

Many experiments have been carried out on the crossing of inbred lines of the same breed or of different breeds. Generally, fertility and viability are restored even in the first crossbred generation. It has been possible with this technique, for example, to produce commercial "hybrid chicks" with superior egg-laying performance. Because the individual inbred lines are poor producers, the hybrid chicks are usually developed by a four-way cross. This is carried out by mating the offspring from crossing of lines A and B with the offspring from crosses between lines C and D, producing in effect a "double" hybrid. In order to obtain the best possible result, a large number of lines are tested in various crosses for combining ability —that is, ability to produce desirable results from crossing. Most of the eggs marketed in North America and western Europe are produced by hybrid chicks.

Similar methods of breeding have been tried also on pigs and cattle, but the results have been less favourable. With these animals the market value of the individual is so high that it is impossible for breeders to keep the costs of developing and testing inbred lines within reasonable limits.

Artificial insemination and egg transplantation. The practical application of artificial insemination in horses, cattle, sheep, and pigs was developed in Russia during the first decades of the 20th century. Semen was collected in

an artificial vagina when the male mounted a female, or a dummy, and methods were developed for dilution of semen and its preservation for several days outside the body. More rapid progress became possible in 1950, when it was shown that bull semen could be deep-frozen at −79° C with solid carbon dioxide and later thawed without serious effect on fertility, provided that a certain amount of glycerol was added before freezing. Later, the use of liquid nitrogen made it possible to store semen at about −196° C. Calves have been produced from semen that has been frozen for more than ten years, and it seems possible to increase the period of storage indefinitely. Theoretically, it should be possible to produce more than 100,000 calves per bull annually; 10,000 or more actually have been produced. In Denmark about 95% of the dairy cows are artificially inseminated; in England and Wales, about 70%; and in the United States, about 50%.

Preservation of semen

The greatest advantage of artificial insemination of dairy cattle from a genetic point of view is that the bulls can be progeny tested with much greater accuracy than in natural breeding, not only because the number of daughters is greater but also because the daughters are spread over many herds with different environmental conditions. A progeny test carried out in any one herd is valid for the conditions in that herd only. Furthermore, by artificial insemination the progeny test can be completed at least one year earlier, on average, than if the bulls were used in natural service.

Artificial insemination is practiced also on beef cattle, horses, sheep, pigs, and poultry, although on a small scale compared to its use with dairy cattle. The application of artificial insemination in pig breeding was delayed by difficulties in deep-freezing boar semen, but these difficulties appear to have been overcome.

Polyovulation

By hormone treatment, production of many ova (polyovulation) can be induced in females. After artificial insemination, the fertilized eggs can be collected from the oviducts or uterus of the donor and transferred to the uteri of recipient females for development into normal fetuses. At the time of transfer, the recipient must be synchronized with the donor with regard to the sexual cycle. Many successful transplantations of fertilized eggs have been made in sheep, and some have been carried out in cattle and pigs. Theoretically, it is possible to harvest thousands of eggs from an individual female within a few years' time and to distribute these eggs to a large number of recipients, thereby multiplying the possibilities for propagation of superior females. To secure the eggs from the donor, however, and transfer them to other females is laborious and costly, even if it can be done without major surgery. Methods of freezing and storing eggs without harmful effects must be developed before the transplantation technique can be applied as a routine procedure in the breeding of farm animals.

The rate of genetic improvement. The rate at which genetic improvement can be carried out within a certain population depends on three factors:

1. The accuracy in estimating the breeding value of the individuals. This accuracy itself depends on the heritability of the trait or traits on which the selection is based. The value used may be based on the animal's own performance, as, for example, the milk yield of a cow, or it may be an index formed by a combination of information on breeding value derived from various sources such as pedigree, collateral relatives, and progeny.
2. The selection differential, which is defined as the difference between the index of the animals selected for breeding and the mean of the entire population or group to which the selected animals belong.
3. The length of the generation interval, defined as the average interval (in years) between the birth of the parents and the birth of their offspring used for breeding. On average, this interval for horses is about nine years, for cattle five, for sheep three and a half, for pigs two and a half, and for chickens one and a half. It is possible, however, to reduce these intervals quite considerably by planned breeding.

The overall rate of genetic improvement, or response to selection, is equal to the accuracy in selection multiplied by the selection differential and divided by the generation interval. In equation form this is expressed as: $Re = HS/Y$, in which Re is the response to selection, H is the heritability, S the selection differential, and Y the generation interval.

Response to selection

From this formula, it is evident that the rate of genetic improvement of the population can be increased by increasing the accuracy in choice of breeding animals, by increasing the selection differential, and/or by decreasing the length of the generation interval. The possibility of increasing the selection differential depends on the rate of reproduction. Pigs have a much higher reproductive rate than cattle, and therefore fewer animals are needed for breeding and a greater selection differential is possible. In dairy cattle it may be necessary to raise about 60% of all heifer calves for breeding, merely to keep the size of the herd constant. For bulls, however, the selection differential can be very high, especially when artificial insemination is used. The size of the population subjected to selection is also important. It has been calculated, for example, that when the number of performance-tested dairy cows in an artificial insemination unit increases from 2,000 to 20,000, it should be possible to increase the rate of genetic improvement for milk yield about 50% by means of efficient progeny testing and selection.

In most Western countries there was a pronounced improvement in many economically important traits of farm animals after World War II. In the United States, for example, the average yield of milk from milk-recorded cows increased from 4,275 kilograms in 1955 to 5,538 kilograms in 1967. This represents an increase of 29.5%, or 2.27% per year. For the Swedish red and white breed, the average milk yield increased from 4,297 kilograms in 1960 to 5,032 kilograms in 1969; *i.e.*, by 17.1%, or 1.71% per year. It has been estimated that the genetic improvement of this breed during the same period corresponded to about 1% per year, or more than half the actual rise in milk yield. It seems probable that the major part of the genetic improvement in this instance was due to a more efficient progeny-testing and selection program that had been started among the tested bulls in the mid-1950s. Similarly, in Danish pig-testing stations, the average daily gain in live weight increased from 678 to 697 grams from 1955 to 1962, the length of carcasses increased from 93.8 to 95.9 centimetres, and the backfat thickness decreased from 3.26 to 2.80 centimetres. The modern broiler chick is an example of the success obtained by crosses between breeds that have been specialized for different lines of production without close inbreeding.

Limitation to selection

When there is a considerable amount of hereditary variation, it is possible to change a breeding population quite considerably in about 5–10 generations of intense selection. Sooner or later, however, the response to selection decreases, and ultimately a selection limit is reached. This may be due to an exhaustion of the genetic variation or, more likely, to a disturbed gene balance, especially if the selection is concentrated on one, or a very few, traits only, without considering the fertility and viability of the animals. The remedy in such a situation is either a cautious introduction of new genes from another population or deliberate crossbreeding to increase the genetic variation. In either case, selection in the gene-enriched population can start again. Another possibility is to relax the selection for a number of generations, giving the population time to recover, but this is a rather time-consuming process.

COOPERATIVE AND GOVERNMENTAL MEASURES FOR THE PROMOTION OF ANIMAL BREEDING

In most countries in which animal breeding has reached a fairly high level, associations were formed long ago (beginning in Great Britain in the late 18th century) with the aim of promoting the development of existing breeds. At first the primary objectives of breeders' associations were to publish herdbooks and to lay down rules for registration of "purebred" animals, to arrange shows and

fairs, and to work for the dissemination of the breeds at home and abroad. Later, these associations started performance testing or worked in close cooperation with other organizations developed for this purpose. In European cattle breeding there is a trend toward concentration into one organization of the various activities, such as performance and progeny testing, artificial insemination, and other services, among which herdbook registration may be only a minor detail. A similar trend has appeared also in the breeding of sheep, pigs, and fur animals. With regard to the breeding of animals for sport or as pets, traditional breeders' associations are, and probably will continue to be, of great importance.

Breeders' associations

Scientific research has been the foundation, and education the impelling force, in the accelerated development of animal breeding in the 20th century. Specialized research institutes and experimental stations, with cooperating agricultural colleges and extension divisions, have arisen in almost every country in which animal production is an important enterprise. Most information based on laboratory research and breeding experiments has arisen in the United States and Great Britain, but Denmark seems to have been the first country to organize an agricultural extension service that reached all levels of farming and animal husbandry.

Organizations have been established in several countries to promote animal science and production; for example, in the United States, the American Society of Animal Science and the American Dairy Science Association; in Great Britain, the British Society of Animal Production; and in Germany, Deutsche Gesellschaft für Züchtungskunde. In 1966 these national organizations formed the World Association for Animal Production, the major objective of which is to arrange periodic world conferences for the exchange and dissemination of knowledge in the field of animal science.

BIBLIOGRAPHY

Books: J.L. LUSH, *Animal Breeding Plans*, 3rd ed. (1945), the classic work on the principles of animal breeding; V.A. RICE *et al.*, *Breeding and Improvement of Farm Animals*, 6th ed. (1967); IVAR JOHANSSON and JAN RENDEL, *Ärftlighet och husdjursförädling* (1963; Eng. trans., *Genetics and Animal Breeding*, 1968); F.B. HUTT, *Animal Genetics* (1964); FRANZ PIRCHNER, *Populationsgenetik in der Tierzucht* (1964; Eng. trans., *Population Genetics in Animal Breeding*, 1969); M. BURNS and M.N. FRASER, *Genetics of the Dog: The Basis of Successful Breeding*, 2nd ed. (1966); E.J. PERRY (ed.), *The Artificial Insemination of Farm Animals*, 4th rev. ed. (1968).

Review articles: H.T. FREDEEN, "Inbreeding and Swine Improvement," *Anim. Breed. Abstr.*, 24:317–326 (1956), "Genetic Aspects of Disease Resistance," *ibid.*, 33:17–26 (1965); IVAR JOHANSSON, "Hereditary Defects in Farm Animals," *Wld. Rev. Anim. Prod.*, 3:19–29 (1965); BENGT LINDHE, "Breeding and Crossbreeding for Beef Production," *ibid.*, 5:51–57 (1969); J.E. LEGATES, "Crossbreeding in Cattle," *ibid.*, 3:69–74 (1966).

(I.J.)

Animal Feed

Animals in general require the same nutrients as man. (This article is concerned primarily with feed for livestock and poultry. Information on pet food will be found in the article PETS.) Some feeds, such as pasture grasses, hay and silage crops, and certain cereal grains, are grown specifically for animals. Other feeds, such as sugar-beet pulp, brewers' grains, and pineapple bran, are by-products remaining after a food crop has been processed for human use. Surplus food crops, such as wheat, other cereals, fruits, vegetables, and roots, may also be fed to animals. In this way such surpluses are converted into meat, milk, and eggs for the human diet.

History does not record when dried roughage or other stored feeds were first given to animals. Most early records refer to nomadic people who, with their herds and flocks, followed the natural feed supplies. When animals were domesticated and used for work in crop production, some of the residues were doubtless fed to them.

Preservation of green forages such as beet leaves and corn plants by packing them in pits in the earth has long been practiced in northern Europe. The idea of making silage as a means of preserving and utilizing more of the corn plant was gradually developed in Europe and brought from France to the United States in the 1870s. When the mature, dried corn plant was fed to cattle in the winter much of the coarse stem was wasted, but when it was chopped and ensiled (made into silage) everything was eaten. The first effort to evaluate feeds for animals on a comparative basis was apparently made by Albrecht Thaer (1752–1828), in Germany, who developed "hay values" as measures of nutritive value of feeds. Tables of the value of feeds and of the requirements of animals in Germany followed and were later used in other countries. Present-day knowledge represents an expansion and improvement of these early efforts.

BASIC NUTRIENTS AND ADDITIVES

The basic nutrients that animals require for growth, reproduction, and good health include carbohydrates, protein, fat, minerals, and vitamins. The energy needed for growth and activity is derived primarily from carbohydrates and fats. Protein may also supply energy, however, if other sources are not adequate or if it is supplied in great excess above the needs of the body.

Animals need a source of energy to sustain life processes within the body and for muscular activity. When the energy intake of an animal exceeds its requirements, the surplus is stored as body fat, which can be utilized later if the energy consumed in food is not adequate.

Proteins. A small amount of protein is needed for the daily repair of muscles, internal organs, and other body tissues. Immature animals also require protein for growth of the muscles and other parts of the body. Since milk, eggs, and wool contain much protein, additional amounts are needed in the food of animals producing these.

Proteins are composed of more than 20 different amino acids, which are liberated during digestion. Animals with simple stomachs, including humans, monkeys, swine, poultry, rabbits, and mink, require correct amounts of the following ten essential amino acids daily: arginine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, trytophan, and valine. In addition to these, poultry need glycine and glutamic acid for growth. High-quality protein supplied by eggs, milk, fish meal, meat by-products, and soybean meal contains correct amounts of the essential amino acids. Poor-quality protein, such as that in corn grain (maize), contains too little of one or more essential amino acids. Feeds having poor-quality proteins are made useful by blending them with other feeds that supply the lacking amino acids.

The quality of the protein in the food is of little importance to ruminants, including cattle, sheep, goats, and the other animals that have four stomachs, because the bacteria that aid in the digestion of food in the rumen (first stomach) use simple nitrogen compounds to build proteins in their cells. Further on in the digestive tract the animals digest the bacteria. By this indirect means, ruminants produce high-quality protein from a food that might originally have contained poor protein, or from urea (a nitrogen compound). Very young ruminants, such as calves, lambs, and kids, however, need good-quality protein until the rumen develops sufficiently for this bacterial process to become established.

Role of ruminants

Carbohydrates and fats. Most animals get energy from carbohydrates and fats, which are oxidized in the body. These yield heat, which maintains body temperature, furnishes energy for growth and muscle activity, and sustains vital functions. Animals need much more energy (and more total feed) for growth, fattening, work, or milk production than simply for maintenance.

The less complex carbohydrates (sugars and starches) are readily digested by all animals. The complex carbohydrates (cellulose, hemicelluloses) that make up the fibrous stems of plants are broken down by bacterial action in the rumen of cattle and sheep or in the cecum of rabbits and horses. Such complex carbohydrates cannot be digested by men, or, to any appreciable extent, by dogs, cats, birds, or laboratory animals. Thus, ruminants and some herbivorous animals obtain much more of the

energy-giving nutrients from the carbohydrates of plants than the simple-stomached carnivores and omnivores, for which fibrous materials have little or no energy value.

Fat in feeds has a high nutritive value because it is highly digestible, and because it supplies about two and one-quarter times as much energy as starch or sugar. While fat has high nutritive value, it can be replaced by an equivalent amount of digestible carbohydrates in the food, except for small amounts of essential fatty acids. Very small amounts of the unsaturated fatty acid linoleic, contained in some fats, are necessary for growth and health. Usual animal feeds, however, supply ample amounts of this acid unless it has been removed by processing.

Minerals. Certain minerals are essential for animal life, among them being common salt (sodium chloride), calcium, phosphorus, sulfur, potassium, magnesium, manganese, iron, copper, cobalt, iodine, zinc, molybdenum, and selenium. The last six of these are poisonous to animals if excessive amounts are eaten.

All farm animals generally need more common salt than is contained in their feeds, and they are supplied with it regularly. Of the other essential minerals, phosphorus and calcium are most apt to be lacking, because they are heavily drawn upon to produce bones, milk, and egg shells. Good phosphorus supplements are bone meal, dicalcium phosphate, and defluorinated phosphates. Egg shells are nearly pure calcium carbonate. Calcium may readily be supplied by ground limestone, ground shells, or marl that is high in calcium.

Small amounts of iodine are needed by animals for the formation of thyroxine, a compound containing iodine, secreted by the thyroid gland. A serious deficiency of iodine may cause goitre, a disease in which the thyroid gland enlarges greatly. In certain regions goitre has caused heavy losses of newborn pigs, lambs, kids, calves, and foals. Goitre can be prevented by supplying small amounts of iodized salt to the mother before the young are born.

Losses caused by goitre

In some areas, soil and forage are deficient in copper and cobalt, which are needed along with iron for the formation of hemoglobin, the oxygen-carrying pigment of the red blood cells. In these areas, farm animals may suffer from anemia unless the deficiency is corrected by means of a suitable mineral supplement.

Iron, used in hemoglobin formation, is amply supplied in most animal feeds, except milk. The only practical problem with iron deficiency occurs in young suckling pigs before they start to consume other feeds in addition to milk.

Though manganese is essential for animals, the usual rations for all farm animals, except poultry, supply sufficient quantities. A lack of manganese may cause the nutritional disease of chicks and young turkeys called slipped tendon (perosis) and also may cause failure of eggs to hatch. Normal rations for swine are often deficient in zinc, especially in the presence of excess calcium. Adding 100 parts per million of zinc carbonate cures zinc-deficiency symptoms, which include retarded growth rate and severe scaliness and cracking of the skin (parakeratosis). A trace of selenium is necessary for normal health of animals; excessive amounts found in forages in some regions poison animals and may cause death. To furnish both calcium and phosphorus, livestock may be allowed free access to such a mixture as 60 percent dicalcium phosphate and 40 percent common salt. Trace mineralized salt is used when copper or cobalt may be deficient. Animals are given access to common salt separately, so they will not be forced to eat more of the other minerals than they require to get the amount of salt they need.

Vitamins. Known vitamins include A, C, D, E, and K, and the B group: thiamine, riboflavin, niacin, pantothenic acid, choline, biotin, folic acid, and vitamins B_6 and B_{12}.

Vitamin A, the one most apt to be lacking in livestock feeds, is required for growth, reproduction, milk production, and the maintenance of normal resistance to respiratory infections. All green-growing crops are rich in carotene, which animals can convert into vitamin A. Vitamin A supplement is often added to animal rations to ensure an adequate supply when they are not on good pasture.

Vitamin D enables animals to use calcium and phosphorus; a deficiency causes rickets in young growing animals. The ultraviolet rays of sunlight produce vitamin D from the provitamin in the skin. Field curing of hay develops vitamin D through the action of the sunlight on ergosterol in the hay crops. Certain fish oils are very rich in vitamin D. Livestock that are outdoors in the sunlight much of the time have a plentiful supply of vitamin D. Under winter conditions in cold regions, cattle, sheep, and horses ordinarily get ample amounts from the hay they are fed; laboratory animals may be deficient unless a supplement is added.

The vitamin B group is not important in the feeding of cattle, sheep, and other ruminants because the bacteria in the rumen synthesize these vitamins. Very young calves, poultry, swine, and other simple-stomached animals require the B vitamins in their diets. Of these, riboflavin, niacin, pantothenic acid, and vitamin B_{12} are most likely to be deficient in ordinary feeds; special supplements are needed by pigs, poultry, and laboratory animals. Choline may also be deficient in poultry feeds.

Vitamin E is necessary for normal hatching of eggs. It plays a role along with selenium in preventing muscle stiffness and paralysis (dystrophy) in lambs, calves, and chicks under certain conditions. Vitamin C, which prevents scurvy in humans and guinea pigs, can be synthesized in the bodies of other animals and need not be supplied in their food. Vitamin K is synthesized by bacteria in the intestinal tract; a dietary supply is usually not important.

Antibiotics and other growth stimulants. Since about 1950, antibiotics have been added to diets of growing animals because they increase the rate of growth and decrease death loss. Those most commonly used are chlortetracycline, oxytetracycline, bacitracin, and penicillin. An antibiotic helps to overcome the growth-depressing effects of an inadequate diet or of imperfect management practices, but its effectiveness differs among animal species.

Diethylstilbestrol, testosterone, progesterone, estradiol, and dienestrol in various combinations are administered orally or subcutaneously to stimulate growth of cattle and sheep for meat production.

COMPOSITION AND VALUATION OF FEEDS

Usual chemical analyses of feeds include dry matter, protein, fat, fibre, and ash. Energy value, mineral elements, and vitamins are also determined; these values are included in complete tables of feed composition.

Determination. Digestion and balance experiments measure the degree to which the various components of a feed are absorbed and retained by the animal body. Protein requirements are expressed as the amounts of digestible protein needed for growth or other body functions. The amounts of energy needed are measured as digestible energy (DE), metabolizable energy (ME), net energy (NE), or total digestible nutrients (TDN). The total gross energy value of a feed is the amount of heat liberated when it is burned in a bomb calorimeter. Only a part of this energy is available to the animal because some passes through the body without being digested. Furthermore, some of the DE is not utilized but is excreted in the urine as urea. There is a further energy loss from the ME as heat of fermentation and as gas produced by bacteria in the digestive tract. This loss is appreciably greater in ruminants than in pigs, chickens, or other nonruminants. The work of eating, digesting, and metabolizing food may also be subtracted from the food energy, resulting in the NE, or useful energy value of a food. The TDN value of a feed represents the sum of the digestible protein, digestible fat x 2.25, digestible nitrogen-free extract, and digestible fibre. This measure is less useful than NE because it considers neither the fermentation and heat losses during digestion and metabolism nor the fact that

Animal protein requirements

energy is utilized more efficiently for maintenance or milk production than for growth and fattening.

Optimization of nutrient–cost ratio. Feed costs vary widely from season to season; it is often possible for producers to realize substantial savings by wise selection of the feed ingredients used to formulate complete rations. It is much easier for large commercial feed companies with operations in different regions to take advantage of low costs than it is for individual relatively small-scale livestock producers.

Least-cost programming of feed mixtures makes it possible to use electronic computers for selecting the correct amounts of the lowest-cost feed ingredients that will fully satisfy the nutrient requirements of a specific type of animal at a particular stage of development. The system has successfully formulated highly palatable rations that yield maximum production at lowest cost. Again the large-scale producer has an economic advantage.

BASIC TYPES OF FEEDS

Animal feeds are classified as follows: (1) concentrates, high in energy value, including (*a*) cereal grains and their by-products (barley, corn, oats, rye, wheat), (*b*) high-protein oil meals or cakes (soybean, cottonseed, peanut), (*c*) by-products from processing of sugar beets, sugarcane, and (*d*) animal and fish by-products; (2) roughages, including (*a*) pasture grasses, (*b*) hays, (*c*) silage, (*d*) root crops, and (*e*) straw, stover (stalks). The composition of a few of the most commonly used feeds is shown in Tables 1 and 2.

Concentrate foods. *Cereal grains and their by-products.* In North America and northern Europe, barley, corn, oats, rye, and sorghums are grown almost entirely as animal feed, although small quantities are processed for human food. These grains are fed, whole or ground, either singly or mixed with high-protein oil meals or other by-products, minerals, and vitamins, to form a complete feed for pigs and poultry or an adequate supplement for ruminants and horses.

Production of grains is seasonal because of temperature or moisture conditions or both. It is necessary to produce a full year's supply in the limited growing season. The grain is dried to prevent sprouting or molding and stored in containers or buildings where insects and rodents cannot destroy it. It is generally desirable to store more than a year's supply of the grains because crop failures sometimes occur.

High-protein meals. Vegetable seeds produced primarily as a source of oil for human food and industrial uses include soybeans, peanuts (groundnuts), flaxseed, cottonseed, coconuts, oil palm, and sunflower seeds. After these seeds are processed to remove the oil, the residues, which may contain from 5 percent to less than 1 percent of fat and 20 to 50 percent of protein, are marketed as animal feeds. Cottonseed and peanuts have woody hulls or shells, which are generally removed before processing—otherwise the resulting by-product is higher in fibre and appreciably lower in protein and energy value.

These feeds are used as supplements to roughages or cereal grains and other low-protein feeds to furnish the protein needed for efficient growth or production. The supplement chosen for a particular ration depends largely upon the cost and availability of supply.

By-products of sugar beets, sugarcane. From the sugar-beet industry come beet tops, used on the farm fresh or ensiled, and dried beet pulp and beet molasses from the sugar factory. Cane molasses is a residue from cane-sugar manufacture. These are all palatable, high-quality sources of carbohydrates. Sugarcane bagasse (stalk residue) is fibrous, hard to digest, and of very low feed value. In some European countries fodder beets and some other roots are grown as animal feed. Citrus molasses and dried citrus pulp, available at low cost as by-products of the citrus-juice industry, are high-quality feeds for cattle and sheep.

Other by-product feeds. By-products or residues from commercial processing of the cereal grains to produce human food supply large quantities of animal feeds. The largest group of these comprises feeds from the milling of wheat in the production of flour including wheat bran, wheat middlings, wheat-germ meal, and wheat-mill feed. In some areas bakery wastes, such as stale bread, rolls, and pastry products, are ground and used as feed for pets and farm animals. Rice bran and rice hulls are obtained from the mills that polish rice for human food. Corn gluten feed, corn gluten meal, and hominy feed are produced as by-products from the manufacture of starch for industrial and food uses.

Brewers' grains, corn-distillers' grains, and brewers' yeast are useful animal feeds from the dried residues of the fermentation industries that produce beer and distilled spirits. Waste products from pineapple-canning plants include pineapple bran or pulp and the ensiled leaves from the plant. By-products from the abattoirs and meat-packing plants that process animals into meat include such feeds as meat and bone meal, tankage, meat scraps, blood meal, poultry waste, and feather meal. Dried skim milk, dried whey, and dried buttermilk are feed by-products from the dairy industry. Various types and qualities of fish meals are produced by the fish-processing plants. These feeds contain 50 percent or more of high-quality protein and the mineral elements calcium and phosphorus. Steamed bone meal is particularly high in these important minerals.

Bone and fish meals

Roughages. *Pasture.* Pasture grasses and legumes, both native and cultivated, are the most important single source of feed for cattle, horses, sheep, and goats. During the growing season they furnish most of the feed for these animals at a cost lower than for feeds that need to

Table 1: Concentrates Commonly Used as Animal Feeds

type of concentrate	composition (%)					
	dry matter (total)	protein	fat	fibre	ash	NFE*
Barley	88.6	12.0	2.1	7.4	3.3	63.8
Beans (*Phaseolus* species)	89.0	23.0	1.2	4.1	3.9	56.8
Bone meal, steamed	95.0	12.1	3.2	2.0	71.8	5.9
Brewers' grains	92.0	25.9	6.2	15.0	3.6	41.3
Brewers' yeast	93.0	44.6	1.1	3.0	6.4	37.9
Citrus pulp	90.0	6.6	4.6	13.0	6.0	59.8
Coconut meal	93.0	20.4	6.6	12.0	6.9	47.1
Corn	86.5	9.1	4.0	1.8	1.2	70.4
Corn distillers' grains	91.6	29.1	8.9	11.5	3.1	39.0
Corn ears (corn and cob meal)	87.0	8.1	3.2	8.0	1.6	66.1
Corn gluten feed	90.0	25.3	2.4	8.0	6.3	48.0
Cottonseed, whole	92.7	23.1	22.9	16.9	3.5	26.3
Cottonseed meal	94.0	41.0	4.3	12.0	6.2	30.5
Fish meal	92.0	63.2	4.4	—	21.7	27.0
Hominy feed	90.6	10.7	6.5	5.0	2.5	65.9
Linseed meal	91.0	35.3	5.2	9.0	5.6	35.9
Meat and bone meal	94.0	50.6	9.5	—	29.1	48.0
Milk, dried skimmed	94.0	33.5	0.9	—	7.6	26.3
Molasses, cane	75.0	3.2	0.1	—	8.1	63.6
Oats	91.0	12.1	4.8	12.0	3.2	58.9
Palm-kernel meal	91.4	19.2	6.7	11.9	3.9	49.7
Peanut meal	92.0	47.4	1.2	13.0	4.5	25.9
Pineapple bran or pulp	88.6	4.2	1.6	18.4	2.6	61.8
Poultry waste	93.1	52.1	13.6	—	8.5	18.9
Rapeseed meal	93.6	37.1	6.3	13.7	6.0	30.5
Rice, grain	89.0	7.3	1.9	9.0	4.5	66.3
Rice bran	91.0	13.5	15.1	11.0	10.9	40.5
Rye	89.0	11.9	1.6	2.0	1.7	71.8
Safflower meal	91.0	19.7	6.0	31.0	3.7	30.6
Sesameseed meal	93.0	47.9	5.1	5.0	9.3	25.7
Sorghum	88.0	10.3	2.6	2.1	1.9	71.1
Soybean hulls	91.3	12.5	2.1	35.5	4.6	36.6
Soybean meal	89.8	50.9	0.8	2.8	5.6	29.7
Soybean seed	90.0	37.9	18.0	5.0	4.6	24.5
Sugar-beet pulp	91.0	9.1	0.6	19.0	3.6	58.7
Wheat	89.0	12.7	1.7	3.0	1.6	70.0
Wheat bran	89.0	16.0	4.1	10.0	6.1	52.8
Wheat-germ meal	90.0	26.2	10.9	3.0	4.3	45.6
Wheat middlings	89.0	18.0	3.6	2.0	2.5	62.9
Whey, dried	94.0	13.8	0.8	—	9.7	69.7

*NFE = nitrogen-free extract.
Source: Data selected from U.S. National Academy of Sciences, National Research Council, *United States–Canadian Tables of Feed Composition*, 1969.

Table 2: Roughages Commonly Used as Animal Feeds

type of roughage	composition (%)					
	dry matter (total)	protein	fat	fibre	ash	NFE*
Dry roughages						
Alfalfa hay	90.0	16.6	2.0	26.8	8.5	36.1
Berseem	90.6	13.4	3.1	24.4	11.0	38.7
Bromegrass hay	92.8	5.4	2.8	31.7	7.6	45.3
Clover hay, crimson (*Trifolium* species)	87.4	14.8	2.0	28.1	8.3	34.2
Corn cobs	90.4	2.5	0.5	32.4	1.5	65.2
Corn fodder	87.2	5.1	1.0	32.4	6.2	42.5
Cottonseed hulls	90.3	3.9	1.4	42.9	2.5	39.6
Lespedeza hay	93.0	14.6	3.7	28.6	5.5	40.6
Oat hay	88.2	8.1	2.7	27.3	6.6	43.5
Oat straw	90.1	4.0	1.9	36.9	7.4	39.9
Peanut hay, without nuts	90.6	10.0	3.2	23.6	9.6	44.2
Rice hulls	92.0	3.0	0.8	40.7	19.1	28.4
Rice straw	92.5	3.9	1.4	33.5	14.5	39.2
Sorghum, without heads	85.1	4.5	1.8	27.7	8.2	42.9
Soybean hay	89.2	14.5	2.7	28.6	7.2	36.2
Timothy hay	87.7	7.6	2.3	29.1	5.4	43.3
Wheat straw	92.6	3.9	1.5	37.0	8.3	41.9
Green roughages						
Alfalfa	21.1	4.3	0.5	5.5	2.0	8.8
Barley pasture	20.0	5.2	0.8	3.7	3.3	7.0
Beet tops	20.7	2.6	0.4	2.8	7.5	7.4
Bermuda grass	36.7	4.2	0.8	9.5	3.8	26.0
Berseem	18.8	2.7	0.7	4.4	2.8	8.2
Bluegrass	30.5	5.3	1.1	7.6	2.2	14.3
Bromegrass	32.5	7.2	1.4	7.3	3.4	13.2
Clover, ladino	18.7	4.7	0.9	2.7	2.1	8.3
Fescue	23.9	2.8	0.8	7.1	2.0	11.2
Guinea grass	26.8	1.4	0.4	11.5	3.0	10.5
Lespedeza	31.1	5.8	0.8	8.5	3.3	12.7
Mangels, beets	10.6	1.4	0.1	0.9	1.1	7.1
Molasses grass	30.9	1.3	0.9	16.0	2.8	9.9
Napier grass	22.0	1.1	0.3	9.0	2.6	11.2
Orchard grass	23.8	4.4	1.1	5.6	2.7	10.0
Para grass	27.8	1.8	0.4	10.0	2.9	12.7
Rape	16.3	2.9	0.6	2.6	2.2	8.0
Rye grass	24.3	4.0	1.0	5.3	3.8	10.2
Sudan grass	17.6	3.0	0.7	5.4	1.6	6.9
Sugarcane	23.2	1.0	0.8	6.8	1.2	13.4
Wheatgrass (*Agropyron* species)	30.8	7.3	1.1	6.8	3.3	12.3
Silages						
Alfalfa, wilted	36.2	6.4	1.2	10.9	3.1	14.6
Clover	39.9	7.9	1.3	12.7	4.0	14.0
Corn	25.6	2.2	0.8	6.6	1.5	14.5
Pineapple tops	21.3	1.6	0.6	4.8	1.5	12.8
Sorghum	28.9	2.3	0.8	7.8	2.2	15.8
Soybean	28.0	4.1	0.9	8.7	2.8	11.5
Timothy	40.8	4.4	1.3	14.2	2.9	17.0
Root crops						
Beets	13.0	1.6	0.1	0.9	1.5	8.9
Carrots (*Daucus* species)	11.9	1.2	0.2	1.1	1.2	8.2
Cassava	32.6	1.1	0.3	1.4	1.0	28.8
Potatoes	24.6	2.2	0.1	0.5	0.9	20.9
Turnip (*Brassica rapa*)	9.3	1.3	0.2	1.1	0.9	5.8

*NFE = nitrogen-free extract.
Source: Data selected from U.S. National Academy of Sciences, National Research Council, *United States–Canadian Tables of Feed Composition*, 1969.

be harvested, processed, and transported. Hundreds of different grasses, legumes, bushes, and trees are acceptable as feeds for grazing animals. The nutritive value of the cultivated varieties has been studied, but information is incomplete for many of the others. The composition of some of the forages most widely used is shown in Table 2.

Hay. Hay is produced by drying grasses or legumes when they approach the stage of maximum plant growth and before the seed develops. This stage gives maximum yields of digestible protein and carbohydrates per unit of land area. The moisture content must be reduced to 22 percent or less to prevent molding, heating, and spoilage during storage. Legume hays, such as alfalfa and clovers, are high in protein, while the grasses (timothy, Sudan grass) are lower in protein but vary considerably depending upon the stage of maturity and level of nitrogen fertilization applied to the crop. Stored hay is fed when sufficient fresh pasture grass is not available.

Silage. Silage is made by packing immature plants in a storage container to exclude the air and allow fermentation to develop acetic and other acids, which preserve the moist feed. Storage may be in upright tower silos or in trenches in the ground. Best quality silage results when the forage is ensiled with a moisture content of 50 to 65 percent. Lower moisture levels cause difficulty in obtaining sufficient packing to exclude air and may result in molding or other spoilage. Too high a moisture content causes nutrient losses by seepage and excessively acid, unpalatable silage. Ensiled forage can be stored for a longer period of time with lower loss of nutrients than dry hay. The nutritive value of silage depends upon the type of forage ensiled and how successfully it has been cured. Corn, sorghums, and grasses, and sometimes leguminous forages, are used in making silage.

Root crops. These are used less extensively as animal feed than was true in the past, for economic reasons. Mangels, rutabagas, cassava, and sometimes surplus potatoes are used as feed. As shown in Table 2, they are lower in dry-matter content than most other feeds. They are low in protein and provide mostly energy.

Straw and hulls. Straws remaining after wheat, oats, barley, and rice crops have been harvested are used as feed for cattle and other ruminants. They are low in protein and very high in fibre; digestibility is low. Straw is useful in maintaining mature animals during periods of shortage of other feeds, but it is too low in quality to be satisfactory without being supplemented with other feeds to supply the protein, digestible energy, and minerals needed for growth and production. Treatment of straw with alkali markedly increases the digestibility of the cellulose, augmenting its value as a source of energy.

Corncobs, corn stalks, cottonseed hulls, and rice hulls can also be used as sources of fibre in ruminant rations. Rice hulls are lower in value, while the others are similar to straw.

BIBLIOGRAPHY. The most comprehensive summaries of information on animal feeds include the UNITED STATES NATIONAL ACADEMY OF SCIENCES, NATIONAL RESEARCH COUNCIL, *United States–Canadian Tables of Feed Composition*, publ. 1684 (1969), *Composition of Cereal Grains and Forages*, publ. 585 (1958); and B.H. SCHNEIDER, *Feeds of the World: Their Digestibility and Composition* (1947). The reference book by A.M. ALTSCHUL (ed.), *Processed Plant Protein Foodstuffs* (1958), is the most complete compilation of information on the production and composition of animal and human foods that supply protein. The report, "Legumes in Agriculture," *F.A.O. Agric. Stud. 21* (1953), is an excellent summary of leguminous forages throughout the world. Selected information on the composition and use of animal feeds is presented by the "Nutrient Requirements of Domestic Animals" series, reports by the U.S. NATIONAL ACADEMY OF SCIENCES on poultry, swine, dairy cattle, beef cattle, sheep, rabbits, minks and foxes, fish, and laboratory animals. The AGRICULTURAL RESEARCH COUNCIL, LONDON, has also published a series of reports, "Nutrient Requirements of Farm Livestock," that gives information on feeds for poultry, ruminants, and pigs. Textbooks and reference works that are useful sources of information on animal feeds include: O.J. KELLNER and M. BECKER, *Grundzüge der Fütterungslehre*, 14th ed. (1966); E.W. CRAMPTON and L.E. HARRIS, *Applied Animal Nutrition*, 2nd ed. (1969); C.M. MCCAY, *Nutrition of the Dog* (1946); F.B. MORRISON, *Feeds and Feeding*, 23rd ed. (1967); and L.A. MAYNARD and J.K. LOOSLI, *Animal Nutrition*, 6th ed. (1969).

(J.K.L.)

Animals and Plants in Myth and Legend

Animals and plants have played important roles in the oral traditions and the recorded myths and legends of the peoples of the world, both ancient and modern. This article is concerned with the variety of relationships noted between man and animals and plants in myths, legends, and popular folk traditions and in primitive and popular systems of classification.

NATURE AND SIGNIFICANCE

The problem of boundaries. Man has always been fascinated by the problem of boundaries: what distin-

guishes him from another man; what marks off his culture from another; what the dividing lines are between man and other beings—*i.e.*, plants, animals, or the gods. At times he has maintained a rigid sense of separation and viewed the breaking of distinctions as transgression. At other times he has sought to cross the boundaries in order to gain power or knowledge. In some myths, he has glorified an age when the categories of distinction had not yet come into existence, and he has yearned for a return to this paradisical condition. In other traditions, he has viewed with horror the monsters that result when different spheres of being are mixed. These questions and attitudes may be found in every culture and during every age; *e.g.*, modern man is both attracted to and repelled by his similarities to and descent from other animals.

Man is a boundary-marking creature. This is both a conceptual endeavour and a pragmatic one. Not only does his being demand that he position himself in a complex system of relationships, but also his social life and his biological survival depend on the making of distinctions. To speak with the gods, to have relations with this or that woman, to take possession of another's territory, or to eat this or that plant or animal involves man in a host of decisions upon which his existence depends. One of his chief resources for answering such questions is that of the myths and legends that constitute a blueprint of the world in which he dwells.

Relationships between man, animals, and plants

Varieties of relationships. Myths and legends concerning animals and plants employ a wide variety of motifs, but they express a relatively limited number of relational options. Man, animals, and plants may stand (1) in a relationship of opposition or difference; (2) in a relationship of descent; (3) in a relationship of mixture or transformation; or (4) in a relationship of identity or similarity. These will be determined by and expressive of the total world view of a people. The hunter, for example, has a far different understanding of the animal than does the agriculturalist or pastoralist; the tuber planter has a different view of plants from that of the cultivator of grains. Even within these broad categories sharp differences occur. The Kalahari Bushman of southern Africa, who, alone, naked, and crawling on the ground, blends in with his environment in order to kill an animal for food, reveals a different understanding of man's relation to nature than does the Masai tribesman of eastern Africa, who, costumed and walking upright as part of a line of chanting hunters in order to slay a lion as a symbol of his manhood, stands forth visibly as the ruler of the world through which he moves. The Cretan bull dancer of ancient Mediterranean culture, playing with the animal by somersaulting over his back, expresses a different conception of man's relation to this powerful animal and the forces of fecundity and death that it symbolizes than does the Spanish bullfighter who slays the beast.

RELATIONSHIPS OF OPPOSITION OR DIFFERENCE

The fundamental religious boundary is that between the sacred and the profane, the sacred being conceived of as a sphere of power superior to or opposed to the mundane. That which is sacred may be either creatively or chaotically powerful. If the former, it is primarily expressed in creation myths; if the latter, in demonic traditions.

Cosmogonies. The notion of a creator deity in animal or plant form is comparatively rare. There are stories of animals, birds, or insects creating the world and of creators with animal attributes or animal companions, but these are isolated and frequently secondary traditions. Even in the widespread motif of the birth of the world from a cosmic egg there is rarely the notion of a bird laying or incubating the egg (the most notable exception is the world egg laid by a beautiful bird in the beginning of the *Kalevala*, the national epic of Finland). There are, however, a number of cosmogonic (origin of the world) motifs that employ animal or plant symbolism: the cosmic tree that supports and nourishes the world; the earth surrounded by a serpent or supported on a turtle or on some other animal's back; the features of the present world created by the actions of some primeval animal—*e.g.*, lakes and rivers were caused by the digging of an animal or the hills were raised by the flapping wings of a bird. Sacrificial motifs abound, such as the world being formed from the cut-up parts of an animal or restored by its primordial sacrifice.

Animals in opposition to creation

The most important traditions associated with animals occur in dualistic creation accounts in which animals oppose creation, acting as a foil to the creator, or creation is accomplished by combat between the creator and animal monsters representative of chaos who must be slain or bound before the world can be established. The widely distributed earth-diver myth is the most familiar example of dualistic creation. There are two types. In the one, a divine animal plunges into the cosmic waters to bring up the first particles of earth from which the world will be formed. In the other more strongly dualistic version, there is antagonism between the creator (frequently in human form) and the diver (usually in animal form), with the latter attempting to gain power over the creation.

Other oppositions occur with respect to the creation of man (see below *Relations of descent*). Perhaps the most frequent myth of the origin of death is that of the "perverted message" or "two messengers." In the one, an animal is sent with a message from the creator that man is immortal, but the animal alters the message to state that man must die. In the other, two animals are sent, one with the word that man is immortal, one with the message that man will die. A mishap occurs to the first, and only the fatal message reaches man.

Animal and plant deities. The belief in sacred plants or animals is universal. Common to all of these is the notion that the plant or animal is a manifestation of the sacred and thus possesses the dual attributes of beneficence (in healing, hunting, or agricultural magic) or danger (as expressed in taboos against their destruction or consumption). More rarely, gods are believed to have animal (theriomorphic) or plant (phytomorphic) forms. Inheriting the Greek polemics against Egyptian religion and the Israelite polemics against "idolatry," Western scholars have tended to speak of such traditions as "animal worship," although it is usually not the animal itself but rather the sacred power revealed by the animal that is being revered. Other deities possess animal or plant attributes or are incarnations associated with particular animals or plants. Here the animals or plants possess an emblematic function. Certain qualities are associated with certain species (*e.g.*, wisdom with the owl, strength with the lion, immortality with the eagle, inspiration with the grape), and the god's possession of these qualities is symbolized by their being identified with the appropriate animal or plant. In other traditions, natural phenomena are associated with the actions of certain species (*e.g.*, wind as a bird, lightning as a snake), and the god who controls such phenomena is identified with the species. At times, the animal or plant achieves a divine identity of its own; *e.g.*, the thunderbird or the earthquake monster.

Western attitudes toward animals and plants as manifestations of the sacred

Hunting and agricultural deities. *Master of the animals.* In the traditions of archaic hunting peoples there is frequently a figure whom scholars term the master of the animals or the protector of game. He is the ruler of the forest, of all animal species, or of only one particular species (usually a large game animal; *e.g.*, the northern master of the caribou). The master controls all game animals (frequently by penning them up). He dispenses a certain number to man as food and can be invoked by a shaman, a medicine man who possesses magical and psychical powers, when he withholds game. He guides the hunter and, in some traditions, avenges the spirits of slain animals, whose souls return to his enclosures when they die. He is sometimes pictured in human form, on occasion having animal attributes or riding an animal; in other traditions, he is a giant animal or can assume animal form.

In a related complex, a deity in animal form demonstrates to man the art of hunting, serving as the first victim (a motif found in some of the American Indian bear

mother or buffalo woman tales). Or the deity appears among men as an animal who must be slain and eaten so that he may return to his heavenly home (*e.g.*, the Ainu Iyomante feast in Japan).

Corn or rice mother, the dema deity. A similar pattern is found among archaic agricultural peoples. An ancestral (*dema*) goddess, at times in plant form, produces food asexually from her body. She is slain by the tribe, and from the dismembered portions of her body crops appear.

Vegetation deities. The archaic pattern of the *dema* deity needs to be distinguished from the widespread tradition among higher agricultural peoples of the bountiful mother earth or the god or goddess of vegetation or special crops. In these instances, the deity, frequently depicted or associated with animal and vegetative characteristics, is the principle of inexhaustible vitality. The god will frequently have a human consort who will participate in a sacred marriage in order to gain fecundity for man (*e.g.*, ancient Mesopotamian religions).

Culture heroes. The master of the animals or corn mother is frequently found in association with animal culture heroes. An animal or trickster who can assume animal form secures for man the various attributes of culture (acting either in consort with or opposition to the gods). These traditions are found in etiological (origins or causes) stories about how man first learned to hunt, discovered tobacco, and other accomplishments. The most frequent motif is that of the animal who stole fire from the gods for man. Frequently, such traditions will lie behind etiologies of specific animal or plant characteristics; *e.g.*, the bat is black and blind because it stole fire and was singed by the flames and blinded by the smoke. In other tales, the animals oppose the acquisition of culture by man, and must be overcome by a human culture hero.

The functions of the culture hero

A closely related theme is the myth of a life-giving tree or other healing magical plants, growing in paradise or some other inaccessible place, to which the culture hero must travel in order to gain a boon for mankind. He will frequently be assisted by or have to overcome supernatural animals.

Demonic plants and animals. Opposed to these positive conceptions of the creative powers of plants and animals is the notion that their sacred power is chaotic or demonic. Rather than aiding man, they are destructive. The most common examples are monstrous plants and animals who figure especially in heroic quests as guardians of boons or threats to be overcome; mythical animals associated with destructive natural phenomena, such as the earthquake monster or the tradition that eclipses are caused by a monster who devours the sun or moon; personifications of evil powers such as death or disease (*e.g.*, the hound of hell) or chaos beasts (such as dragons) whose release marks the end of the world or who will be slain in a final battle by a saviour deity. A universal phenomenon is the association of certain species of animals with sorcerers and witches. The most frequent form of this belief is that of the familiar—an animal whose soul is bound up to that of the sorcerer, whose form the sorcerer can assume, and who may be commanded to serve his evil master.

Dragons and familiars

Some species (*e.g.*, animals such as the serpent and various narcotic plants) exhibit the ambivalence of the sacred—they are conceived as being both beneficent and dangerous.

Taboo. One of the clearest indications of a sacred animal or plant that stands in a relation of opposition to man, whether conceived positively, negatively, or ambivalently, is the association of a taboo (proscription) with the species against killing, touching, or eating. The species is conceived as being radically different from man, and man enters into its sphere or takes it into himself only at great peril (see below, *Relations of mixture and transformation*).

RELATIONS OF DESCENT

One of the major ways man has of organizing his world is through genealogy or relations of descent. In theogonies, or tales of the origin of gods (*e.g.*, that by Hesiod, a Greek poet of the 8th century BC), or in legendary lists of human offspring (such as the ethnic catalog in the Old Testament), relations of descent and the association of characteristics, territories, and spheres of influence with descendants provide a total blueprint of the cosmos and the human world (see below, *Classificatory systems*). In traditions concerning animals and plants, relations of descent are most prominent in myths of the origin of man and in totemic (animal-clan relational) materials. Central to both is the figure of the plant or animal ancestor.

Cosmogonies. In most cosmogonies, the creation of man and his world occurs after the creation of the cosmos. Man is seen as the last (and, in some traditions, as the culminating) creation. The cosmos has a "prehistory" that is presupposed before man's appearance. Man and the human, cultural world often represent a decisive break with or alteration of this previous creation. This frequently finds expression in a three-stage periodization of creation: the world of the gods and other primordial beings; the world of the ancestors; and the world of man. In some traditions the gods create man or an ancestral pair from whom men descend; in others, the ancestors are transformed into men. Common to both patterns are relationships expressing man's affinity to and distinction from the divine, ancestral, animal, and vegetable worlds.

Man as the last and culminating aspect of creation

Creation of man from plants or animals. A widespread motif, especially among archaic peoples, is that man was descended from plants or animals. These descent traditions usually name a particular species as man's ancestor, and the tribe frequently takes its name from the plant or animal.

In some myths, an asexual mode of creation is implied; a child, for example, appears from the bud of a tree or from a split fruit, or man is a featherless bird sent from the sky. Even the motif of the birth of man from an egg is predominantly an asexual motif. Other traditions, particularly agricultural ones, see man as the product of the mating of a plant or animal species. In some myths, fabrication rather than descent is emphasized. Man is fashioned from a plant or animal by the gods or his parts are modelled after other species. In these descent traditions, the primal man who results is usually the progenitor of a particular people. Other peoples are created from different or less favourable species. These traditions persist in folkloristic accounts of the birth of individuals from plants or animals. Such myths express a close relationship between man and the animal and plant world. Man does not represent a new type of being but rather a new manifestation or form.

Ancestors. The widely distributed notion of animal or plant ancestors places far more emphasis on transformation (see below, *Relations of mixture and transformation*). The ancestral myths describe a primeval time of creation (or successive creations) followed by a decisive alteration in the conditions of life in the shift from the ancestral to the present human mode of being. Compared to the "fixed" characteristics of the present period, the ancestral era was one of flux, lacking definite boundaries. Animals, plants, and men were much the same: they could speak with each other, have sexual relations with each other, and engage in other relationships. The ancestors were polymorphic (many formed). Frequently depicted as emerging from the ground, their movement toward the surface is one of progressive discrimination from compound hybrids to forms somewhat resembling present species. But even on the surface, the ancestors remained relatively fluid: some resembling plants, others animals or men, and all with shared characteristics and the power to change their form at will.

The ancestral era

The ancestors were primordially powerful beings, but due to a variety of causes, their world was transformed, and the present order of things came into existence. Human culture and the decisive features of the world as man now knows it were established during the transformation: man's labour, sexuality, and death are due to some action of the ancestors; the topography of the land is the "tracks" left by the ancestors; men, animals, and plants received their present form after the ancestral age.

Other relations of descent. *Totemism.* The relations to an animal or plant ancestor are frequently associated with the complex phenomenon of totemism. Totemism is primarily a social relationship. It expresses the belief that there is a connection between a group of persons, on the one hand, and a species of animal or plant, on the other. Beyond this simple statement, little generalization is possible. The relationship to the totem (animal or plant symbol) occurs in a variety of forms; associated phenomena (*e.g.*, exogamy, or marriage outside the clan, and taboos against killing the totem species) may or may not be present. The myths associated with such traditions narrate the origin of social groups and the discovery of its totem.

Guardians. A phenomenon that has, at times, been confused with the social relations of totemism is that of the individual guardian. This is a relationship between a particular person and a particular species, usually revealed to the individual in a vision, such as in the vision quests among the Plains Indians in the United States. These guardians become a source of knowledge and good fortune for the individual. To these traditions, more folkloristic motifs may be related, such as the birth of various individuals from intercourse between humans and animals or plants; the animal wife; animal nurses; or the ability of certain persons to understand or converse with animals or plants.

Hierarchy. The fluidity of boundaries that is characteristic of relations of descent raises important questions as to the status of man or of culture in relation to nature. Are animals or plants essentially like men or not like men? Is the human world superior to or inferior to the natural sphere? Such questions lie behind a variety of motifs associated with descent traditions: that the first men were undeveloped, amorphous, or resembled animals; that animals resisted the creation of man; and that the primordial man is the ruler of the natural world. The characteristic fluidity of the descent traditions persists in traditions such as the wild man and in relations of transformation and identity (see below).

The status of man and culture in relation to nature

RELATIONSHIPS OF MIXTURE OR TRANSFORMATION

Boundaries and the maintenance of distinctions guarantee the continued existence of the cosmos as an integrated totality. There are rituals that periodically recount the establishment of the original divisions of the cosmos (especially at festivals of the New Year); other rituals foster remembrance of the decisive deeds of the ancestors in fixing the present state of things; ritualized social structures (such as the caste structure of India) maintain a complex system of distinctions; and religious ideologies (such as astrology) foster the notion of spheres of power that control all members of a class, be they gods, planets, animals, plants, minerals, or men. In such societies, to be real is to affirm and repeat the structures of the cosmos. Each being is called upon to dwell in a limited world in which everything has its given place and role to fulfill. To be sacred is to remain in place. To break out, to cross boundaries, is to open the world to the threat of chaos, to commit transgression. Associated with this world view is the notion that the mixing of realms is the result of evil influence and leads to monsters, hybrids, and uncleanliness. An alternative view in the history of religions sees positive sacred power to be gained from the violation of the given boundaries of the world, from the transcendence of the way things ordinarily are. Each being, in such a view, is called upon to challenge its limits; to break them and to create new possibilities for existence, to achieve freedom. Associated with this view is either the necessity for a periodic loosening of restraint or the celebration of gods or sacred persons who have achieved freedom. Religiously expressed, for the one view, the sacred is the ordinary, that which remains in place; for the other, the sacred is the extraordinary, that which transcends place.

Taboo. These two points of view, *i.e.*, that power comes from conformity to class or freedom from class, may be illustrated by the widespread category of taboo. Research in the second half of the 20th century has emphasized that taboo is primarily a taxonomic (classificatory) system. Those things that are forbidden involve the crossing of boundaries or are beings that fall between classes (see below, *Classificatory systems*). Thus, one may not with impunity enter other spheres (*e.g.*, the realms of the gods) or touch sacred objects, transport an object from one realm to another, cross sexual or class lines, or have relations with a being not of one's class. Many of the food taboos have been shown to reflect taxonomic anomalies. An animal such as the bat is tabooed because it has fur like a mammal but flies like a bird; it has wings like a bird but has fur rather than feathers—and therefore is neither mammal nor bird and must be shunned. On the other hand, the consumption of forbidden foods or engaging in forbidden sexual practices (including homosexuality and bestiality) is part of the ritual of transcendence in many cultures. If an individual can survive the crossing of boundaries he will obtain extraordinary sacred power (*e.g.*, the practitioners of Tantrism, an Indian esoteric and magical system, who violate both eating and sexual taboos; the Jewish magicians mentioned in the Old Testament Book of Isaiah, chapter 65, who eat the forbidden swine and say, ". . . do not come near me, for I am holy").

Proscriptions against crossing boundaries

Monsters and hybrids. From the earliest cave paintings to the most recent and private dreams of individuals, human imagination has been preoccupied with monsters. Monsters are chaos beasts, lurking at the interstices of order, be they conceived as mythic creatures who preceded creation, survivals from an archaic era, creatures who dwell in the dangerous lands that are remote from human habitation, or beings who appear in nightmares. Though the forms and types of monsters are numberless, a single principle underlies the majority of them: a monster is out of place, conforming to no class or violating existing classes. This is most frequently expressed by the monster's having hybrid form (the result of a mixture of species, attributes, sexes, and other categories), being the result of a transformation, or having dislocated or superfluous parts. The dragon, for example—perhaps the most widespread monster in myth and folklore—is born through a mixture of species: it is a serpent born asexually from a rooster's egg that is incubated in manure; by the transformation of an animal; or by the joint generation of a man or worm and a metal. Its form is a compound of species: the body of a serpent or crocodile with the scales of a fish; feet, wings, and occasionally the head of a bird; the forelimbs and occasionally the head of a lion; or, in another dominant type, the ears of an ox, the feet of a tiger, the claws of an eagle, the horns of a deer, the head of a camel, the eyes of a demon, the neck of a snake, the belly of a mollusk, and the scales of a fish. In other types of dragons, organs or attributes of the snake, lizard, fish, mollusk, toad, elephant, horse, pig, ram, deer, eagle, falcon, octopus, or whale will predominate. In many traditions, the dragon has the power to transform itself at will. It possesses superfluous organs, most frequently being many headed, and it has both subterranean and aerial characteristics and habits.

Generation of the dragon

Compound monsters. These same features are present in the majority of monstrous forms. The most common form consists of a mixture of attributes from differing species—*e.g.*, the Centaur (horse-man), the Minotaur (bull-man), Echidna (snake-woman), Pegasus (horse-bird), Sphinx (woman-lion-bird), Siren (bird-woman), and Empusa (animal-metal) of Greek mythology, and the griffin (lion-eagle), mermaid (woman-fish), the vegetable lamb (plant-animal), the barnacle goose (mollusk-bird), and mandrake (plant-man). In other instances, the characteristics are juxtapositions of different species—*e.g.*, the tree that bears human heads as fruit; horses born from eggs; flesh-eating mares; milk-producing birds.

Multiplication or lack of organs. Because modes of locomotion and other bodily characteristics are prime modes of classification, the superfluity or lack of organs removes the monster from the ordinary taxonomic divisions.

Protean figures. The most extreme form of the fluidity that is characteristic of monsters is the Protean figure

who can change into any form or combination of forms at will.

In all of these monstrous forms, the central notion appears to be the danger associated with beings that are out of place or are fluid. Some contemporary anthropologists have argued the opposite conclusion; *i.e.*, rather than being threats to the classificatory system, monsters, through their startling combinations and juxtapositions, force men to think more clearly about and distinguish more sharply between the different boundaries of their world. In this interpretation, the monsters are ultimately supportive of order rather than a destructive threat to it.

Union of disparate features expressing a creative totality

Reconciliation of opposites. In some traditions, the union of disparate features or opposites does not express a chaotic hybrid but rather a creative totality (the *coincidentia oppositorum*, or the "coincidence of opposites"). Though most frequently expressed by bisexuality, either in traditions of an androgynous (man-woman) creator or first man, the theme is present in some animal and plant traditions as well (*e.g.*, the emergence of man from the androgynous *rīvās* plant in Iranian mythology). Although it occurs in cosmogonic settings (*e.g.*, the tree that unites heaven and earth), motifs of the reconciliation of animal and plant opposites more usually occur in paradisical imagery that promises the harmonious mingling of realms (*e.g.*, the "peaceable kingdom" of Isaiah 11:6–8 or Virgil's IV *Eclogue*).

Transformation. One of the largest groups of animal and plant traditions in folklore and religious material is that of transformation. Familiar stories—such as Beauty and the Beast; the transformation of a man into an ass in the *Metamorphoses* by Apuleius, a Roman writer of the 2nd century AD; the frog king or the swan maiden, as well as such well-known traditions as that of the werewolf, the vampire, or leopard men—testify to the wide dissemination of this theme. Every permutation and combination exists: man into animal, bird, fish, insect, reptile, amphibian, or plant; animal into man, bird, fish, insect, reptile, amphibian, or plant; and animal into another species of animal or plant into animal. There are also partial transformations resulting in hybrid forms as well as alternating transformations; *e.g.*, animal, man, or tree by day and the reverse at night. Another great series of transformations concern the dead, who either transmigrate into or return in animal and plant forms.

The power to compel another to change form, or to cross boundaries oneself at will, may be judged good or evil depending on the assessment of order in the world view of the particular culture. In the majority of instances of transformation of another, the transformation is considered to be the result of evil magical powers, and most tales conclude with the disenchantment of the subject and his return to his original form. Many of the instances of self-transformation are for the positive purpose of transcendence.

Cultic transformations. Several of the motifs present in the folklore of transformation suggest cultic procedures (*e.g.*, transformation into an animal by putting on its skin). Cultic experiences probably lie behind and lend credibility to many such tales.

Ritualized moments of transition

In many societies, ritualized moments of transition are accompanied by a liminal period in which boundaries are broken and chaos rules, only to be overcome as order is restored. This is prevalent in the Saturnalia (orgiastic festivals) that are a part of the New Year scenario and in rites of passage, such as initiation. Animal and plant transformations play a significant role in such ceremonies, both as negative symbols of chaos (*e.g.*, the dead return in animal form and mingle with the living; ritualized combats against the primordial dragon) and as positive symbols of the breaking through of bounds and the release of the forces of life (*e.g.*, the presence in many of these ceremonies of young, virile, semitheriomorphic males who engage in sacred sexual intercourse). Prominent in such Saturnalian traditions are deities such as Dionysus, a Greek god, who can assume vegetable, animal, or human forms at will, and who is a god of sudden, dramatic epiphanies (manifestations) and license, and whose devotees, through orgiastic rituals, participate in his freedom to break all bounds in order to recover the boundless vitality and fecundity of primordial chaos. A new life for the cosmos, society, and the individual is obtained through the abolition of the order of the old.

The initiatory scenarios make use of transformations to a somewhat different end. The initiant receives new birth by the dying of his old self after a series of ordeals. Antagonists, frequently in masked animal form, torment him, and his "death" and rebirth is analogous to the hero's successful fight against monsters. Alternatively, the culmination of initiation is frequently the narration of the myths of the ancestors and the vision of them. Masked men, in mixed animal and plant forms, appear to the initiant as his true origin in contradistinction to his biological parents. Other ritual uses of masks achieve the same effect: the ritual transformation of the "actor" into the sacred animal, plant, or deity.

The role of the shaman

Within many cultures there are religious specialists in the breaking of bounds. Perhaps the most widespread example is that of the shaman who is able to journey at will to heaven or the underworld, mingling with both the gods and the dead. His journey occurs through magical flight (frequently in the form of a bird), with animal psychopomps (soul conductors) or guardians, or by ascending the sacred tree that connects heaven and earth. The shaman may transform himself into an animal, and he knows how to converse with them. Similar phenomena are the nagualism of Central America or the leopard societies of Africa in which the practitioner is believed to be able to transform himself into an animal that is frequently considered to be his incarnate "second self."

RELATIONSHIPS OF IDENTITY OR SIMILARITY

Works on primitive mentality published in the 19th and early 20th century usually presumed that primitive man could not distinguish between plants, animals, and men. This "hazy vision," as it was often called, was believed to lie at the root of religious phenomena such as animism (belief in souls) and totemism. More recent studies have demonstrated the presence of complex taxonomies among archaic peoples, although they do not usually employ the criteria of a modern biologist (see below, *Classificatory systems*). Relations of identity, when compared with the other forms of relationship already described, are comparatively rare, occurring most frequently in traditions about the soul. Relations of similarity are more common, usually in literary settings, such as plant or animal fables.

Soul. The most common expression of identity relates the soul of man to that of animals or plants in a variety of ways.

Soul-stuff. Although man tends to associate the soul with personal survival or continuity after death, there is an equally archaic view that emphasizes the continuity of life. This view, to which the Dutch anthropologist A.C. Kruyt gave the term soul-stuff (*Seelenstoff*, as opposed to the postmortem soul, or *Totenseele*), is chiefly found among the rice cultivators of the Indonesian culture area, although it is also witnessed elsewhere. Central to this belief is the circulation of vitality throughout different levels of existence. The soul-stuff is created by the deity as an indestructible reservoir of life. It is eternally reborn —either by returning to its creator, who will redistribute it, or by transmigrating into an embryonic human, animal, or plant. Whatever form it assumes, the same "stuff" is common to all beings.

Death or postmortem soul. The majority of traditions concerns the postmortem soul that leaves the body or comes into existence only after death. A number of motifs reflecting different assessments of the nature of life and death occur. The soul may assume an animal or plant form or there may be animal psychopomps, most frequently a winged creature such as a bird or butterfly. The soul may transmigrate into or be reincarnated as an animal or plant. These traditions need to be distinguished from those concerning spirits of the dead who reappear in bestial form. Related to these are traditions about the separable soul, which is capable of removing itself or being removed from a person while still living. This most

usually occurs in sleep. While detached it may be placed in or assume the form of an animal or, more rarely, a plant. In general, where the notion of soul-stuff predominates, relations of identity are prevalent; where the notion of a death soul is present, the traditions are more closely akin to relations of transformation (see above).

Psychic interrelations

Plural souls. A more complex pattern, of wide distribution, is that of the plurality of souls. Man's vitality and personality are viewed as the result of a complex set of psychic interrelations. A classic example is that of the Apapocuva-Guaraní of Brazil, as described by the anthropologist Curt Nimuendajú (1914): a gentle vegetable soul comes, fully formed, from the dwelling place of the gods and joins with the infant at the moment of birth. To this is joined, shortly after birth, a vigorous animal soul. The type of animal decisively influences the recipient's personality: a gentle person has received a butterfly's soul; a cruel and violent man, that of the jaguar. Upon death, the vegetable soul enters paradise; the animal soul becomes a fierce ghost that plagues men. The plurality of souls provides a complex taxonomy accounting for and relating the distinctive character traits of plants, animals, and men.

The alter ego, or life index. Other religious and folkloristic traditions view the life of man as bound up with that of a plant or animal: if the one is destroyed the other will die as well. In some traditions, this is confined to the familiar or guardian of a witch or shaman; in others, it is an individual relationship possible for any man. More rarely, there is a relation between an entire tribe and a particular plant or animal. In some societies, a ritual of identification is performed, usually at birth (*e.g.*, planting a tree or burying the placenta at the roots of a tree). In others, an individual has a vision or undertakes a vision quest to identify his alter ego.

Types of perceptions of relations. Relations of similarity between human beings and plants or animals usually depend upon the perception of an attribute or aggregate of attributes that they have in common. This process is apparent in colloquial expressions such as when someone is called a "cool cat," a "bitch," a "clumsy ox," a "greedy pig," or "foxy." A similiar process appears to lie behind many of the so-called totemic names or theriophoric or phytophoric personal names (*e.g.*, Swift Deer; Bold Eagle) and is concealed in a number of familiar Western names (*e.g.*, Leo, "the lion"; Deborah, "the bee"; and Jonah, "the dove"). The reverse process, the giving of human names to plants or animals, also depends in a majority of instances on the discernment of character similarities. Care must be used, however, in the interpretation of proper names. Every plant or animal name does not necessarily reveal the perception of similarity. For example, the Seminole Indians combine a character name with a shape and animal name in an arbitrary fashion that appears to pay no attention to their meaning. One well-known Seminole medicine man's name translates as "crazy, spherical puma!"

Function of fables

In the universal literary form of plant and animal fables, this same process is at work. The fable is a didactic tale centred upon one character trait. It is told about a plant or animal, but the intention is that the hearer or reader perceive the similarity between himself and the actor in the tale. He possesses the same characteristics, and hence his actions will lead to the same consequences as the plant or animal in the narrative. The story depends for its point upon the association of a limited number of characteristics that each animal or plant has and on the commonality between these characteristics and human personality traits. More complex forms, verging on allegories, such as the beast epic and the debates between various plants and animals as to which are superior, also exist. The popular *Physiologus* ("Naturalist"), a Greek work from the 2nd century AD, and the medieval bestiary traditions draw morals particularly from monstrous or wondrous animals and plants. Both the fable and the bestiary traditions contributed to the formation of the stereotyped bird, beast, and flower emblems that figure in heraldry and religious iconography.

The process of discovering similarities of personality between plants and animals, on the one hand, and human beings, on the other, also plays a significant role in certain archaic sciences. Physiognomy, which studies the correspondences between bodily features and psychological characteristics, often makes use of such relations of similarity. The earliest Western systematic treatise, the Aristotelian *Physiognomonica*, maintains that people with facial characteristics that resemble certain animals have analogous temperaments (*e.g.*, the male is compared to the lion, the female to the leopard; and humans who have noses with slight notches are impudent like the crow). These views persist in popular figures of speech, such as "bulldog jaw." The same structure underlies the use of plants and animals in archaic healing practices, alchemy, and astrological tables in which animals, plants, and minerals, as well as human personality traits, are associated with the birth signs of the Zodiac or planets.

CLASSIFICATORY SYSTEMS

One of the most exciting and profound developments in the study of archaic cultures, myths, folklore, and rituals in the 20th century has been the increased focus on structural elements and systems of logic, order, and classification. The beginning of this new perspective may be dated from the early part of the 20th century with the works of the French Sociological school, especially the essay by Émile Durkheim and Marcel Mauss, *De quelques formes primitives de classification* (1903; *Primitive Classification*, 1963). More recently, the works of some of the British social anthropologists (*e.g.*, A.R. Radcliffe-Brown, Edmund Leach, and Mary Douglas), U.S. ethnoscientists (*e.g.*, Harold C. Conklin and Charles O. Frake), and French structural anthropologists (*e.g.*, Louis Dumont and, especially, the important studies of Claude Lévi-Strauss) have continued and developed these concerns. In earlier anthropology, "primitive mentality" was characterized by the inability to make distinctions, by a sense of "mystic participation" or identity between man, his cosmos, and all other beings in contradistinction to the logic of modern man. All of the data described in the previous sections were interpreted as resulting from this primitive "haziness." Beginning with the complex kinship systems and continuing by exploring other taxonomies, the new perspective argues to the opposite conclusion: the primitive is, if anything, obsessed with the making of distinctions; his taxonomies reveal a complexity and sophistication that rival those of modern man.

Basic types of systems. A taxonomy is a hierarchical system of ordering. Taxa (classes) at the same level differ or exclude one another; taxa at a higher level include lower categories in a relation of similarity. The lowest division is thus the most different ("the least species," *infima species*); the highest is the most inclusive (the *summum genus*). Biological taxonomies enumerate a number of divisions (*e.g.*, kingdom, phylum, class, order, family, genus, species) on the basis of morphological and genetic characteristics that exhibit relations of exclusion and inclusion. Thus, on the species level, the white oak (*Quercus alba*) differs from the black oak (*Q. velutina*). At the genus level, they are both included as oaks. The oaks differ from the maples at the genus and species level, but they are included in the same class and phylum. The oak and algae differ at every level of classification save for their inclusion in the plant kingdom. All taxonomies, whether biological or not, share these same characteristics: each is a hierarchy of subdivisions; each subdivision excludes each other; each is included in the next higher category that differs from all other classes at the same level.

A related form of classification is that of the tree (the logical tree of Porphyry, a Neoplatonic philosopher of the 3rd–4th centuries AD, is the most famous example). This focusses on features or attributes of the thing being classified rather than the class to which it belongs. The tree operates on the principle of binary opposition with a yes–no, present–absent decision being required at each level for a single feature. If man asks of a plant whether or not it has a fruit, this question provides the basic category. If it has a fruit, is it fleshy or dry? If it is dry, does

it open at maturity or not? If it opens at maturity, does it have one carpel or more? If it has two or more, is a portion of the stem axis included in the capsule or not? If it is, then one is probably describing an iris.

These basic types of classification (with the addition of the paradigm; see *Classification of animals and plants*, below) occur in both modern scientific and popular classificatory traditions, with the scientific utilizing the taxonomy to a greater extent.

Variations in distinctions

Fundamental categories. The basic distinction made by all biological classificatory systems is between the plants and animals with some intermediate category. In popular traditions, an equally important distinction has been between man and the animals with a third category, the "missing link" or monsters, for those that fall in between. Popular traditions draw further distinctions that are in part morphological and in part qualitatively distinguishing, such as between trees and shrubs (tall and short woody plants), and between animals that live on land, in the water, or are aerial. Other distinctions are cultural or utilitarian: domestic or wild animals; plants (cultivated species) as opposed to weeds (uncultivated or undesirable plants); species that are useful for food or for medicine and other distinctions. The Navajo Indians of the American Southwest, for example, make a fundamental division between beings that have speech (man) and beings that lack speech (plants and animals). The animals are further subdivided by means of their mode of locomotion (running, flying, crawling) and their time of greatest activity (day or night). Plants are subdivided according to sex (a symbolic rather than a biological category in this instance), their possession or lack of medicinal qualities, their appearance (*e.g.*, thorny), and by height or size (large, medium, small).

Relativism in popular classifications. A characteristic of popular as opposed to scientific classifications is its relativism. As opposed to the fixity of scientific nomenclature, universally recognized categories and rules for classification, popular traditions frequently have many names for a single species or a single name for many species depending on what characteristic is being lifted out as interesting to the classifier. Much depends on the point of view of the individual: what is a weed to a gardener is an herb to a curer; a rose in a cornfield is a weed to the farmer. Not only are the categories fluid but the complexities of the classification also vary widely according to occupational or social class. In the same way as a layman might distinguish between red, white, rosé, dry, and sweet wines, while a connoisseur would distinguish several hundred varieties and vintages, the Bedouins of North Africa have close to 6,000 words descriptive of camels, whereas those coming from a European background possess, at best, half a dozen; the Eskimo distinguish some 500 types of snow and ice; the Nuer of Africa and other cattle-breeding tribes distinguish between minute differences in cows.

Classification of animals and plants. In treating the classification of animals and plants, it is necessary to distinguish between the systems that order animal and plant species and those systems that use animals and plants as classificatory devices (*e.g.*, the identification of constellations by animal names; the designation of class by animal and plant names). This section will be concerned only with the former.

Cosmogonic origins of categories

In popular taxonomies, the origin of the categories is usually given in the cosmogony. At the beginning, the fundamental divisions of the environment and the appropriate plant and animal species into water, land, and air were established (*e.g.*, chapter 1 of Genesis in the Old Testament). The unique origin of man, who is usually not treated as an animal, is likewise given in the creation story, and his special characteristics are described (*e.g.*, chapters 1 and 2 of Genesis). Other distinctions that follow from the origin myths are divisions into sacred or tabooed animals and plants, totemic species, helpful or domesticated animals and plants, and other distinctions. Thus, the Desana-Tukano Indians of the northwest Amazon in South America distinguish three categories of animals: (1) mammals and birds created by the sun father (chiefly species that are hunted); (2) fish, amphibians, and reptiles that came with the snake-canoe, which carried the first humans, and therefore represent ancestral creations; (3) inoffensive animals—*i.e.*, small animals who "created themselves" when the world was formed. Around this fundamental division are clustered associated species, sex categories, occupations, environmental differences, and other distinctions.

Once this basic cosmic map is established, there is a tendency in popular classifications to associate other cosmic elements with a variety of species (*e.g.*, cardinal directions, primary colours, sex, geographical features). One of the key differences between modern, Western systems and popular traditions is the propensity of the latter for making multidimensional associations. Thus, among the Hopi Indians of the American Southwest, the area of the northwest is associated with yellow, the puma, the oriole, the Douglas fir, the green rabbit bush, the mariposa lily, yellow corn, and the French bean; the southwest with blue or green, the bear, bluebird, white pine, sagebrush, larkspur, blue corn, and butter beans. Often the possession of a single attribute is common (in this case, colour) or the resemblance of parts is sufficient to link disparate elements into a single class. The Nuer, for example, relate the bee and the python because of their similar bodily markings; the Desana classify all animals according to whether they make whistling, humming, or percussive noises, linking the woodpecker, the wild boar (because of the clicking of his tusks), and the jaguar (because of its rhythmic breathing) together as percussive.

Another feature of popular classification is that relations are frequently described in terms of domains and spheres of influence, rather than in isolated, artificial classes. These will frequently be keyed to environmental features: upriver and downstream or desert as opposed to swamp.

Hierarchy, exclusion, and inclusion

Beyond these features, the taxonomy, with its characteristics of hierarchy, exclusion, and inclusion, is common. The Hanunóo of the southern Philippines, for example, identify 1,625 mutually exclusive plant forms. These are grouped into 890 categories (by way of comparison, a Western botanist would identify the same collection as consisting of 1,100 distinct species and 650 genera). The distinctions made are scientific; they do not merely reflect utilitarian concerns but rather the desire to rationally structure and order the diverse phenomena that make up their world (less than 600 of the species are edible, only 406 are claimed to have any medicinal value). The Hanunóo make a basic distinction between things that can and things that cannot be named. Objects that can be named are divided into two classes: things, and animals and persons. To this basic scheme are added further subdivisions. Thus, if the Hanunóo wish to describe a particular variety of pepper known as "cat's penis," they would state that it is a thing that can be named, a thing (as opposed to a person or animal), a plant (as opposed to a rock or manufactured object), a herbaceous plant (as opposed to a woody plant or vine), a pepper plant (as opposed to all other types of plant), a house-yard pepper plant (as opposed to a wild pepper), a house-yard chili-pepper plant (as opposed to other cultivated pepper plants; *e.g.*, green peppers), and a cat's penis house-yard chili-pepper plant (as opposed to other house-yard chili-pepper plants such as the "cock's spur" variety). The tree kind of classificatory scheme may be illustrated by the Subanum of the eastern Philippines, who initially distinguish three primary types of plants by which they classify 1,400 species: woody plants, herbaceous plants, and vines. These are distinguished by the presence or absence of two characteristics of stem growth: woodiness and rigidity. The woody plants are woody and rigid; herbaceous plants are nonwoody and rigid; vines are nonrigid. Further subdivisions would follow the same basic pattern (*e.g.*, with or without thorns, with or without fruits) until the individual species was reached.

Paradigm

An additional form of classification that is not found in Western scientific biological systems but frequently occurs in popular traditions, particularly in linguistic or kinship materials, is the paradigm. A paradigm relates

simultaneously a number of features that intersect on a grid. Thus, if one wanted to map a farmer's description of poultry (chickens, turkeys, ducks, and pigeons), he might describe them with a paradigm using the two axes of sex and maturity (see Table, below). An adult male chicken is called a rooster or cock, a turkey is a tom, a duck is a drake, and a pigeon, if distinguished at all, is a cock. Adult females are called hen, hen, duck, and hen, respectively. A neuter chicken is a capon (a term that is used for the other species as well). Adolescent female chickens are called pullets; adolescent pigeons are squabs without distinction as to sex; the others lack special terminology. Baby chickens are chicks, baby turkeys are poults, baby ducks are ducklings. Thus, roosters, toms, and drakes are classified together according to the sex axis in that they are adult male poultry as opposed to female or neuter. No other relationship is claimed but that of intersection. The advantage of this classification for the interpreter is that it enables him to see what is unclassified and raise certain questions (*e.g.*, why is no sexual distinction made for baby poultry, why no special terminology for adolescent turkeys or ducks).

Farmer's Classification of Poultry

	male	female	neuter	no sexual distinction
Adult	rooster or cock	hen	capon	—
	tom	hen		
	drake	duck		
	cock	hen		
Adolescent	—	pullet	—	squab
Baby	—	—	—	chick
				poult
				duckling

The most widespread use of the paradigm is in India, where everything from food to diseases and castes may be plotted on the two axes, colour (red to green) and heat (hot to cool).

CONCLUSION

Man's relationship to nature and the divine

In the wealth of traditions associated with animals and plants, the common concern, whether implicit or explicit, has been for establishing relations of similarity and difference. The ultimate question has been whether human beings are a part of nature, more closely akin to the gods, or a separate and distinct mode of being. The chain-of-being structure that dominated Western thought through the Middle Ages could answer this question by placing man in between. Man was the highest of the animals, the lowest of the gods. Man's body was like that of the animals: corporeal, sensate, and mortal. Man's spirit or intellect resembled the gods: incorporeal, rational, and immortal. The great surge of ethnological and biological data and theories from the 16th century on tended to disrupt this point of view. New forms of men were encountered (*e.g.*, the "savage") who seemed to their first describers closely akin to the brute; new biologies were proposed that placed man wholly within the animal kingdom, merely as one species among many, and postulated man's descent from the beast. More recently, psychology and ethology have emphasized the irrational (or brutish) elements in man and pointed to the close analogies between animal and human behaviour. Since the 18th century, the question of the place of man has been formulated in a new, nonbiological way: man as a cultural being as opposed to the natural realm. There have been many forms of this dichotomy: man is the only being who has a language, who uses symbols, who employs tools, who freely plays, who is self-conscious, or who possesses a history. Man, in short, creates himself as a cultural being in contradistinction to the animal or plant who is created by his environment or heredity.

Over and against this modern view that man makes himself is the more archaic view that man was formed by the gods. His history is given in the myths of the primordial establishment of things, and his solemn responsibility, along with every other living thing, is to fit himself within this given world. This does not mean that archaic man lacks distinctions. Among the African Lele, for example, animals are distinguished from man by their lack of manners, their immense fecundity, and by their sticking to their own sphere and avoiding contact with humans. Animals that violate this third characteristic are understood to be human-animals, either the result of sorcery or metempsychosis. But the distinctions and the etiquette required by them have different values (most usually that of sacred and profane, clean and unclean, and other similar ritualistic values).

In folkloristic material, four basic forms of relationship were described, ranging from opposition to identity. In each, the basic description was provided by the cosmogony. Each has rules governing man's responsibility toward this relationship, and each has cultural means for transmitting and enforcing these rules. In the more explicit systems of classification, the cosmogony played an equally important role. Though there was clearly a more speculative element, this speculation was related to a determination of the way things really are and the need to conform to and respect this basic pattern. Thus, though the nomenclature for plants and animals is learned by a child from birth, the logic of the system is frequently revealed only at initiation, at which point the initiant, as an adult, become responsible for the proper observance of all the boundaries required by his society (*e.g.*, among the Senufo of Africa, 58 figurines are presented to the initiant in a carefully prescribed order that provides an inventory of the basic classes of animals, men and human activities, and social distinctions).

Myths and folklore about animals and plants are not simply entertainment, although this element is nowhere lacking. They constitute, as do the systems of classification, a repository of truth that man must learn in order to survive. To violate the structures revealed in them through ignorances or design is to threaten life itself.

BIBLIOGRAPHY. The most valuable source for traditions regarding animals and plants is STITH THOMPSON, *Motif-Index of Folk-Literature*, rev. ed., 6 vol. (1955–58). See also A.A. AARNE, *The Types of the Folktale*, 2nd rev., trans. and enl. by STITH THOMPSON (1961), esp. types 1–299. Useful surveys are provided in the articles in J. HASTINGS (ed.), *Encyclopaedia of Religion and Ethics* (1910–22), particularly those by N.W. THOMAS, "Animals" and by T. BARNS, "Trees and Plants." *Studium Generale*, 20:69–361 (1967), contains a valuable collection of articles on plants and animals in religion, folklore, and iconography in various cultures. Each article has extensive bibliography and examples. For important treatments of classificatory systems, see CLAUDE LEVI-STRAUSS, *Le Totémisme aujourd'hui* (1962; Eng. trans., *Totemism*, 1963); *La Pensée sauvage* (1962; Eng. trans., *The Savage Mind*, 1966); and M. DOUGLAS, *Purity and Danger: An Analysis of Concepts of Pollution and Taboo* (1966).

(J.Z.S.)

Animation

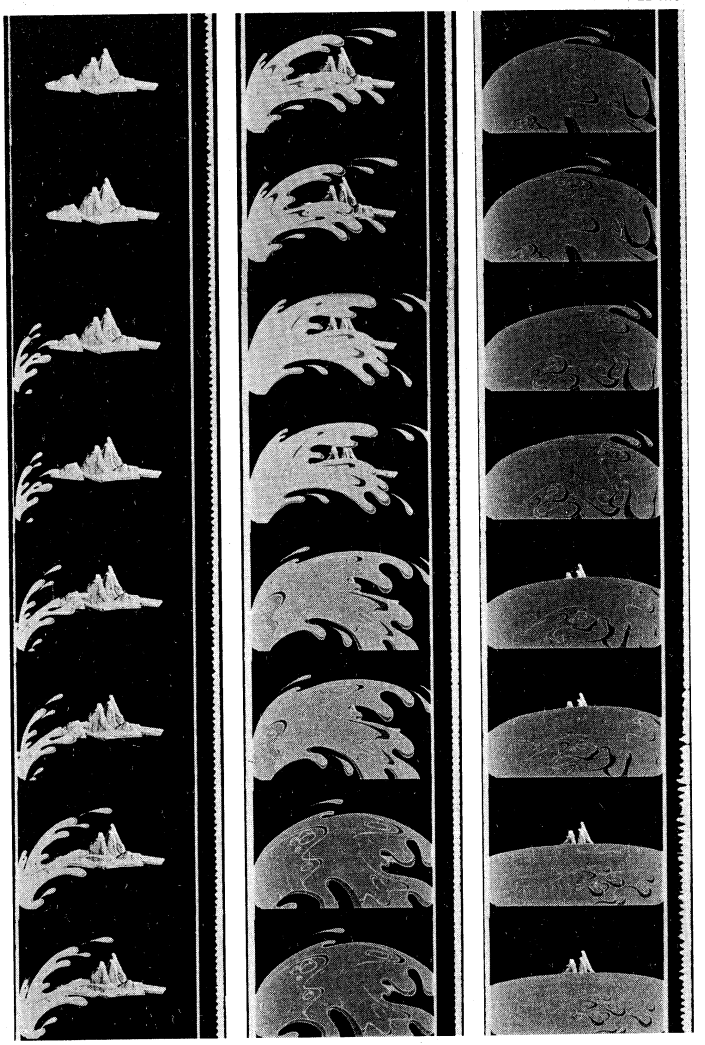
Animation is a method of creating the illusion of life or movement in inanimate objects or drawings, an extension of graphic art into the field of cinematography. It brings the free imagination of the artist or the puppeteer into direct association with the dynamic art of the film. Puppets or drawings of figures can walk, talk, or mime, and the characterless package on a television commercial can open, close, or distort itself. Through animation the drawing or the caricature, comic or serious, comes to life to create a dramatic fable, to make a satiric comment, or to reveal by mobile diagram the intricacies of some technological or scientific process. Or it can exist simply to amuse, and the best known works of animation are undoubtedly the cartoon comedies. The potentialities of animation have constantly fascinated film makers, and experiments in the medium have been made since the earliest days of the motion picture.

Animated films, whatever their length or purpose, have a motion and characterization more succinct than live action. They can more freely contract or distort both time and space and defy the forces of gravity; they can be as realistic or as fantastic as their designers wish.

With mobile sculpture, the animated film offered new dimensions in the visual arts. Short film loops have been designed to bring apparent movement to works of painting and sculpture, and in 1956 Pablo Picasso's feature film *Le Mystère Picasso* employed a series of drawings and paintings specially created to evolve in time as well as space on the screen.

The basis of all animation is the building up, frame by frame, of the moving picture by exact timing and choreography of both movement and sound. All film movement is achieved by projecting during every second of time a certain number of frames, normally 24, each a still photograph minutely varied from its predecessor, which record the successive phases of the subject's movement before the camera. The same motion, of a stylized

EB Inc



Twenty-four frames, the length of film projected each second, from *If You Could See the Earth*, an educational film. The jagged white line to the right of each frame is the soundtrack.

or caricatured version of it, can be achieved by "stop-motion" or "stop-action" cinematography, the frame-by-frame photographing of a similarly phased series of drawings or the phased movement of such objects as puppets, marionettes, or commercial products. And, as in live filming, the camera itself can create movement by tracking into a scene or panning across it.

TECHNIQUE AND PRODUCTION

The range of animation techniques is broad. The basic form is the simple, outlined figure that moves against a simple, outlined background.

Figural basis of animation. The development of cel (or cell) animation permitted the phased movements of the figures to be traced onto a succession of transparent celluloid sheets and superimposed, in turn, onto a single static drawing representing the background. With this technique the background could be drawn in somewhat greater detail and tonal qualities introduced through shading, while the figure itself became a black silhouette, blotting out the background when the cels were superimposed. Multiple cel animation—the superimposition of several cel layers, each carrying different figures or parts of figures requiring special care in animation—allowed increased complexity in the image with minimum work load for the artist-animators. With the more modern forms of colour film introduced in the early 1930s, opaque paints and coloured inks could be used on the cels. Cel animation required the use of a so-called rostrum camera, which photographs downward onto the background with its series of superimposed cel layers pegged into place to secure accurate registration.

Cel animation

Other forms of animation include silhouette animation, developed by Lotte Reiniger in Germany during the 1920s. It used jointed, flat-figure marionettes whose poses were minutely readjusted for each photographic frame. Movement was similarly simulated in puppet animation, which used solid three-dimensional dolls in miniature sets. In another development, abstract mobile patterns were painted directly onto and along the celluloid reel itself, thereby omitting the photographic process altogether. These basic animation techniques, as well as some experimental forms rarely used, had been fully developed by the 1930s. Subsequent developments have involved many refinements in procedure, and, more recently, computerized techniques.

The great majority of animated films are short and have always been so for obvious reasons. When each second of action requires, for the fullest animation, 24 adjustments of the cel series, a minute's action may call for many hundreds of drawings. Flat-figure and puppet animation require adjustment on a similar scale to fulfill the needs of stop-frame cinematography at its most refined.

Planning. The preparation of these films, whatever their length or form, follows a similar process. First comes the story, plot, action, or situational idea, which may be a written treatment with or without supporting sketches. It describes the continuity of what it is proposed should take place on the screen, the nature of the cartoon or puppet characters, the graphic stylization of the film as a whole, and similar considerations. Such a treatment, perhaps very brief, precedes any fuller scripting or other elaboration that may take place.

Storyboard. Since visual emphasis is the key to animation, and sound its close counterpart, the sooner ideas are translated into pictures the better. The "storyboard" provides the continuity of the action, which is worked out scene by scene simultaneously with the animation script. In the storyboard the story is told and to some extent graphically styled in a succession of key sketches with captions and fragments of dialogue, much like a cartoon strip but with much fuller treatment. A feature-length film could easily require a final continuity of several hundred such sketches.

Sketches. Meanwhile, an animation director is also preparing modelling drawings for the principal characters and drawings establishing the backgrounds, or settings, for the film. These begin to indicate the general graphic style and, when colour is involved, the colour scheme and decor to be used. The modelling drawings must indicate the nature and temperament of the characters as well as their appearance when seen from a variety of angles and using a number of characteristic gestures. These will act as guides for the key animators, who with their assistants must bring the figures to dramatic life through the succession of final drawings created on the drawing board.

Synchronization. Animated films are, in effect, choreographed; since mobility involves time, the movements must be exactly timed and so deployed through the right number of successive drawings, like notes in music deployed through bars in a score. When the characters

speak or sing, their lip movements must be synchronized with the words they appear to utter. When sound tracks, both dialogue and music, are prerecorded, the animators have an exact time scheme to follow; if the tracks are not prerecorded, then the "scoring" of the action will control the subsequent timing of the speech and music at recording stage. The timing in either case is predetermined on paper in a workbook, which grades the progression of the animators' drawings frame by frame with the same precision as a musical score. A similar control in the form of a time chart may be created by the director as a guide for the composer. A third control, the so-called dope sheet or camera exposure chart, guides the rostrum cameraman in the frame-by-frame setups and sequence of cels or backgrounds.

Execution. When the exacting labour of animation is underway, difficult moments in the choreography of the figures may be "line-tested"; that is, summarily outlined in pencil, photographed, and tested out on the screen for rhythm and characterization. The key, or senior, animators, draw, or "cartoon," the highlights, or salients, of the movement, perhaps the five or more drawings out of the 24 per second that will give the special edge of liveliness or characterization to the movements. Assistant, or junior, animators close the gaps by completing the intermediate drawings. The smaller the animation unit the greater the burden each artist has to bear in the preparation of final drawings. These drawings, the backgrounds of which remain on drawing paper, are transferred to the cels by specialized artists, who trace the animators' work and paint over it with opaque colouring. The work of tracing and painting can be saved when the animators draw directly on the cels with coloured chinagraph pencils, which they can rub out or correct without harm.

Silhouette and puppet animation

Preparation for silhouette or puppet animation is not dissimilar. Drawings are replaced by solid objects—the flat figures with their jointed limbs to be posed under the rostrum camera or the malleable, three-dimensional puppets with their sets of masks showing different expressions—that must have their carefully phased movements adjusted before each successive frame is exposed. When the picture track and the sound track with speech, sound effects, and music dubbed together are completed under the control of the director and his editor, a "married print" can be made, with the track recorded optically.

Newer techniques. Later developments in technique include the application of the computer to the more routine forms of animation. The Bell Telephone Company first used this technique in the mid-1960s to prepare instructional films in mathematics. Later, computerization made possible films that demonstrated highly complex movements, and advanced forms of precision animation became available for technological demonstration and research. On a humbler technical level, modern rostrum cameras are to a large extent automated. This contributes to the ever-increasing complexity of the medium and the achievement of many special effects. Once, animation could be combined with live action only by complicated laboratory work; now, by a process using "travelling mattes," the camera can record both simultaneously.

USES OF ANIMATION

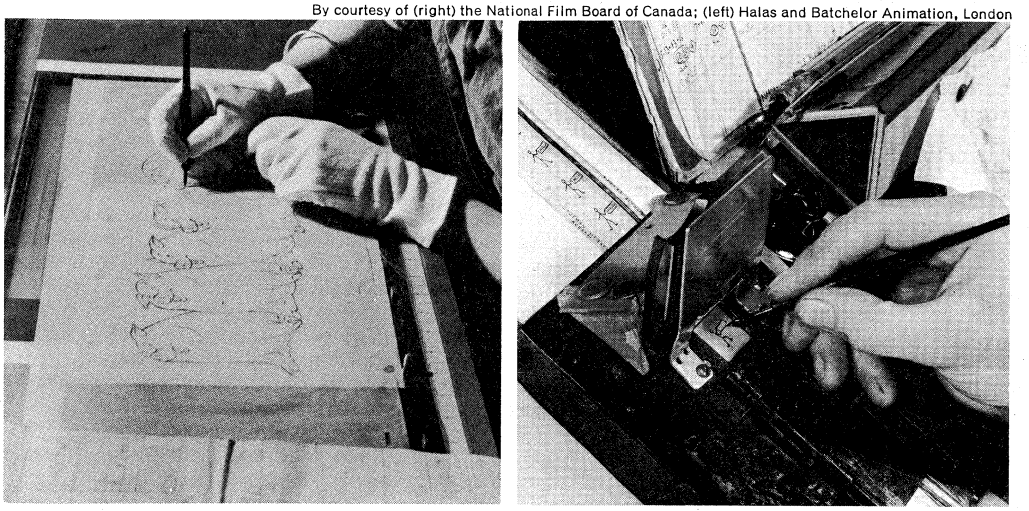
The general public tends to think of animation only in terms of the cartoon "comic" whose tradition goes back to such knowing animal characters as Felix the Cat, by the American animator Pat Sullivan, which was popular in the 1920s. The cartoon by its very nature stylizes and caricatures human and animal behaviour, often presenting animals in anthropomorphized form in the manner of Aesop's fables.

Early animated cartoons

Moving drawings had been projected onto a screen by a series of mirrors and lanterns as early as 1877 by Émile Reynaud in Paris. Animated cartoons on film date from 1906, when J. Stuart Blackton made *Humorous Phases of Funny Faces* for Vitagraph. That company in 1911 began Winsor McCay's *Little Nemo* series, the first of many based on newspaper comic-strip characters. Others included *Mutt and Jeff*, *The Katzenjammer Kids*, *Krazy Kat*, *Happy Hooligan*, and *Bringing up Father* in the late 1910s and 1920s. These silent cartoons were usually shown in movie houses to live music. In France, Émile Cohl's work with film was begun with *Fantasmagoria* (1908).

Disney and other storytellers. The fuller development of the entertainment cartoon short, and later of the animated feature film, such as Walt Disney began to produce in the 1930s, belonged essentially to the era of sound films. The first important sound feature made with puppets was *The New Gulliver* (1935), in which the Russian animator Aleksandr Ptushko combined puppets and a live actor. Disney's work dated from 1923 with series based on such themes as *Alice in Wonderland* and *Oswald the Rabbit*. In 1928 he produced the first animated sound cartoon, *Steamboat Willie*, which launched the career of Mickey Mouse; his first colour cartoon was *Flowers and Trees* (1932). His feature-length *Snow White and the Seven Dwarfs* (1937) contained about 477,000 photographed drawings; its sound track was dubbed into 13 foreign languages. The graphic style of these films followed at first that of the cartoon strips of the period—hard outlines, little or no modelling, and the flat application of paint when colour was introduced—and the medium achieved tonal subtlety and sophistication only at a much later period.

By courtesy of (right) the National Film Board of Canada; (left) Halas and Batchelor Animation, London



Techniques of animation art.
(Left) An artist draws onto a cel that is to be coloured and combined with background to be photographed as one frame; from Halas and Batchelor's *Animal Farm.* (Right) Norman McLaren, noted Canadian animator, draws directly onto film.

Use of serious subject matter

The symbolic political fable *L'Idée* (1934), by the Austro-Hungarian animator Berthold Bartosch, represented a rare use of animation to present a serious, as distinct from a comic or legendary, subject; the approach was first achieved at feature length by the British husband-wife team of John Halas and Joy Batchelor in their adaptation of George Orwell's novel *Animal Farm* (1954). More recently, cartoons have been used on occasion in certain Communist countries to make social statements in the guise of highly stylized anecdotes or fables. Some of the most advanced, most inventive work of this kind has been done in such countries as Yugoslavia, Czechoslovakia, and Poland. In the U.S., highly individual styles were pioneered in the 1940s by the United Productions of America (UPA) group under the direction of Stephen Bosustow. Among their creations were the series that featured Mr. Magoo and Gerald McBoing Boing.

Later, various kinds of graphic fantasy found expression through animation, and the older style of cartooning associated with Walt Disney was completely overtaken by artists using a diversity of contemporary graphic styles, often extremely elaborate. The work of Jiří Trnka and Kařel Zeman exemplified the long-established Czech leadership in animated puppetry. Trnka's *Old Czech Legends* (1951) and *The Good Soldier Schweik* (1955) use stylized puppets in naturalistic settings to dramatize social satires, folk legends, and classics. Zeman's work has both puppets and live actors moving within real sets.

Abstract animation. Inked or painted directly onto the film, abstract animation was pioneered in the 1920s with the individual and collaborative work of the German Hans Richter and the Swede Viking Eggeling, and continued in the 1930s with the films of Len Lye, a New Zealander also known for his abstract sculpture. Abstract film can be realized through orthodox animation techniques, as in the work of the painter-animator Oskar Fischinger in Germany that was choreographed to classical music, a style used by Disney in *Fantasia* (1941). But the specific form of abstract or semiabstract animation of which Norman McLaren of Canada became the principal exponent was achieved by painting directly on film rather than photographing drawings. McLaren's films encompassed the widest range of experimental techniques, including the so-called pixilation—a technique that makes human movements look like those of animated automatons—of *Neighbours* and the beautiful evolving multiple images of ballet dancers in *Pas de deux*. The mobile patterns characteristic of McLaren's earlier films were frequently accompanied by a score composed of sounds that originated as a series of elaborate symbols stencilled directly onto the sound track, as distinct from conventional sound recording.

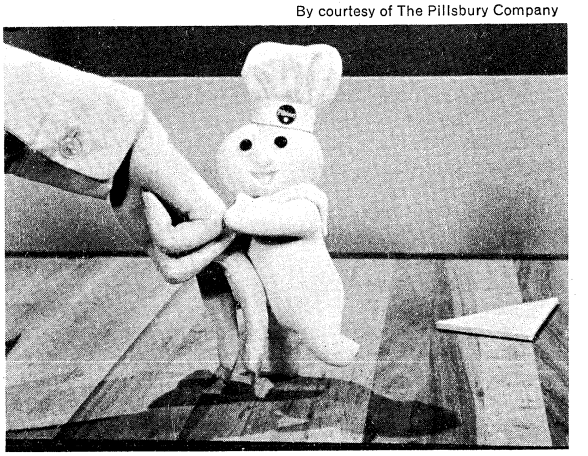
Instruction and promotion. Animation is constantly used for instruction and for public relations and other forms of promotion. On the lighter side it can, if conceived imaginatively, make a dry or difficult subject—*e.g.*, a topic of the day such as road safety—more interesting to the public. Although in the 1960s the cartoon film appeared to lose ground as pure entertainment except in the Communist countries, on the whole it gained ground in public relations and, above all, in the field of instruction. One of the more important extensions has been the introduction into classrooms and lecture halls of the 8mm casette-loaded loop projector. This device enabled the instructor to show students a brief, silent film illustrating some immediate point; he can repeat it at will or freeze any frame while commenting more fully on its pictorial content. Film loops of this kind are "published" and sold along with the textbook; the older style instructional film comprising one or more reels each about ten minutes in length can still be rented to provide broader background information or to elaborate on a wider scale, with commentary, some major aspect of a subject. Though live-action loops can be just as relevant to such integrated audiovisual demonstration, animation has frequently proved to be clearer, more precise, and more analytical for technical or scientific exposition. The mobile, coloured diagram or schematic drawing can show clearly, quickly, and repeatedly such processes as the actions of the heart or the movements of the internal combustion engine, and animation may be superimposed upon, or alternate with, a live-action image. By animation the normal speed of a movement can be slowed down to achieve greater clarity. Animated maps can demonstrate, for example, the movements of population, of imports and exports, the disposition of armies in battle. Even company balance sheets have been explained entertainingly by means of animation.

Film loops

The human figure has a "universal" quality when schematized with slight or extreme caricature. Promotional films often achieved wider international acceptance by advancing their arguments through this type of figure, which was not tied down to any single nationality. Thus industrial films on oil, banking, computerization, and related subjects were able to promote their subject on an international, as distinct from a merely national, level.

Animation on television. The use of three-dimensional, puppet, or object animation was limited in the instructional film because pure graphics are more fluid. The animation of objects, a process known as pack animation, was frequently used in television commercials. Animation and accompanying commentary can make advertising points at lighting speed: live action, unless the figures are pixilated, is necessarily slower. Animation is more abstract, more schematized, and, if desired, more amusing by virtue of its visual sleight of hand, its technical proficiency, and the frank artificiality of its image in time and space.

By courtesy of The Pillsbury Company



Puppet animation.
The puppet is manipulated slightly from frame to frame to give the illusion of continuous movement in this television advertisement.

Entertainment cartoons on television

In addition to the continuous use of animated commercials, television sponsored many series of new entertainment cartoons, most directed to children. These were at first reminiscent graphically of the old, hard-outline cartoon films of the silent screen: they had to be as cheaply made as possible, and every shortcut in the animation process was used. Later, the refinements of style and technique familiar to the better big-screen animation began to appear in television animation, though the image for the smaller screen had to be kept as simple and visually readable as possible. As the cartoon film became less important in the theatre, the television of all nations showed increasing interest in the medium. And computerization appeared to offer the possibility of lowering the cost of high-quality animation for entertainment.

In the 1970s closer links seemed likely between the automated processes of graphic animation and the work of newer workers in all visual arts. There was little doubt, however, that these artists would come to regard the animated film as only one of many techniques available for the expression of the rapidly expanding interest

in kinetic sculpture and other forms of art that incorporated mechanized movement.

BIBLIOGRAPHY. R.D. FEILD, *The Art of Walt Disney* (1942), a richly illustrated account of Disney's films from the 1920s to the early 1940s; JOHN HALAS and ROGER MANVELL, *The Technique of Film Animation*, 2nd ed. (1968), a fully illustrated account of animation techniques and production procedure for professionals; *Design in Motion* (1962), a survey of design in international animation at a period having a wide variety of experimental styles and techniques, fully illustrated, including examples in colour; and *Art in Movement: New Direction in Animation* (1970), the link between contemporary animation design and techniques and kinetic art forms, together with outstanding work by animators throughout the world, including many illustrations and examples in colour; ROGER MANVELL, *Animated Film* (1955), a full account, illustrated by working drawings, storyboard sketches, workbook, and other technical diagrams, of the production of the animated feature *Animal Farm;* RALPH STEPHENSON, *Animation in the Cinema* (1967), a brief but detailed and authoritative history of animation.

(R.M.)

Animism

Animism is the belief in innumerable spiritual beings concerned with human affairs and capable of helping or harming men's interests. Animistic beliefs were first competently surveyed by Sir Edward Burnett Tylor in a work, *Primitive Culture* (1871), to which is owed the continued currency of the term. While none of the "great" religions of the world is animistic (though they may contain some animistic elements), most of the "little" religions, those of the tribal (or "primitive") peoples, are. For this reason an ethnographic understanding of animism, based on field studies of the tribal peoples, is no less important than a theoretical one, concerned with the nature or origin of religion.

IMPORTANCE IN THE STUDY OF MAN AND RELIGION

Definition and limits of the term animism

The term animism denotes not a single creed or doctrine but a view of the world consistent with a certain range of religious beliefs and practices, many of which may survive in more complex and hierarchical religions. Modern scholarship's concern with animism is coeval with the problem of rational or scientific understanding of religion itself. After the age of exploration, Europe's best information on the newly discovered peoples of the Americas, Africa, Asia, and Oceania often came from Christian missionaries. While generally unsympathetic to what was regarded as "primitive superstition," some missionaries in the 19th century developed a scholarly interest in beliefs that seemed to represent an early type of religious creed, inferior but ancestral to their own. It is this interest that was crystallized by Tylor in *Primitive Culture*, the greater part of which is given over to the description of exotic religious behaviour. To the intellectuals of that time, profoundly affected by Darwin's new biology, animism seemed a key to the primitive mind—man's intellect at the earliest knowable stage of cultural evolution. Present-day thinkers consider this view to be rooted in a profoundly mistaken premise. All contemporary cultures and religions are regarded as comparable in the sense of reflecting a fully evolved human intelligence capable of learning the arts of the most advanced society. The religious ideas of the "stone-age" hunters interviewed in this century have been far from simple.

Since the "great" religions of the world have all evolved in historic times, it may be assumed that animistic emphases dominated the globe in the prehistoric era. In societies lacking any doctrinal establishment, a closed system of beliefs was less likely to flourish than an open one. There is, however, no ground for supposing that polytheistic and monotheistic ideas were excluded. But what is plain today—that no historically given creed has an inevitable appeal to the educated mind—had scarcely gained a place in scholarly argument 100 years ago.

THEORETICAL ISSUES

Tylor's theory of animism. For Tylor the concept of animism was an answer to the question, "What is the most rudimentary form of religion which may yet bear that name?" He had learned to doubt scattered reports of peoples "so low in culture as to have no religious conceptions whatever"—he thought religion was present in all cultures, properly observed, and might turn out to be present everywhere. Far from supposing religion of some kind to be a cornerstone of all culture, however, he entertained the idea of a pre-religious stage in the evolution of cultures and believed that a tribe in that stage might be found. To proceed in a systematic study of the problem, he required a "minimum definition of religion" and found it in "the Belief in Spiritual Beings." If it could be shown that no people was devoid of such minimal belief, then it would be known that all of mankind already had passed the threshold into "the religious state of culture."

But, if animism was ushered in as a "minimum definition," it became the springboard for a broad survey. Although anthropology in Tylor's day was mainly an "armchair" science, through wide and critical reading he had developed a good sense for what was credible in the ethnographic sources of his day. He assembled an array of cases and arranged them in series from what seemed to him the simplest or earliest to the most complex or recent. In this way he taught that religion had evolved from a "doctrine of souls" arising from spontaneous reflection upon death, dreams, and apparitions to a wider "doctrine of spirits," which eventually expanded to embrace powerful demons and gods. A fundamental premise was "that the idea of souls, demons, deities, and any other classes of spiritual beings, are conceptions of similar nature throughout, the conceptions of souls being the original ones of the series." Tylor asserted that men everywhere would be impressed by the vividness of dream images and would reason that dreams of dead kin or of distant friends were proof of the existence of souls. The simple belief in these spiritual beings independent of natural bodies would, he thought, expand to include more elaborate religious doctrines, accompanied by rites designed to influence powerful spirits and so control important natural events.

While Tylor offered no special theory for this expansion and so avoided most of the traps of early social evolutionism, he taught that cultures moved, though not along any single path, from simpler to more complex forms. The direction of movement was shown by the survival of animism in muted but recognizable forms (including most "superstitions" and many expressions such as "a spirit of disobedience" or common words like "genius") in the advanced civilization of his own day. This "development theory" he championed against the so-called degradation theory, which held that the religion of remote peoples could only have spread to them from centres of high culture, such as early Egypt, becoming "degraded" in the process of transfer. Tylor showed that animistic beliefs exhibit great variety and often are uniquely suited to the cultures and natural settings in which they are found.

In retrospect, Tylor seems more balanced in his judgments than later writers who constructed the problem of "minimal religion" in a narrower frame. Tylor's greatest limitation was self-imposed, since he narrowed his attention to what may be called the cognitive aspects of animism, leaving aside "the religion of vision and passion." Tylor took animism in its simplest manifestation to be a "crude childlike natural philosophy" that led men to a "doctrine of universal vitality" whereby "sun and stars, trees and rivers, winds and clouds, become personal animate creatures." But his cognitive emphasis led him to understate the urgent practicality of the believer's concern with the supernatural. Tylor's men are "armchair primitives" (the creatures of "armchair anthropologists"), not real men caught in the toils of discord, disease, and fear of perdition.

Counter theories: totemism and magic. Tylor thought the idea of the human soul must have been the elementary religious idea and the model for all other supernatural beings. Later scholars, responding to evidence of simpler beliefs that yet entailed a properly religious awe toward the sacred, began to debate the probability of a

The search for a pre-animistic stage

"pre-animistic stage" of theological evolution. Corresponding to this turn in religious studies was a shift in anthropology toward a concern with "primitive thought" and, in particular, the explanation of religion as intellectual error. Émile Durkheim, a French sociologist, in his *Elementary Forms of the Religious Life* (1915), held that religion originated in totemism, conceiving that identification with a totem animal could result from an irrational projection of men's expectations of security in the bosom of society. He thought such collective projections were more solidly based in the human condition than the "hallucinations" (dreams) that Tylor had supposed must lead to the ideas of soul and supernatural being. Durkheim has been criticized for not seeing totemism as only one animistic cult among many, with special implications as an organizational schema but not "elementary."

English and German theorists conceived the invention of religion in more pragmatic terms: attempting to extend his control of nature beyond the limits his crude science imposed, man had invented supernatural power—magic. The most prolific scholar of this persuasion was Sir James G. Frazer, who argued in his massive work *The Golden Bough* (1890–1915) that "the magic art" had arisen as a pseudo-science, probably had achieved universality before the emergence of religion, and was more firmly rooted than religion in men's beliefs. He thought intelligent men had become disillusioned with earthly magicians and had invented infallible ones—gods. Frazer's work ranged over classical mythology and savage custom without distinction. Finding parallel traditions everywhere, he compiled a massive testament to the psychic unity of mankind. The myriad structures of both magic and religion that he surveyed all could be reduced to transparent intellectual error. The apparent mystery of religion was virtually explained away, for, if magic had become man-centred religion and religion god-centred magic, there was little to choose between them.

The predominance of human over supernatural spiritual powers

Since Frazer accepted the common notion that the sign of religion is man's humility before the gods and held that magic put man in the ascendancy, it followed that wherever the supernatural beings could be tricked, bribed, or otherwise mastered the system of beliefs was magical. This obscured Tylor's clear "minimum definition of religion" and threw an odd light on what he had called "lower animism," the belief in spirits that men with ritual knowledge can master. Any self-confident ritual act—for example, the Eskimo hunter's ritual control of game spirits or the shaman's cure of a grave affliction—had become magical and so transparently egoistic. The result for an ensuing generation of anthropologists was loss of focus upon the religions encountered in the field. Tylor had found animistic beliefs generally devoid of ethical content even when centred in men's urgent needs. Frazer, interjecting the image of the primitive magician with illusions of unlimited power, made it difficult to grant animistic religion even Tylor's minimum of dignity.

Frazer had identified Melanesia as an area in which, by his terms, magic dominated over religion. An admirer, Bronisław Malinowski, was able to accomplish a thorough ethnographic study of the Trobriand Islanders in Melanesia during World War I, and it was Malinowski who dominated European ideas about the intimate life of "primitive man" in the following decades. Viewing his islanders within the frame of ideas Frazer had provided, Malinowski pictured them as secular in outlook. In numerous works on the Trobriand Islanders, published through the 1920s and 1930s, there was to be scarcely a mention of religion as such. The belief in spirits appeared only as mythical background to magical practices connected with gardening and seafaring and with a ceremonial cycle in which the competition for prestige was dominant. The effect was to reduce religion to its pragmatic and social aspects, thus de-emphasizing the very peculiarities of human belief and experience that first attracted men such as Tylor to the study of "primitives."

ANIMISTIC PHENOMENA IN THEIR SOCIAL CONTEXTS

While it is futile to seek cases of animism in "pure," "minimal," or "elementary" form, some social contexts are undeniably simpler than others, and it may be tempting to suppose that the religions found in those contexts would follow suit. On that principle, however, nomads such as the Australian Aborigines might be supposed (as they were supposed by Durkheim) to enjoy an uncomplicated religious life, but this is emphatically not the case. What complicates Australian religions is an elaborate ceremonialism not usually found in nomadic societies. Ceremonialism generally can be treated as an emphasis in the area of expressive behaviour, usually consistent with the animistic world view and unlikely to displace it. While it is an emphasis most common among agriculturists, its presence among nomads is by no means confined to Australia. Though there is no reason to suppose that ceremony is of any more recent origin than any other way of expressing man's relation to the spirit world, animistic religions (religious systems in which animism plays an essential role) can be sorted into those with and those without a ceremonial emphasis, and, in this formal sense, the latter are the simpler. The salient characteristic of all animistic religions is their particularism, a quality opposite to the universalism of the "great religions," which conceive man as subject to global powers and personal destiny.

Ceremonialism and particularism

Particularism is evident in the number and variety of spirits recognized and in the peculiar scope attributed to each. The pre-Christian Lapps of Scandinavia have sometimes been called fetishists because they propitiated nature spirits as well as personally named gods and demons. The nature spirits were generally benevolent and always localized. They could be addressed in particular objects, such as stones or posts, which men would set up in likely places. The few personally venerated spirits (gods) were identified with thunder, sun, moon, hunting, childbirth, and the winds. Evil spirits might be incarnate in animal or monstrous forms and could cause disease or other misfortune. In Minnesota–Ontario the Ojibwa world was animated by a great number of eternal spirits (manitous), all of about equal rank, represented in trees, food plants, birds, animals, celestial bodies, winds, and wonders of every description. Besides these esteemed spirits were other categories, which were dreaded: ghosts, monsters, and the windigo (a crazed man-eating ogre), who brought madness (a cannibalistic psychosis). The list of creatures, places, attributes, and events that could be treated as totems in Australia would be quite similar. The Buryat of Lake Baikal in Siberia, living on the fringes of empire (Mongolian and Chinese), developed an elaborate social order and viewed the spirit world as the twin of their own, organized in the same way into noble, commoner, and slave ranks. At death a man passed over to the other world, assuming his proper rank and acquiring fresh power over men, which he might exercise well or ill in accordance with his character in life. Evil men, as it were, became devils and great men gods.

In particularistic religions there is a range of spirits, from sojourning ghosts and mortal witches to perennial beings, whose natures and dispositions to man are attributed by categories (*e.g.*, mermaids and leprechauns are both usually pictured as irresponsible), but in action individual spirits are independent of one another. If some spirits may be called gods, they do not constitute a ruling pantheon, for men do not conceive that any supernaturals enjoy comprehensive control of events. In animism, spirits represent particularistic powers and must be handled accordingly. Typically, men's primary emphasis is on avoidance of trouble, and this is the meaning of the many taboos and propitiatory observances, of an almost mechanical nature, that abound in some societies. When trouble is at last encountered, the responsible witch, demon, or disgruntled spirit must be identified, and this is the task of the diviner. The cure may rely upon ritual cleansing, propitiation, or even the overpowering of the malevolent force through supernatural counteragency—the specialty of the shaman. Judging that an animal will not mind being killed if it is not offended ritually, Eskimos take various precautions before, during, and after the hunt. The rationale lies in the belief that animal spirits exist independent of bodies and are

The primary emphasis: avoidance of trouble

reborn: an offended animal will later lead his companions away so that the hunter may starve. If, in spite of their precautions, men are left without game, a shaman may be called to discover the transgression that has offended an animal spirit—or perhaps he may find that he must do battle with a malevolent being controlled by a rival shaman willing the community harm.

Ceremonialism, when its emphasis is upon feasting, exchange, and display, may be secular in its emphasis, as is the case in much of Melanesia and New Guinea; or if religious, it may be associated with totemic or ancestral cults, as in Australia or Africa, the expressive emphasis of which is on social ties rather than on the quality of relations between men and the supernaturals. Finally, ceremony may be used directly to dramatize the role of the spirits in men's lives, as it is by the Pueblo peoples of Arizona–New Mexico. At their height, the Pueblo ceremonial cycles were as rich as any in the world. Supernaturals were elaborately impersonated by the dancers, and the human condition was portrayed as one of dependency. But, for all this, particularism was not greatly compromised. The supernaturals were many and were represented in a realistic manner emphasizing their differences from ordinary men. The style was that of mummery and conjuring, consciously put on by grown-ups as a sort of morality play. There was no sense of incongruity in the fact that neighbouring pueblos cultivated other sets of spirits. In some pueblos, separate clan societies had complete charge of the ceremonial calendar and formally controlled communication with the supernatural, even to selecting the member who might be curer in case of an illness. But such a step toward ecclesiasticism in a very small community could not greatly affect its animistic premises, and witchcraft prevailed without the blessing of the ceremonial societies.

Openness of doctrines

When the fullness and versatility of all these religions is considered, without any need to press them into simplified categories or evolutionary stages, it can be seen that openness, not narrowness, of doctrine is a general feature of animism. Wherever it is found, it is a grass-roots religion, not a doctrinaire one imposed from above. Ecclesiasticism may coexist with animism, as in China or Burma, where there are no pre-eminent gods whose universal claims presuppose mastery of the whole supernatural world. But the most likely context of animism is an uncentralized social order in which secular power is not developed and each local settlement is at the focus of its own world.

THE ANIMISTIC WORLD VIEW

Part of the conceptual difficulty experienced both in anthropology and in the history of religions, when animism is to be placed among other systems of belief, springs not from the early association of animism with a speculative theory of religious evolution but directly from the huge variety of animistic cults. As a category, Tylor's concept is more general than either polytheism or monotheism, and its meaning is harder to delimit—the word applies broadly to most of the "little religions" but suggests nothing of their varieties. For this reason, much use is made of subordinate labels, such as shamanism, totemism, or ancestor propitiation. These cults do not, in any case, constitute the whole religion of a people. They are, however, institutions that are not bound to one culture area —an Australian totemic cult does bear a "family resemblance" to an African one, though their differences also are many. Shamanism, with its reliance on ecstasy, is found from Greenland to Mysore, and the propitiation of ancestors is not restricted to Africa and the Far East. It has long been recognized that the frequent recurrence of institutions fitting a certain pattern implies that there is a radically limited number of possible patterns, and, in this case, the premises of animism evidently have imposed the limitation. Animism attributes importance to categories of supernatural being whose individual members are attached to particular places and persons or resident in particular creatures and are autonomous in their dealings. In such a system, each human encounter with the supernatural must work itself out as a distinct episode. Even where ceremonialism emphasizes an enduring moral relationship to certain supernaturals, men are likely to conceive of alternative powers to which they might seek at need. In a crisis, loyalties may shift: in West Africa, gods have been sold to neighbouring villages, and, in Melanesia, a vision of European trade goods has inspired a series of new millenarian cults. The quality of openness lends itself to change and eclecticism, hardly ever to religious chauvinism.

A communication with supernatural beings about practical life needs

Animistic creeds have in common an undertaking on the part of men to communicate with supernatural beings, not about metaphysics or the dilemmas of the moral life but about urgent practicalities: about securing food, curing illness, and averting danger. It is characteristic that genuine worship of a supernatural hardly is found. Creator gods often appear in myth but not in cult. In ancestor cults the most recently dead are the most vividly conceived—the original clan ancestor, for all his symbolic importance, is remote both from men and from godhead. If animistic spirits anywhere exercise authority, they do so in particularistic, even egoistic fashion, sanctioning men for ritual neglect or breaking taboos, not for acts of moral neglect or secular offense. Animistic religions do not readily coalesce with systems of political authority and probably do not favour their development. When it is asked whether the association of animism with smaller and simpler societies proves it the natural (original) religion, the answer can only be that it is not known (and perhaps not knowable) what a prehuman or panhuman religion would be like. The problem is as difficult as reconstructing protohuman speech. If religion is taken as a pattern of serious relations between men and supernaturals, then societies devoid of religion have not been found, and it may perhaps be concluded that religion is usually close to the vital centre of a culture, where the credibility of institutions is determined. The view of all nature as animated by invisible spirits—be it shades, demons, fairies, or fates—with which men could cope in meaningful ways may belong to the past, but philosophies that attribute powers of initiative and responsiveness to nature have not gone out of currency. The lesson of the study of animism is perhaps that religion did not arise, as some of Tylor's successors believed, out of *Urdummheit* ("primal ignorance") or delusions of magical power but out of men's ironic awareness of a good life that they are unable, by earthly means, to grasp and hold. Animistic beliefs have everywhere engaged men's susceptibility to private vision and enabled them to cope with it at the level of accepted meaning.

BIBLIOGRAPHY. The original and classic source is E.B. TYLOR, *Primitive Culture*, 2 vol. (1871 and later editions, reprinted 1958). For the "pre-animism" thesis, see R.R. MARETT, *The Threshold of Religion*, 4th ed. (1929); and SIR JAMES G. FRAZER, *The Golden Bough*, 3rd ed. rev., 13 vol. (1955; one-volume abridged ed., 1922; reprinted 1955). An extended analysis and review of theories may be found in ADOLF E. JENSEN, *Mythos und Kult bei Naturvölkern*, 2nd ed. (1960; Eng. trans., *Myth and Cult Among Primitive Peoples*, 1963); or for a short review, see E.E. EVANS-PRITCHARD, *Theories of Primitive Religion* (1965). Still to be recommended for an overview is ROBERT H. LOWIE, *Primitive Religion*, new ed. (1970). An ample source for the animism of China is the 5th volume of J.J.M. DE GROOT's monumental study, *The Religious System of China* (1907, reprinted 1969); and for Burma the best introduction is M.E. SPIRO, *Burmese Supernaturalism* (1967). Ethnographic studies of animistic peoples abound; for an annotated bibliography, see WILLIAM A. LESSA and EVON Z. VOGT, *Reader in Comparative Religion*, 2nd ed., pp. 640–645 (1965). A provocative consideration of the relation of animism to modern world views is presented in JACQUES MONOD, *Le Hasard et la nécessité* (1970; Eng. trans., *Chance and Necessity*, 1971).

(G.K.P.)

Ankara

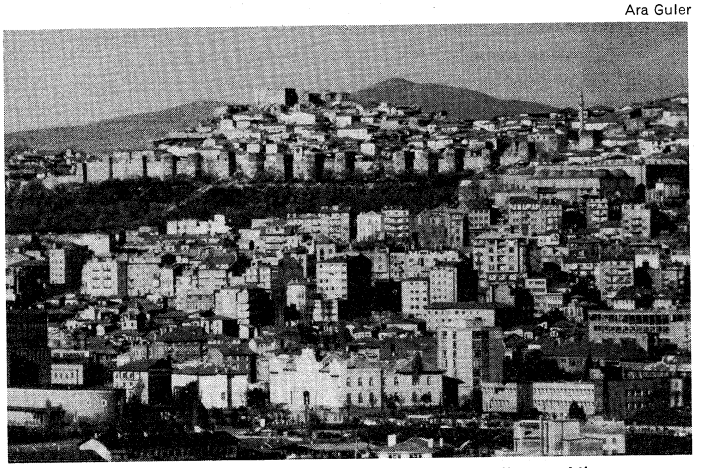
Ankara is the capital of the Republic of Turkey and the administrative centre of Ankara *il* (province). The city, at the northern edge of the central Anatolian Plain and about 125 miles (200 kilometres) south of the Black Sea coast, lies close to the confluence of the Hatip, İnce Su, and Çubuk streams. A railroad leads to Istanbul 360

miles west, and a modern paved highway also connects the city with Istanbul, via Bolu. In 1969 Ankara's population was 1,157,000.

Origin of the city's name

Controversy surrounds the origin of the name of Ankara. In Greek and Latin it was Ancyra. Legends connect the name with the settling of the area at a place where an anchor was found, or else with an anchor supposedly captured from the Egyptians in the 3rd century BC. A more convincing explanation is that the name comes from a Phrygian word meaning "gorge" or "ravine." Arab geographers called the city Qal'at as-Salāsil (Fortress of the Chains). From the 11th to the 13th century the Seljuq Turks called the city Engüriye, but under Ottoman rule (from the 14th to the 20th century) the name was Anqarah. In 1930, the name Ankara officially replaced the Europeanized name Angora, associated with Angora mohair, which is produced by the white, long-haired goats of the region.

Ara Guler



Ankara, showing the walled citadel on the hilltop and the shopping and residential district below.

History. While the date of the city's foundation is uncertain, archaeological evidence indicates habitation at least since the Stone Age. Twentieth-century excavations have revealed that a thriving Phrygian town was located in the area at the end of the 2nd millennium BC. Alexander the Great conquered Ankara in 334 BC, and in the 3rd century BC the town served as the capital of the Tectosages, a tribe of Galatia (the ancient name for the region around Ankara). Later, in the 1st century BC, the city came under the control of Mithradates VI the Great, ruler of the kingdom of Pontus on the Black Sea coast. In 25 BC Ankara was definitively incorporated into the Roman Empire by the emperor Augustus; a number of great buildings were then constructed. In AD 51 the Christian Apostle St. Paul visited Ankara. In his subsequent Letter of Paul to the Galatians, he reprimanded the Christian community he had founded there for its continuing attachment to pagan rites.

As a city of the Byzantine Empire, Ankara was attacked by both the Persians and the Arabs. In 620, a Persian king of the Sāsānid dynasty, Khosrow II Parvīz, took Ankara, but he had to abandon it in 627. The Arabs held the city for a short time in 654. Hārūn ar-Rashīd, the 'Abbāsid caliph of Baghdad, besieged Ankara in 806, as did his son al-Mu'taṣim in 838. Arab threats to the city continued in the 10th century, but by the 11th century the main danger to Ankara's Byzantine rulers was posed by the Turks.

The Seljuq era

In about 1073, Ankara fell to the Seljuq Turks, but the crusader Raymond IV of Toulouse drove them out again in 1101. The Byzantines, however, were unable to maintain their control. Ankara then became a bone of contention between the Seljuqs and their rivals among the Turkish frontier lords. After 1143, Seljuq princes fought among themselves for possession of the city.

After the collapse of the Seljuq state early in the 14th century, Ankara came under the sway of the Il-Khans, a Mongol dynasty that ruled in Persia. With the decline of Mongol power, Ankara then came under the domination of the Eretnids of Sivas (east of Ankara), one of a number of Turkish groups contesting for power in Anatolia. In the city itself, authority was in the hands of the Ahis (Akhis), an organization of influential merchants and guild leaders.

In 1356 the city was captured by Orhan (Orkhan), the second sultan of the Ottoman dynasty, and became a part of the Ottoman domains in 1360. Ankara was besieged during the Anatolian campaign of Timur (Tamerlane), the famous 14th- and 15th-century Tatar conqueror, and was the scene of a Turkish defeat in 1402. In 1403 it again became subject to Ottoman rule and, in subsequent centuries, became a commercial centre on the caravan route to the East.

After World War I, Mustafa Kemal Atatürk, the Turkish nationalist leader, made Ankara the centre of the resistance movement against both the government of the Ottoman sultan and invading Greek forces; he established his headquarters there in 1919. The Turkish Grand National Assembly declared Ankara the capital of Turkey in 1923.

The contemporary city. *Site and environment.* The city is situated at the foot and along the slopes of a hill whose summit rises some 3,200 feet above sea level but only 400 to 500 feet above the high Anatolian Plateau. North, south, and east of the city are mountain ranges whose highest peak, Elma Dağı to the south, rises to slightly more than 6,000 feet above sea level.

Because Ankara is encircled by mountains, its atmosphere is relatively dry. On the average there are 98 rainy and 25 snowy days a year. January and February are the rainiest and snowiest months. January is also the coldest month, the lowest temperatures averaging about 28° F (−2° C). July and August are the warmest months, the highest temperatures averaging 86° F (30° C), but because of the elevation the evenings are cool.

Ulus, or the old city

The city plan. The old city, which grew up on the slope around the citadel, is known as Ulus and contains Ankara's commercial centre, with shops, banks, and offices. Narrow winding streets with wooden and mud–brick two-story houses are characteristic of the older residential areas high up the hill.

Ankara, however, underwent rapid development after its designation as the capital of Turkey. The nucleus of the new city is Yenişehir, which developed to the southeast. It has broad avenues, smart shops, multistory hotels, theatres, cinemas, restaurants, apartment buildings, and governmental offices. Atatürk Bulvarı (Atatürk Boulevard), the main avenue of the new city, becomes "embassy row" where it begins to be flanked by the impressive compounds of the foreign embassies. Other important districts, all to the south of Yenişehir, are Bahçelievler (distinguished by neat private homes with well-kept gardens), Kavaklıdere (an affluent residential district of apartment houses and shops on the slopes of Çankaya Hill), and Çankaya (containing the presidential palace), and, to the southwest, Maltepe (at the foot of the hill on which stands the impressive Mausoleum of Atatürk).

Transportation. Today, as in antiquity, Ankara is an important crossroads for trade and forms a major junction in Turkey's road network. In 1967 the *il* (province) of Ankara had about 850 miles of state highways, of which about 430 were asphalted. The city has a large bus terminal; more than 30 bus lines dispatch some 400 buses daily. Licensed motor vehicles in the city totalled about 55,000 in 1967, and traffic jams had become frequent.

The railway reached Ankara in 1892. The city is on the main east–west rail line across Anatolia, from which branch lines extend northward to the Black Sea and southward to the Mediterranean. Esenboğa Airport, 16 miles from Ankara, is international; in 1969 it was handling an average of 45 airplane arrivals a day and 490,000 passengers a year. Türk Hava Yolları (THY), the Turkish national airline, provides service to Europe, as do a number of foreign airlines. There are also four military airports in the *il* of Ankara.

Within the city itself transport services are provided by motor buses and electric trolleybuses, as well as by the

characteristic *dolmuş* taxis (*i.e.*, taxis operating on fixed routes with set fares).

Architecture. The architecture of the city reflects its varied history. Remains from the Roman era include a bath, the Column of Julian, and the Temple of Roma and Augustus. Byzantine remnants are represented by the citadel and by a cemetery. The Alâeddin Mosque near the walled citadel—square with one minaret—dates from the Seljuq era. Ottoman buildings are numerous and include the Hacı Bayram Cami (1429); the Mahmud Pasa market building (1464), which has been converted to house the Hittite Museum; and the Mehmed Pasa Market, a two-storied 15th-century building that now houses the Archaeological Museum.

The new city is built mostly of cement and stone. Its most striking buildings are the huge Atatürk Mausoleum complex and the high-rise hotel and office building that dominates Kızılay Square in downtown Ankara.

Steady population increase

Demography. When Atatürk reached Ankara in 1919, the population numbered between 25,000 and 30,000. Since then the population has grown steadily, reaching 120,000 in 1935 and 290,000 in 1950. By 1965 the city's inhabitants totalled 906,000, while the entire provincial population of Ankara numbered over 160,000. In 1970 the population of Ankara's metropolitan area was 1,440,000, and that of the entire province (including the city) was 2,023,000. Males make up 60 percent of the city's residents. Approximately half the inhabitants are of age 24 or under, a quarter are in the age group from 25 to 45, and the remaining quarter are more than 45 years of age. In the 1965 census, 890,000 people gave Turkish as their mother tongue; among the other listed mother tongues, Kurdish, Albanian, and Arabic were prominent.

The city, like the Turkish nation as a whole is overwhelmingly Muslim, with 98 percent of the inhabitants adhering to Islām. Gregorians, Catholics, and adherents of the Greek Orthodox Church make up most of the remaining 2 percent. There were some 600 Jews in Ankara in 1965. Density of population in the city's metropolitan area in 1970 was about 1,100 persons per square mile (424 persons per square kilometre), compared with 170 per square mile for the province as a whole.

Economic life. Government is Ankara's main business. Industry lags, but in 1965 Ankara's industry surpassed that of the Mediterranean port of Izmir, thus making it Turkey's second most important industrial city after Istanbul. Factories for the production of wine and beer, flour, sugar, macaroni products, biscuits, milk, cement, terrazzo (mosaic paving), construction materials, and tractors are well established. Service industries were expanding rapidly in the early 1970s, as also were facilities for tourism. The *il* of Ankara produces wheat, barley, rice, corn, fruit, and vegetables. The Ankara district is famous for its pears, producing about 10 percent of Turkey's pear crop. Livestock, especially sheep, are also important. The city has several open-air markets, which are held in different districts on different days.

Governmental institutions and public utilities. Ankara municipality is governed by a mayor, who is elected for a four-year term, and by a city council also elected for a four-year term. In 1972 the council had 68 members. The city is subdivided into 190 districts. Major administrative concerns are the provision of water, gas, and electricity, as well as civil defense, sanitation, fire, security, and social welfare services. In 1967 electricity was supplied to some 200,000 subscribers, gas to 75,000, and water to 100,000.

Health and safety. Ankara's health facilities have been vastly improved as a result of the establishment of Hacettepe University and its medical school, college of nursing, and related clinics. These supplement the work of municipal, governmental, and private hospitals, as well as of sanitariums, clinics, and dispensaries. The Numune Hospital, Ankara Hospital, and Atatürk Sanitarium together have a capacity of 2,300 beds.

The universities and institutes

Education. Ankara is well served by a variety of educational and cultural institutions. The University of Ankara, founded in 1946, has a number of faculties; in 1967 it was staffed by 1,600 faculty members and had 20,000 students. Hacettepe University, founded in 1967, has a graduate school of education in addition to its medical facilities. The Middle East Technical University (METU) was established in 1956 with the assistance of the United Nations; the language of instruction is English, and students are attracted from many countries of the region. Turkey's national conservatory for training in music (including opera), ballet, and drama is centred in Ankara.

The Gazi Educational Institute is one of the country's leading teacher-training establishments. Some 60 technical institutes and secondary schools in 1967 had a combined enrollment of more than 80,000 students.

Cultural life. Ankara is the home of the state theatre, founded in 1949; in 1967 it consisted of six theatres, presenting almost 100 different productions a year. The State Opera and Ballet, also founded in 1949, became part of the state theatre in 1966. The Presidential Philharmonic Orchestra is a spiritual descendant of the Ottoman sultan's orchestra for which the Italian composer Gaetano Donizetti composed music in the 19th century. Ankara boasts a number of private theatres and more than 70 cinemas, including open-air facilities for summer viewing.

The National Library (Mıllı Külüphane), founded in 1946, has more than 500,000 books, in addition to many periodicals, manuscripts, and materials on microfilm. The largest university collection is that of the Faculty of Letters, with some 200,000 volumes. Neighbourhood circulating libraries are expanding rapidly to bring the populace into contact with the world of books.

Interest in Ankara's museums, which present a panorama of Anatolian history, is often heightened by housing them in renovated Ottoman buildings. The most important are the Archaeological Museum (with its world-renowned Hittite collection) and the Ethnographic Museum (with its holdings related to Turkish history, folklore, and art). Part of the Atatürk Mausoleum is devoted to the Atatürk Museum, which displays many of Atatürk's personal effects, including his official cars.

The media. Radio Ankara, founded in 1927, operates on long, short, and medium wavelengths. In addition to its regular programs, it broadcasts in Turkish to Turkish workers abroad in Europe. Programs are also broadcast in 14 languages for home consumption. Television programs also are broadcast. More than 60 newspapers were published in Ankara in the early 1970s; about 20 of these were dailies. Close to 200 magazines and journals are published, a dozen of which are weeklies.

Recreation. Water scarcity and the high lime content of the soil make the development of parks difficult, and those that exist are therefore all the more prized. The main one is the Gençlik Parkı (Youth Park) north of Yenişehir, with its large boating lake. The park in the citadel grounds and Emniyet Park are other recreation sites, as is Ankara Zoo, on Atatürk's model farm.

There is much enthusiasm for sports (especially soccer), and Ankara has more than 80 organized athletic clubs. The 19th of May Stadium, located (together with other major sports installations) in the Sports City between the railroad station and the racetrack, has a capacity of 35,000. The racetrack is also a major attraction.

BIBLIOGRAPHY. The best guide to Ankara in English is ROBERT BOULANGER, *Turkey* (1960). *Ankara il Yilliği, 1967* (1968), is the most detailed description of the city and province, with photographs, charts, and statistics. NURETTIN CAN GULEKLI, *Ankara Rehberi: The Guide of Ankara* (1949), is an early history-cum-guidebook in both Turkish and English. The most reliable population statistics are contained in the Turkish government's publication *Genel Nüfus Sayimi Nüfusun Sosyal ve Ekonomik Nitelikleri 24/10/1965* (Census of the Population, Social, and Economic Characteristics of the Population) (1969). The definitive history of Ankara is still to be written. A thoughtful summary is available in F. TAESCHNER's article "Ankara," in the *Encyclopaedia of Islam*, vol. 1, pp. 509–511 (1960). Excellent works on the pre-Ottoman period include WILLIAM RAMSAY, *The Historical Geography of Asia Minor* (1890, reprinted 1962); PAUL WITTEK, "Zur Geschichte Angoras im Mittelalter," in *Festschrift Georg Jacob* (1932); and CLAUDE CAHEN, *Pre-Ottoman Turkey* (1968). SPEROS VRYONIS, *The Decline of Medieval Hellenism in Asia Minor and the Process of Islamization from the Eleventh*

Through the Fifteenth Century (1971), contains a good deal of information on Ankara dispersed throughout the work.
(N.It.)

An Lu-shan

An Lu-shan was a soldier of mixed Iranian and Turkish descent who rose to be a commanding general on the Chinese frontier and who, in AD 755, rebelled and attempted to found a new dynasty to replace the T'ang dynasty. Though the rebellion was in the end suppressed, the dynasty (618–907), during which China reached great heights of cultural and political development, was seriously weakened both externally and internally. Forces of economic and social change that had been accumulating during the earlier years of power and prosperity were released, and the subsequent period of greater decentralization saw the evolution of new social forms and institutions, which came to fruition two centuries later with the rise of the Sung dynasty (960–1279). For such reasons the rebellion of An Lu-shan is looked upon as a watershed between distinct historical epochs and has an importance for historians that transcends the personalities involved. These were, at the same time, romantic and colourful in themselves and have continued to provide subject matter for popular Chinese fiction and drama to this day.

Early life and career

An Lu-shan was born in AD 703. The family name An was derived from the Chinese name for Bukhara in Sogdiana (present Uzbekistan). The given name Lu-shan is a Sinicized form of the Iranian *rowshān*, "light." An Lu-shan's ancestors belonged to a group of Sogdians who had been incorporated into the Eastern Turks, and his mother was from a noble Turkish clan.

The Eastern Turks, whose ascendancy in Mongolia dated from the 6th century, had been conquered by the Chinese emperor T'ai-tsung at the beginning of the T'ang dynasty but had made themselves independent and were enjoying renewed prosperity at the time of An Lu-shan's birth. The death of their ruler, Qapaghan Qaghan, in 716, however, led to disorder and strife, and the Ans sought refuge in China. Just at that period the frontier policies of the emperor Hsüan Tsung (reigned 712–756) were providing opportunities for men such as An Lu-shan, his cousin An Ssu-shun, and other soldiers of non-Chinese origin to serve in the Chinese armies. Their rise to positions of command was further aided after 736 by the dictatorial rule of China's chief minister, Li Lin-fu, who was unwilling to appoint native Chinese as generals for fear that they would gain prestige that would enable them to rival his own position at court.

An Lu-shan's military career took place on the northeastern frontier in what is now southern Manchuria. The first occurrence of his name in the T'ang annals is under the year 736, when, as a reconnaissance officer, he lost his force through rash conduct and was condemned to death. He was pardoned and, thereafter, rose rapidly in rank, receiving his first independent command in 742. As a military governor he became a political figure, made frequent visits to the capital, and became a personal favourite of the Emperor and his consort, the celebrated beauty Yang Kuei-fei. An Lu-shan, an enormously fat man, was adept at playing the buffoon in order to ingratiate himself. Such was his favour at court that once, three days after his birthday, he was taken into the women's quarters of the palace (wrapped in an enormous baby diaper) and put through a mock ceremony of adoption by Yang Kuei-fei. Indecorous conduct of this kind led to rumours of improper relations between him and Yang Kuei-fei, which have added spice to the later legend, but his position at court depended at least as much on the Emperor himself as on his consort.

By the time of Minister Li Lin-fu's death in 752, An Lu-shan had accumulated three frontier provinces under his command and was the most powerful general in the empire. After the dictator's demise an intense struggle developed between An Lu-shan and Yang Kuo-chung, the cousin of Yang Kuei-fei, who attempted to take over Li Lin-fu's position. Though Yang Kuo-chung could attack and destroy An Lu-shan's supporters at court, he was unsuccessful in his attempts to establish a countervailing military base in the provinces or to undermine An Lu-shan's position in the northeast, and he only succeeded in exacerbating An Lu-shan's suspicions and frustrations. Finally, toward the end of 755, alleging that he had received a secret command from the Emperor to get rid of Yang Kuo-chung, An Lu-shan turned his large, war-hardened army inward and marched on the capital.

An Lu-shan's rebellion

Within a month the rebels had taken the eastern capital, Lo-yang, and at the beginning of the lunar year corresponding to 756 An Lu-shan proclaimed himself emperor of the Great Yen dynasty.

Meanwhile, loyal T'ang forces had been mobilized and had taken up defensive positions in the narrow pass up the Huang Ho leading into Shensi. For six months the rebels were unable to advance. There was great suspicion and rivalry between Yang Kuo-chung and Ko-shu han, the general in charge of the defense of the eastern approaches to Ch'ang-an, capital of China. Fearing a coup against himself, Yang Kuo-chung goaded Ko-shu han into abandoning his defensive posture and moving eastward against the rebels. The T'ang army was routed, and the way to Ch'ang-an was left open. The Emperor hastily abandoned the city and fled westward. At Ma-wei, a small village west of the capital, his guard mutinied, assassinated Yang Kuo-chung, and demanded the death of the Emperor's favourite, Yang Kuei-fei. The Emperor reluctantly consented, and her assassination became the theme of the great poet Po Chü-i's "Everlasting Remorse" and of countless other works of art.

Although An Lu-shan's forces occupied Ch'ang-an, he himself remained behind in Lo-yang. By this time he was seriously ailing, perhaps with diabetes. He was nearly blind and suffered from extreme irascibility, which made his attendants go in constant fear of their lives. At the beginning of 757 he was murdered by a eunuch slave with the connivance of his own eldest son, An Ch'ing-hsü, and others. The rebellion dragged on for several years, first under An Ch'ing-hsü, then under a former subordinate, Shih Ssu-ming, then under Shih Ssu-ming's son, Shih Ch'ao-i. Finally in 763 it officially came to an end with the defeat and death of Shih Ch'ao-i. A major role in the defeat of the rebels was played by contingents sent by the Uighurs, who had replaced the Eastern Turks as masters of the eastern steppes and who were traditional allies of T'ang. By this time the T'ang government was too exhausted to undertake a thorough pacification of the rebel territory, and the surviving rebel generals were allowed to surrender and remain in command of the regions they occupied in return for a nominal submission to T'ang. Other parts of the country fell under the control of T'ang generals who were almost as insubordinate as the former rebels. The second half of the T'ang dynasty and the following Five dynasties period were troubled by a chronic warlordism that only came to an end with the rise of the Sung dynasty in 960.

After his death, An Lu-shan became the object of a cult on the northeast frontier among the non-Chinese soldiers who constituted the bulk of the army and whose aspirations he symbolized. Conversely, the non-Chinese, "barbarian" character of the rebellion was influential in stirring up the xenophobic sentiments of the Chinese that increasingly characterize the second half of the T'ang dynasty, in marked contrast to the receptive, cosmopolitan attitudes of the first half.

BIBLIOGRAPHY. E.G. PULLEYBLANK, *The Background of the Rebellion of An Lu-shan* (1955), places the man and his rebellion in historical context. HOWARD S. LEVY, *Biography of An Lu-Shan* (1960), is a translation of the official biography of An Lu-shan; and R. DES ROTOURS, *Histoire de Ngan Lou-chan* (1962), a scholarly and thoroughly annotated translation from a large collection of source material on An Lu-shan.
(E.G.P.)

Annelida

The Annelida, or segmented worms, constitute a major invertebrate phylum of the animal kingdom. The group numbers more than 9,000 species and includes three classes: the marine worms (Polychaeta), which are divided into free-moving and sedentary, or tube-dwelling,

forms; the earthworms (Oligochaeta); and the leeches (Hirudinea). Annelids, named for the transverse rings, or annulations, of the body, are segmented animals with a body cavity (coelom) and movable bristles (setae). The coelom is reduced in leeches, and setae are lacking in a few specialized forms, including leeches.

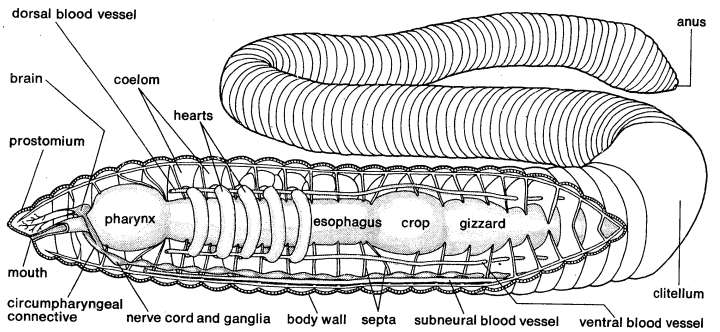


Figure 1: Cutaway view of an earthworm.

Annelids are of importance to man. Large earthworms, or night crawlers (*Lumbricus terrestris*), are cultivated and sold as bait for freshwater fishes and as humus builders in gardens. The sludge worm *Tubifex*, abundant near sewer outlets and thus an indicator of water pollution, is collected and sold as food for tropical fish. Polychaetes play an important role in turning over sediment on the ocean bottom.

The medicinal use of leeches, which dates from antiquity, reached its peak in the first half of the 19th century. The European species *Hirudo medicinalis* formerly was exported throughout the world, and native species also were used. Hirudin, an extract from leeches, is used as a blood anticoagulant.

The estuarine mud flats of Maine and Nova Scotia are the principal sources of the bloodworm (*Glycera dibranchiata*), which is used as bait for saltwater fishes. Reproductive parts of the palolo worm (*Leodice viridis*), which break off and are found in great numbers during the reproductive period, are used as food in the Samoan Islands of the South Pacific.

GENERAL FEATURES

Distribution and abundance. Annelids are found worldwide in all types of habitats, especially oceanic waters, freshwaters, and damp soils.

Most polychaetes live in the ocean, where they either float, burrow, wander on the bottom, or live in tubes they construct; their colours range from brilliant to dull, and some species can produce light. The feather duster (*Manayunkia speciosa*) inhabits the Great Lakes and some rivers of the United States. The polychaetes include more than 5,000 known species, which are about evenly divided between free-moving and sedentary, or tube-dwelling, forms. The oligochaetes number about 3,250 known species. Oligochaetes, including earthworms, burrow into soil; certain small oligochaetes are found in freshwater, and a few are marine, usually inhabiting estuarial or other shallow waters. Leeches, which number about 300 species, inhabit freshwater or humid environments and are carnivorous or parasitic on other organisms—*e.g.*, all marine leeches are parasitic on fish.

Size range and diversity of structure. The length of annelids varies from a fraction of an inch to more than six metres (about 20 feet). The width may exceed 2.5 centimetres (about one inch) in the contracted state. Free-moving polychaetes and earthworms include the largest species. Leeches attain lengths of about 0.4 metre (about 16 inches) in the contracted state.

Polychaetes. The body of free-moving polychaetes (see Figure 3) consists of a head, or prostomium, which may bear two or more eyes; a preoral segment, with such appendages as antennae, tentacles, and palpi (fleshy sensory projections); a trunk divisible into distinct segments; and a tail, or pygidium, which may bear anal cirri (fleshy projections) or plaques and a terminal anus. Each body segment following the second segment (peristomium) usually has paired parapodia; *i.e.*, fleshy, lateral outgrowths used in feeding, locomotion, or breathing. The parapodia, generally prominent in free-moving polychaetes, bear bundles of setae, which can be extended, and acicula (needlelike structures), used for support.

The head of sedentary polychaetes (see Figure 3) may be distinct or indistinct. Forms with a distinct head generally lack head appendages. Branchiae, or gills, which serve for respiration and as food-gathering organs, are well developed in many of the tube-dwelling, or sedentary, forms. Some have tentacles at the anterior (front) end, and gills arise from the dorsal (upper) surface of a few anterior segments; in these species food is gathered by the tentacles; respiration is confined to the gills, which often have two or more eyes. The rest of the body is divided into thoracic and abdominal regions. Parapodia, if present, are generally simple lobes; frequently the setae project directly from the body wall. Many sedentary polychaetes construct tubes made from a substance secreted from cells that comprise the epidermis, or skin. Tubes may consist of calcium carbonate, parchment, or mucus to which sand or fine sediment adheres. The anus is at the posterior tip. Tube dwellers generally have an external fecal groove along which fecal material passes forward. Eyes are occasionally present along the sides of the body or on the pygidium in sedentary forms that do not live in tubes.

Tube construction

Oligochaetes. The body of oligochaetes (see Figure 1) is uniformly segmented and has conspicuous segmental lines. The prostomium is usually a simple lobe overhanging the mouth and generally lacks appendages. The microscopically small eyes are scattered over the body. The clitellum, a saddle-shaped thickening of the body wall, is present at sexual maturity. The anus is at the posterior tip. Setae project directly from the body wall and generally from the ventral (under) surface of the body.

Leeches. Leeches (see Figure 4) have 34 segments; elongation can occur by subdivision of these segments. Leeches, elongated or elliptical in shape, have a small sucker at the anterior end and a large sucker at the posterior end. A clitellum is present in the midregion during the reproductive period. The poorly developed eyes are paired structures at the anterior end. Setae are absent.

NATURAL HISTORY

Life cycle. *Reproduction.* Reproduction is similar in the free-moving and sedentary polychaetes. The usually separate sexes, indistinguishable in the immature state, lack permanent gonads (sex organs). Prior to sexual maturity, gametes (reproductive cells) begin to form from one of the three embryonic tissue types—namely mesoderm—from which are derived linings around the digestive tract, the coelom (the main body cavity), and the parapodia. The developing eggs or sperm are shed into the coelom, where they are nourished by special nurse cells (eleocytes) or other coelomic bodies. Nourishment of gametes is often achieved by the breakdown of muscle tissue, which is absorbed by the eleocytes and passed on to the developing eggs. Ripe eggs and motile sperm may leave the body through gonoducts, or tubes for the passage of reproductive cells; through excretory, or nephridial, pores; or through ruptures of the body wall.

The majority of polychaetes shed their gametes into the water. Various major body changes may precede the emission of gametes, the two most profound being epitoky and stolonization. In species of *Nereis* and *Syllis*, both free-moving polychaetes, epitoky involves the region in which the sex products are concentrated. Morphological changes include enlargement of the eyes, enlargement of a specific number of parapodia, replacement and alteration of setae, and development of an anal organ (rosette) for the emission of sperm. At sexual maturity these polychaetes leave their burrows and swim before releasing gametes.

Epitoky and stolonization in polychaetes

Surgical removal of the brain of nereids that normally undergo epitoky causes the development of morphological changes without the subsequent release of eggs or sperm. Nereids that normally do not undergo epitoky are

unaffected by the surgical removal of the brain. Apparently the substance that stimulates epitoky, a hormone not yet identified, is present only in species that normally undergo this phenomenon. In syllids, stolonization, the development of stemlike growths, may produce one or more stolons, or stems, containing developing gametes. Stolonization is also apparently controlled by a hormone, which is thought to be formed in the digestive tract.

Another type of epitoky occurs in some sedentary polychaetes; in these worms, the parapodial lobes develop long, thin setae. Either the entire animal or only the posterior portion (*e.g.*, the palolo worm) leaves the tube or burrow and swims before releasing gametes.

Hermaphroditism—that is, the production of eggs and sperm in one individual—rarely occurs in polychaetes. The free-moving polychaete *Ophryotrocha*, however, shows marked sexual variability; individual males or females may exist in association with hermaphroditic forms. Experiments with *Ophryotrocha* suggest that the age at which transition from one sex to another occurs may differ among groups within certain species.

Asexual reproduction is known in a few sedentary polychaete species. In some genera—*Ctenodrilus*, *Pygospio*, and *Sabella*—fragmentation of the body occurs, sometimes forming single segments, from which new individuals can develop.

Reproduction in oligochaetes is primarily hermaphroditic; the number, arrangement, and location of the male and female gonads and their pores vary considerably among the various species. Lower oligochaetes (Microdrili) have one pair of testes and one pair of ovaries in successive segments. The Haplotaxidae have one pair of testes in segment 10 and another in segment 11 and one pair of ovaries in segment 12 and another in segment 13. Higher oligochaetes (Megadrili) retain the two pairs of testes in segments 10 and 11 and the pair of ovaries in segment 13. Developing sperm are frequently stored in seminal vesicles before transfer to female receptacles. Sperm ducts lead from the seminal vesicles to male pores located one or more segments behind the testes. The ovaries are simple outpouchings (ovisacs) with oviducts leading to female pores in the next posterior segment.

Copulation in oligochaetes

Copulation in oligochaetes is reciprocal—that is, both sperm and eggs are exchanged—and takes place in a head to tail position, with the two ventral surfaces in contact (see Figure 2). In lower oligochaetes, the male

From W. Kukenthal and T. Krumbach (eds.), *Handbuch der Zoologie*; by permission of Walter de Gruyter, Berlin-New York

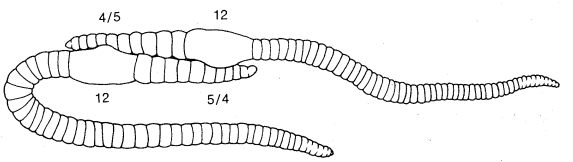


Figure 2: *Enchytraeus albidus* in copulation, the male pore, on segment 12, being directly opposed to the spermathecal pore of the other worm, between segments 4 and 5.

pores of one worm and the female pores of another are opposite each other, and sperm pass directly from the male pores into the seminal receptacle of the female. In some forms, sperm transfer occurs in the region of the clitellum, which is some distance from the male pores, and the sperm are carried along external grooves.

An earthworm lays eggs in a case, or cocoon, which is ovoid and of varying size, depending on the species; the cocoon slips forward and receives sperm as it passes the male pores of another worm. Fertilization, therefore, takes place within the cocoon. The clitellum not only secretes the cocoon that protects the eggs but also a material that serves as nourishment for the young and a mucous substance that aids in copulation.

Asexual reproduction is common in aquatic oligochaetes; indeed, sexual reproduction is virtually unknown in certain naidid species. Some oligochaetes divide to form a chain of two or more individuals that later break off as young worms. In a few species new individuals form by fragmentation of a worm, such as occurs in certain polychaetes. In many genera, individuals lay self-fertilized eggs capable of development. Others exhibit parthenogenesis—the production of young without fertilization—a phenomenon associated with polypoidy (multiple sets of chromosomes) in earthworms and accompanied by degeneration of male gonads.

All leeches are hermaphroditic, and reproduction is always sexual. The testes, from four to ten pairs, are arranged by segments, beginning with segment 12 or 13. The testes on each side of the body are connected with the vas deferens, a duct that leads indirectly to the male pore. The female reproductive system consists of one pair of ovisacs containing the ovaries, which, although located in front of the testes, may extend some length posteriorly, depending on the animal. The ovaries connect to form an oviduct that forms either a female pore or, in those species that copulate, a vagina.

In one leech family (Gnathobdellae), sperm are transferred by the penis of one animal into the vagina of another. In two other families (Rhynchobdellae and Erpobdellidae), sperm are transferred by sperm capsules, or spermatophores, onto the body of the leech, after which the sperm leave the spermatophore and enter the ovary through the female pore to unite with the eggs. Leech eggs, numbering from one to more than 100, are usually deposited in cocoons, which may be oval or elongated in shape and are generally attached to rocks or vegetation. Glossiphonids produce a membranous cocoon and attach it to their ventral surface, where development takes place. A clitellum, which forms only during the reproductive period, secretes the cocoon and material (albumin) to nourish the developing young.

Development. Annelid eggs, like those of flatworms and mollusks, exhibit so-called spiral, or determinate, cleavage. Development is termed determinate in annelids because early differentiation of various regions occurs; in indeterminate cleavage in echinoderm and chordate eggs, early differentiation does not occur.

In annelids, the first four cells (blastomeres) give rise, by alternating clockwise and counterclockwise divisions, to a cap of smaller cells, called micromeres, at one end of the egg and a cap of larger cells, called macromeres, at the other end. The first somatoblast—that is, a cell that gives rise to a specific type of differentiated tissue—which forms from a division of the macromeres at the 16-cell stage, gives rise to components, including the nervous system, derived from the embryonic tissue known as ectoderm. The second somatoblast forms at the 64-cell stage and gives rise to all future mesoderm, from which components such as muscle are formed. The cells that give rise to the endoderm, the forerunner of the digestive tract, are also produced at the 64-cell stage.

All cells divide simultaneously during the early stages of annelid development; during later stages, however, macromeres divide more slowly than micromeres. As a result, a ball of cells (solid gastrula) forms as the micromeres grow over the macromeres. The gastrula may form by either invagination (inpocketing of cells), epiboly (overgrowth and lengthening), or both processes. Some of the micromeres become arranged in a characteristic pattern, known as the annelid cross.

Trochophore larva of polychaetes

A free-swimming immature form called the trochophore larva develops only in the polychaete annelids but also occurs during the development of certain other invertebrate groups—mollusks, sipunculids, echuirids, and lophophores. The trochophore larva of polychaetes is typically diamond shaped with a circle of short, hairlike projections (cilia), called the prototroch, around the thickest part of the body. Cells bearing the prototroch develop from specific micromeres at the 16-cell stage. The four cells of the annelid cross frequently give rise to a so-called apical tuft of cilia at the anterior end. A tuft of cilia (the telotroch) may appear later at the posterior end, and a second ring, the mesotroch, may develop behind the prototroch.

The upper half of the trochophore—that portion above or in front of the prototroch—will become the postomium (head) containing the brain, the eyes, and the prostomial appendages, if present in the adult polychaete.

The lower half of the trochophore contains the digestive tract, excretory organs (protonephridia), and other internal organs; it is also the site of future segmentation. Whether or not the digestive tract functions during the trochophore period depends largely upon the size of the egg and the species. The mouth and anus form during the trochophore stage.

Typically, the first three segments form almost simultaneously in the lower half of the trochophore. Development of the parapodial lobe and the appearance of larval bristles (setae) follow shortly thereafter. The body grows in length by the addition of new segments from the preanal segment (pygidium or tail), which is the site of all additional segment formation. The setae generally fall off the first segment, which becomes the adult peristomium, or first postoral segment. The parapodial lobes of this segment either develop peristomial appendages or atrophy, depending upon the polychaete species. Adult setae gradually replace larval setae on the second and third segments.

Although the features of the trochophore larva are relatively uniform among species within an order, the nature of the larva also depends upon other factors—*e.g.*, egg size and larval ecology. Polychaete trochophores may be classified into those lacking a pelagic (ocean-dwelling) stage; those with a nonfeeding, short pelagic stage; those with a nonfeeding, long pelagic stage; and those with a feeding, long pelagic stage. Species lacking a pelagic trochophore stage show special adaptive features—*e.g.*, protection by the parents, formation of an egg capsule, the discharge of eggs within one of the parent's tubes, or viviparity (birth of young instead of eggs). The young of species with a short pelagic larval life—a few days or less—either are protected by a parent throughout much of larval life or are hatched from small eggs, with little or no yolk. Few species are characterized by a nonfeeding, long pelagic life; generally, such species have a large amount of yolk stored in the egg.

The most common polychaete trochophore feeds and has a long pelagic life; food consists of microscopic organisms such as diatoms or dinoflagellates. Structural modifications, usually large numbers of setae or bands of cilia around each newly formed segment, facilitate the long pelagic life. Some polychaetes, such as spionids, develop more than 20 segments before settling on a suitable surface. The long pelagic life may continue several months, during which the larva (*e.g.*, *Ophelia*) seeks out a particular sediment upon which to settle. After settling, the young polychaete quickly loses its trochophore characteristics and begins to resemble the adult.

Development of oligochaetes in cocoons

Development in oligochaetes takes place entirely within the cocoon; there is no free-living larval stage. The cocoon of the aquatic lower oligochaetes contains large eggs and relatively little albumin. The cocoon of the terrestrial higher oligochaetes contains small eggs but large amounts of albumin, which nourishes the developing embryos. The oligochaetes undergo a highly modified form of spiral cleavage. The ectoderm, endoderm, and mesoderm, however, arise in the same manner as in the polychaetes. The elongated gastrula has a ventral mouth at the front end and a posterior anus. At the gastrula stage, earthworms begin to feed on the albumin, the embryo elongates, and the mesoderm bands break into units to form the walls, or septa, of individual segments; the worm then leaves the cocoon and begins to construct a burrow nearby.

Early development in the leeches is similar to that of the oligochaetes. The mesodermal bands form the individual segments as in the oligochaetes, beginning anteriorly. As these mesodermal bands hollow out to form the body cavity, or coelom, mesenchymal cells from the lining of the coelom begin to form one large coelomic cavity extending the entire length of the worm; as the number of mesenchymal cells increases, however, the coelom becomes filled with them. This is the characteristic state of adults. Young leeches hatch from cocoons after feeding upon albumin.

Life-span. Since both the polychaetes and oligochaetes are able to regrow lost parts—*i.e.*, regenerate (see next section)—it has been commonly believed that they are essentially ageless. Few longevity studies have been carried out with polychaetes because they are difficult to maintain under laboratory conditions. Most of the adults of species studied have a characteristic number of segments, which form rapidly during early life and prior to the appearance of gametes. Many polychaetes, especially among the nereids, reproduce only once. Under laboratory conditions, survival after spawning is about one day. In nature these worms, usually quite sluggish after spawning, are eaten by fish and other animals. Nereids may live only a few months; at least one other species, however, requires years for maturation of gametes. Species that form stolons (stems), such as the syllids, or whose posterior end breaks off, such as the palolo worms, are capable of repeating the process; but the number of times and the length of time they are able to do so have not yet been established. Most sedentary polychaetes survive following spawning, but, again, it is not yet known how often this process can be repeated.

The life-span of oligochaetes is better established because they are so frequently used in laboratory experiments. Asexual reproduction for 130 generations has been reported in one aquatic species. Some earthworms are believed to live as long as ten years. Senescence, or aging, is known to occur in oligochaetes; *Eisenia*, for example, lives beyond a reproductive period with a progressive loss of weight. Aging oligochaetes darken in colour, largely as a result of an increase in pigment deposition. In addition, the metabolic rates decrease, and their physiological processes slow.

Very little is known about the life-span of leeches. One species of *Erpobdella* requires a year to reach sexual maturity, after which it lays cocoons one time and dies. Another species breeds once a year for two years and dies the third year.

Regeneration. It has been said that the Annelida contain the most highly organized animals with the power of complete regeneration. The powers of regeneration are greater in the polychaetes and lower oligochaetes than in the higher oligochaetes; leeches lack the ability to regenerate. Most polychaetes and oligochaetes can regenerate a new tail. The ability to replace an amputated part is usually restricted to the anterior end, where lost segments are replaced either by the same number or fewer; if fewer segments form, internal reorganization of the organ system follows. Regeneration from a single segment occurs naturally in the polychaetes *Ctenodrilus* and *Dodecaceria*.

Steps involved in regeneration

The process of regeneration occurs in a series of steps. First the wound seals over; then a structure (blastema) forms on the surface of the wound. New tissue probably arises from pre-existing parent tissue, although mesodermal regenerative cells known as neoblasts, which migrate to the site of the injury, are found in polychaetes and lower oligochaetes. As healing begins, ribonucleic acid (RNA) accumulates at the wound site, first in the epidermal cells and later in mesodermal cells. Actually, RNA begins to accumulate in cells several segments away from the injury but becomes restricted to the injury site when the neoblasts migrate there. The amount of glycogen, a complex carbohydrate that serves as an energy source in animals, in the oligochaete *Eisenia* decreases markedly near the point of injury, returning to normal only after regeneration is complete. Oxygen consumption in *Eisenia* drops immediately after an injury but increases twofold after the cells begin to divide.

There is evidence that specific chemical substances (hormones) control regeneration in both polychaetes and oligochaetes. A hormone from the posterior part of the brain is essential for posterior regeneration; its presence is apparent only after the second or third day following injury. A mature *Nereis* is unable to regenerate unless brains from young worms with tails removed are implanted in its coelom.

Many annelids do not feed during the early stages of regeneration. For earthworms, the primary energy source during this period is probably protein since the excretion of nitrogen-containing compounds (products of protein

breakdown) is higher than normal. Body protein also serves as an energy source during the regenerative process in *Dodecaceria*, which can completely regenerate itself from one segment before feeding.

Reversal of anterior–posterior polarity has been obtained in the Oriental earthworm (*Perionyx excavatus*). A piece removed from the anterior end regenerates a head at both cut ends if the cuts are made simultaneously. If the new anterior head then is removed, the posterior head becomes dominant and evokes tail regeneration at the surface from which the new anterior head was removed. Such phenomena reflect a physiological dominance of the anterior end.

Locomotion. The basic features of locomotion in annelids are most easily observed in the earthworm because it lacks appendages and parapodia. Movement involves extending the body, anchoring it to a surface with setae, and contracting body muscles. When the worm begins a forward movement, circular muscles at the anterior end contract, extending the head forward. At the same time the anterior end lifts from the surface to facilitate forward movement. A wavelike contraction originating in the circulatory muscles then passes toward the posterior end. When the wave of contraction nears the midregion of the body, longitudinal muscles contract, thereby shortening the region. A wave of contraction of longitudinal muscles follows, and the cycle is repeated. The setae of a segment are extended by certain body muscles to prevent backward movement of the segment during the contraction of the longitudinal muscles. The setae are retracted during the circular contraction period. Muscular movement is aided by the compartmentalization of the segment—coelomic fluid, confined by the segment walls, provides a substance against which the muscles can function. The earthworm is capable of reversing the direction of its movement; the waves of contraction pass forward.

Locomotion in free-moving polychaetes is accomplished by circular, longitudinal, and parapodial muscles and by coelomic fluid. When a worm such as *Nereis* moves slowly, the contractual force comes from the sweeping movement of the parapodia. The parapodia of each segment are not aligned, and the effective stroke is the backward one, in which the acicula (needlelike processes) are projected beyond the parapodium and come in contact with the crawling surface. In the recovery, or forward, stroke, the acicula retract, and the parapodium lifts free of the surface. When a parapodium ends its backward stroke, the next parapodium initiates one. Body undulations, which help the worm to move rapidly, are produced by the contraction of longitudinal muscles stimulated by the backward stroke of parapodium of a particular segment.

Locomotion in the burrowing sedentary polychaetes, especially those forms lacking anterior appendages, is similar to that of the earthworm. In tube-dwelling sedentary forms, such as the Serpulimorpha, locomotion is restricted to movement within the tube. In this group, the parapodia are reduced or absent; specialized setae, the uncini, function in much the same way as do parapodia in free-moving forms.

Locomotion in the leech may be compared, in part, to that of the inchworm (immature members of the moth family Geometridae); the anterior and posterior suckers serve as points of contact. When the posterior sucker attaches to a surface, the circular muscles contract, beginning at the posterior end. The leech thus elongates and the anterior sucker fastens to the surface. When the posterior sucker is released, a wave of contraction of the longitudinal muscles moves in a forward direction; this completes one cycle. During swimming, the dorsoventral muscles maintain a contracted state, and undulations of the body are produced by waves of contraction of the longitudinal muscles.

Ecology. *Habitats.* Marine habitats containing specific polychaetes do not exist as they do for mollusks and echinoderms. Many species, such as *Neanthes arenaceodentata* and *Capitella capitata*, cosmopolitan in distribution, are found throughout the world. Aquatic oligochaete species are widespread in suitable environments; terrestrial forms are less widely distributed, except for the earthworm and others that have been transported to new habitats, generally inadvertently, by man. The distribution of leeches is similar to that of oligochaetes, with the aquatic forms more widely distributed.

Some oligochaete species can secrete a tough mucous covering to protect themselves against either summer heat or winter cold. Some terrestrial burrowing forms burrow deeper into the ground during periods of adverse conditions. Some aquatic leeches burrow deep into the bottom of a pond or stream during the warm months. Polychaetes have no known mechanism for adapting to adverse conditions.

Commensalism and parasitism

Associations with other organisms. Commensalism, a beneficial relationship between two types of organisms, is common among certain scale worms (Aphroditamorpha, an order of polychaetes). These worms may be found in the tubes of sedentary polychaetes, in the mantle cavity of mollusks, such as chitons and limpets, on starfishes, in the rectum of sea cucumbers, and in the burrows of various other types. The scale worm *Arctonoe*, which normally lives on starfishes, is attracted to water flowing from the host starfish but not to that from other starfish species. It has been established that the attractant in the water is a chemical from the host, but its nature is as yet unknown. Tube-dwelling polychaetes, such as *Chaetopterus*, may be the host to scale worms, pea crabs, or fish, which eat material carried in by water currents produced by the host. Commensalism occurs in some aquatic oligochaete species. The posterior end of *Aspidodrilus*, for example, is modified as a large sucker for attachment to other worms. *Chaetogaster* attaches to the body of aquatic snails and feeds on a larval stage of trematode flatworms.

Parasitism is rare in polychaetes. Myzostomida, an atypical polychaete group, are commensal or parasitic either on the surface of or inside echinoderms, primarily sea lilies. Polychaete species that live on the surface feed on fine particles carried to the mouth of the sea lily. Parasites that live within sea lilies may be found in the body wall, the coelom, or the digestive tract. Parasitic infestations by polychaetes are frequently severe enough to cause wartlike growths on the surface of the host; such growths have been noted on the surfaces of fossil crinoids of the Paleozoic Era (more than 225,000,000 years ago), indicating these parasites established themselves early. Some forms, such as *Iphitime*, are parasitic in the branchial chamber of crabs; others, such as *Ichthyotomus*, are parasitic on marine eels. The young stages of the cosmopolitan polychaete species *Arabella iricolor* develop in the coelom of various species of another polychaete (*Diopatra*). Some acquatic oligochaetes live in the ureters of toads or in the eyes of frogs. All members of the family Branchiobdellidae (order Prosopora) are parasitic in the brood chambers of isopods or on the gills of crayfish, where they suck blood. Many leeches, all of which feed on blood, attach to the host only during feeding. Marine leeches, however, attach permanently to their fish host. Host specificity is rare in leeches, although some species feed only on one species. *Hirudo* feeds primarily on mammals but also sucks blood from snakes, tortoises, frogs, and fish; when young, it may eat oligochaetes.

Behaviour. Various polychaetes (*e.g., Syllis, Chaetopterus, Cirratulus, Terebella*) are bioluminescent—that is, capable of producing light. The phenomenon occurs within the cells of *Polynoe;* the lower surfaces of some scale worms (*Halosydna*) have special photocells that produce light when stimulated. *Odontosyllis* light production is related to sexual maturity and swarming, which is influenced by lunar cycles. The female produces a bright luminescence that attracts the luminescent male; light production decreases in the female following the production of gametes. In chaetopterids the process, which involves the discharge of a luminescent secretion from certain segments and from the antennae, is under nervous control; in *Chaëtopterus*, light can be produced in the parapodia by stimulating the ventral nerve. The significance of light production in this genus is unknown, however, because it lives in a tube through which light rays

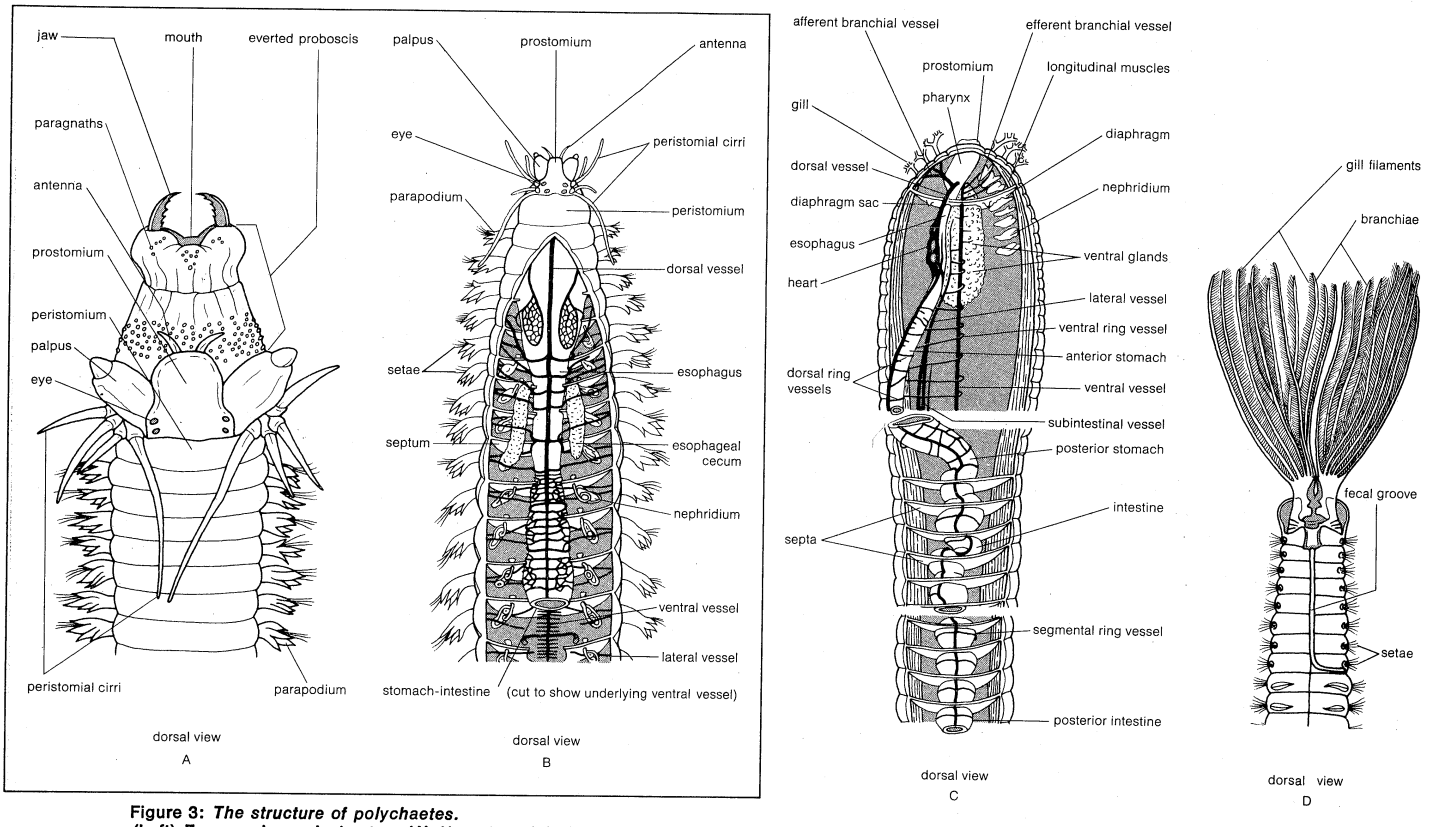


Figure 3: ***The structure of polychaetes.***
(Left) Free-moving polychaetes. (A) ***Neanthes*****, (B)** ***Nereis*****. (Right) Tube-dwelling (sedentary) polychaetes. (C)** ***Amphitrite*****, (D)** ***Sabella*****.**
(A,D) Adapted from original drawings by Janice L. Findley; (B,C) from F.A. Brown, *Selected Invertebrate Types* (1950); John Wiley and Sons, Inc.

cannot pass. Some earthworms (Oligochaeta), when stimulated, produce a luminescent slime from the mouth, anus, dorsal pores, or excretory pores; it is possible that the light is produced by bacteria living in the worm. Luminescence is unknown in leeches.

Polychaetes, especially the tube-dwelling Serpulimorpha, generally respond to changes in light intensity by withdrawing into their tubes.

Aggressive behaviour has been reported in several species of nereids (a group of free-moving polychaetes); they respond to a stimulus by extending the proboscis (feeding organ) to expose the jaws. *Neanthes arenaceodentata* will fight members of its own sex but not those of the opposite sex. The response may be related to spawning since this species does not swarm but lays gametes in the tube of another individual; fighting thus would prevent the occupation of one tube by two individuals of the same sex.

Both polychaetes and oligochaetes can learn to choose between favourable and unfavourable environments. Experiments have shown that individuals can learn to avoid a tube giving an electrical shock. In other experiments, earthworms will try about 12 times to bring into their burrow a leaf made immobile by attachment to some object; when an unattached leaf is presented to the worm, it turns to it and ignores the immobilized leaf thereafter.

FORM AND FUNCTION

Morphological characteristics. *External features.* The body of an annelid is often described as a tube within a tube. The inner tube, or digestive tract, is separated from the outer tube, or body wall, by the coelom. The head region (prostomium) is followed by a series of segments similar to each other in appearance. The body of many species, especially in the sedentary polychaetes, is separated into two or three regions. The cells comprising the epidermis (outermost cell layer) are usually simple columnar epithelial cells covered by a cuticle; parts of the body may be ciliated, especially in smaller forms. The cuticle consists of thin layers of protein similar in composition to that of the collagen found in some vertebrate tissues.

Polychaete structure

The body form of polychaetes (see Figure 3) varies, depending on whether the polychaete is free moving, sedentary, or pelagic (ocean dwelling).

The first segment, the prostomium, is in front of the mouth and may be a simple lobe or a highly developed projection. The next segment, the peristomium, surrounds the mouth and is followed by a series of segments, the total number of which may be limited or unlimited. The parapodia, fleshy outgrowths on each segment following the peristomium, contain bundles of setae (movable bristles), which differ in structure and function among species and thus provide a key to species identification. A seta consists of a basal portion within a follicle and a shaft projecting from the follicle; it is secreted from an epidermal cell, which encloses both the ciliary apparatus from which the seta arises and the lacuna in which the seta develops and through which it pushes to the outside. Composite, or pointed, setae are formed from two or more epidermal cells. In many species the shaft has a blade that can be moved by muscles found at the basal portion. New setae form in reserve follicles and move forward to replace old ones, which are discarded.

Branchiae, or gills, lacking in polychaete species that breathe through the body wall, are simple filaments or tufts near the anterior end of the worm. A mass of feeding structures in sabellid and serpulid polychaete worms, called a tentacular crown, functions both for food gathering and for respiration. Polychaete sensory receptors include eyes, lateral organs, dorsal ciliated ridges, statocysts (organs of balance), taste buds, papillae (blunt-shaped projections), and stiff hairs. The eyes, which range in complexity from simple pigment spots to eyes with lenses, may be found on the prostomium, on the peristomium, on the pygidium, along the sides of the body, or on the tentacular crown. Touch receptors consisting of either stiff hairs or modified cirri may be found over the entire body surface.

Oligochaete structure

The oligochaete body (see Figure 1) is usually cylindrical, is sometimes flattened, and rarely has projecting structures. Segmental lines are usually conspicuous, and secondary segmentation may occur in larger forms. The number of segments varies from seven in some aquatic species to 600 in the earthworms. Setae, embedded in the body wall, may be simple, S-shaped, forked, or hairlike. Except the first, each segment may have either two pairs of S-shaped setae or a circle of setae. Many transitional forms of setal arrangement occur, and copulatory setae are found on some segments in certain species. All sexually mature oligochaetes have a clitellum (a glandular structure derived from the epithelium), which secretes the egg capsule; it may be saddle-shaped or ring shaped. In lower oligochaetes it consists of a single layer of modified epithelial cells; in higher forms, such as earthworms, it may have many layers.

The ends of nerves, which probably respond to touch, heat, and pain, branch among the epidermal cells of oligochaetes. Epithelial sense organs resembling taste buds occur in the skin and mouth cavity; they probably function as chemoreceptors (*i.e.*, smell and taste receptors). Photoreceptors, or light-sensitive organs, are abundant at the anterior and posterior ends of earthworms. Earthworms respond negatively to strong light but are attracted to weak light. All oligochaetes are strongly thigmotactic (attracted to surfaces). Some forms have pressure receptors, sensory hairs, and pits.

A leech (see Figure 4), which has 34 segments, may increase in length as a result of subdivision and elongation of the annuli, or rings, which divide each segment. The typical number of annuli per segment in the midregion is three to five. The anteriorly located eyes usually vary in number from one to five pairs. The clitellum, which is present during reproduction, extends from segments 10 through 12. The most conspicuous of the external features of the leech are the small anterior and the large posterior suckers.

From D.J. Reish, *Laboratory Manual for Invertebrate Zoology*

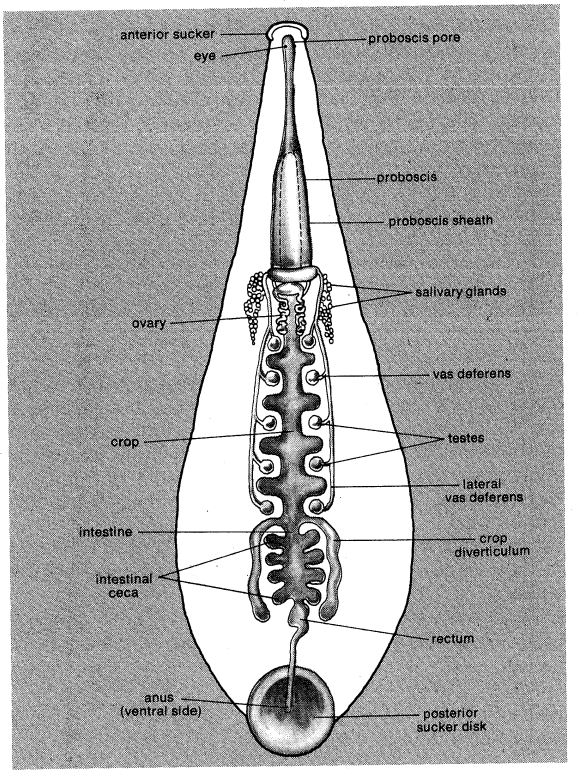


Figure 4: The internal structure of leeches.

Internal features. The body cavity of annelids is lined by epithelium. Successive body segments are separated by walls that correspond to the external rings. In grooves between the segments of some oligochaetes are dorsal pores through which coelomic fluid may be discharged. The coelom is filled with connective tissue. Internal features of the three annelid classes are shown in Figures 1 (oligochaetes), 3 (polychaetes), and 4 (leeches).

Digestive systems

The polychaete digestive system is generally a straight tube; a mouth leads into an esophagus, which is followed by the intestine and the anus. Some free-moving forms have a proboscis that can be thrust forward by being turned inside out—that is, the proboscis is eversible. In oligochaetes such as the earthworm, the mouth opens into a muscular pharynx, which opens to the esophagus and then to a muscular gizzard. The intestine, which extends most of the length of the worm, terminates in an anus. In leeches, the mouth, surrounded by the anterior sucker, opens into the esophagus; the crop and intestine follow—each with minute pockets (diverticula)—then the rectum and anus.

The circulatory system in the lower oligochaetes consists of a dorsal vessel that arises from a blood sinus or capillary network surrounding the intestine and conveys blood forward; a ventral vessel that conveys blood backward; and connective vessels between the two. The blood vessel walls consist of an outer membranous (peritoneal) layer containing muscle fibres, a middle region of collagenous material, and an inner lining of thin cells (endothelium). In higher oligochaetes, one or more pairs of hearts connect the dorsal and ventral vessels and propel the blood. In free-moving polychaetes the dorsal vessel is the chief propulsive force, and networks of small vessels connect the dorsal and ventral ones. In some leeches the blood is propelled by a dorsal vessel connected by loops at both ends to a ventral one.

Blood is moved by wavelike contractions of the blood vessels, by the beating of cilia, or by pumping provided by hearts. In *Arenicola* and the earthworm the heartbeat apparently is initiated in nerve cells rather than in muscle tissue, as occurs in vertebrates. The blood apparently carries nitrogen-containing products to the nephridia for excretion. The only blood cells are amoebocytes, which are free-moving cells that engulf particles.

The basic units of the annelid excretory system are either protonephridia (solenocytes), which have tubules that end blindly within cells, contain flagella (whiplike projections), and are joined to a common duct that drains to the outside; or metanephridia, which are funnel-shaped structures containing cilia (short hairlike processes) that open to the outside.

The nervous system of free-moving polychaetes, similar to that of oligochaetes, consists of a dorsal brain, or supra-esophageal ganglion, a discrete mass of nervous tissue, in the prostomium; a pair of nerves united ventrally to form the ventral subesophageal ganglion; and paired nerve cords with one ganglion per segment. In sedentary polychaetes, the brain may become highly modified.

Physiological characteristics. The coelomic fluid of annelids plays a role in many important functions—*e.g.*, locomotion, regulation of fluid transfer through the body wall (osmoregulation); many metabolic processes occur in the coelom, which also serves as a site for temporary food storage, for excretion of nitrogen-containing wastes, and for maturation of gametes. The coelomic walls of earthworms contain cells, called chloragocytes, that apparently store and metabolize oil and glycogen and also produce ammonia and urea. Granules within the chloragocytes, which eventually disintegrate in the coelomic fluid, are taken up by amoebocytes, which increase in size, becoming large brown bodies that are never eliminated from the body.

Digestion. Most annelids, except leeches, either lack or have poorly developed diverticula, minute pockets that serve as digestive glands. Instead, the gut lining contains secretory cells (concentrated in the foregut) and absorptive cells (concentrated in the hindgut). Digestive enzymes are most active in the gut. Digestion within cells has not been demonstrated in annelids. A lengthwise fold, the typhlosole, hangs downward in the intestinal cavity of oligochaetes. The absorptive surfaces of the typhosole and of the anterior intestine may have a brush border; fats are absorbed only in this region.

Calciferous glands, found only in certain earthworms, apparently excrete calcium by secreting granules of calcium carbonate that are transformed into calcite crystals in the intestine.

Respiration. Gas exchange generally takes place through the skin, but it may occur through gill filaments in some polychaetes or through the rectum of aquatic oligochaetes. Although oxygen may be transported directly in the blood, it is usually carried by respiratory pigment, either hemoglobin or chlorocruorin. Hemoglobin, the most common pigment, is present in most free-moving and some sedentary polychaetes and in most oligochaetes and leeches. Chlorocruorin is found in several polychaete groups (Flabelligerida, Terebellidomorpha, and Serpulimorpha). A few free-moving polychaetes, some oligochaetes, and rhynchobdellid leeches have colourless blood. The blood of the polychaete *Serpula vermicularis* contains both pigments, the young having more hemoglobin and the old more chlorocruorin.

Respiratory pigments

Annelid hemoglobin molecules have several properties in common with the hemoglobin found in vertebrates but differ in molecular weight (based on the weight of a hydrogen atom as 1) and in the relative amounts of certain constitutents. Chlorocruorin differs from hemoglobin in having a lower affinity for oxygen and in being green in dilute solutions, red in concentrated ones.

The properties of annelid respiratory pigments are associated with the mode of life of the worm. The hemoglobin of the lugworm *Arenicola*, a polychaete, releases oxygen to the tissues only under conditions of extreme oxygen deficiency. The hemoglobin of some earthworms takes up oxygen from a normal atmosphere but releases it only when tissue oxygen is low and, thus, may protect the worm from oxygen poisoning.

Excretion. Ammonia is the chief nitrogen-containing end product of protein metabolism in aquatic annelids; earthworms, which are adapted to living in the soil, excrete larger quantities of another nitrogen-containing compound, urea, probably as part of a mechanism for the control of salt and water balance in the worm. The sea mouse *Aphrodita*, a polychaete, excretes 80 percent of its nitrogen as ammonia, which is also the primary nitrogenous excretory product in leeches (smaller amounts of urea also are excreted). Part of the ammonia excreted by leeches may come from bacteria living within part of the leech's excretory system (nephridial capsules). The ability of leeches to withstand high concentrations of ammonia is believed to result from a protective effect provided by high levels of calcium in their cells.

Three aspects of nephridial function in annelids correspond to those of the vertebrate kidney—filtration, resorption, and secretion. Coelomic fluid filters through solenocytes (see above *Internal features*). The ciliated funnels of metanephridia retain minute particles and those of moderate size. In oligochaetes, whose coelomic fluid contains proteins, particles are actively absorbed in the ciliated region of the tubule. The tubules of earthworms also resorb inorganic ions such as sodium and calcium and can selectively eliminate excretory products from both the coelomic fluid and the bloodstream.

Osmoregulation. The fluids of marine polychaetes have the same salt balance as (*i.e.*, are isotonic with) seawater and thus can tolerate no more than a moderate change in the salt (*i.e.*, ion) content of the water. Coelomic fluids contain little or no protein. Certain aquatic oligochaetes, which can live exclusively in freshwater, are capable of regulating the internal medium because, although their coelomic fluid contains fewer salts than does that of polychaetes, it contains more proteins. Osmoregulation has been studied in freshwater leeches, which have mechanisms similar to oligochaetes.

The body wall of a typical marine polychaete, such as *Perinereis cultrifera*, which cannot adapt to salinity fluctuations of seawater, swells and bursts if salinity is reduced to 20 percent that of seawater because the worm has no physiological mechanism for the control of water intake. On the other hand, certain individual *Nereis diversicolor* worms are capable of tolerating intertidal changes of salinity because they have enlarged nephridia that enable them to excrete excess water. Some groups of *N. diversicolor* can adapt to salinity differences as extreme as that between the Black Sea (about 60 percent that of seawater) and a nearby lake (about the same as seawater).

Neurosecretion. The brain contains several types of cells whose secretory activities relate to phases of the life cycle, especially those of reproduction, growth, and regeneration.

Neurosecretory cells, which are nerve cells that produce hormones, are found in the brain; their structure, similar to that of nonsecretory nerve cells, consists of fine projections (an axon and neurofibrillae) and a cell body. The secretions of neurosecretory cells, which terminate in the walls of a blood vessel, in other fluid systems, or in the epidermis, are in the form of microscopic droplets or granules. Neurosecretory cells seem to be derived from epidermal secretory cells that have been incorporated into the central nervous system.

Inhibitor hormones are known in some Phyllodocemorpha, and a stimulator substance has been identified in Drilomorpha, both of which are polychaete groups. Inhibitor hormones in nereids and syllids have been described above (see *Reproduction*). The maturation of gametes is apparently inhibited in nephtyid polychaetes by neurosecretions of the brain. The brain of the lugworm *Arenicola* stimulates maturation of gametes.

The brain has been shown to play a role in the regeneration of the posterior end of the body of polychaetes such as nereids and nephtyids, but the effect may be an indirect one involving the genital inhibiting hormone. Neurosecretory cells occur in the brain and subesophageal ganglia of several terrestrial and aquatic oligochaete species. Removal of the brain from sexually maturing earthworms causes degeneration of the clitellum and prevents gamete formation. The brain also plays a role in osmoregulation, as indicated by the increase in chloride concentration in the urine of oligochaetes lacking a brain. The neurosecretory cells in the brain of leeches control gamete formation.

Role of the brain in regeneration

Coordination. The muscles of annelids are coordinated by both the ventral nerve cord, which is composed of two strands and extends the length of the worm, and the location, within each segment, of a ganglion and nerves. The nerves within each segment carry impulses away from the ganglion (motor nerves) or toward it from a sensory receptor (sensory nerves). The cell bodies of sensory nerves are located beneath the surface epithelium; those of motor nerves are either within the ganglion or in separate parapodial ganglia. Each segmental nerve innervates those components of the body wall, parapodia, and the digestive tract found in its segment.

The nerve cord of many annelids has giant nerve fibres (neurocords), which may have either a simple or a compound structure. Simple neurocords are very large single nerve cells; their axons arise from cells situated in either the brain or a segmental ganglion. Compound neurocords are multiple structures; each axon is connected to numerous cell bodies along its course. The function of the giant nerve cord is the rapid transmission of impulses from one end of the worm to the other; this enables the longitudinal muscles of each segment to contract at about the same time. The value of rapid contraction is evident in the escape reaction of tube-dwelling sedentary polychaetes.

Role of giant nerve fibres

Some giant nerve fibres convey impulses as fast as vertebrate nerve fibres (about 70 feet per second); annelid fibres, however, are larger in diameter (1.5 m in *Myxicola*) and lack a thick insulating sheath (myelin). Not only is recovery from the passage of impulses slower in giant nerve fibres than in other annelid nerves but the former are also the last component to develop in the nerve cord of a growing worm. The nerve cord of *Myxicola* contains one giant nerve fibre, which is used to study the properties of the nerve impulse. In *Myxicola*, an impulse may be conducted in either direction along the nerve, unlike *Nereis* or the earthworm; may be initiated at any level; and is an all or nothing action.

The contraction of annelid muscles occurs in a manner

similar to that of other animals—that is, the proteins actin and myosin are involved.

EVOLUTION AND PALEONTOLOGY

Evolutionary hypotheses. The annelids are considered to have evolved in the sea, perhaps from an ancestral flatworm through the trochophore larva, the characteristic early stage of polychaetes. The oligochaetes are thought to have developed from polychaete stock; the leeches, with the clitellum in common with the oligochaetes, probably evolved from the latter. (See Figure 5 for a diagrammatic representation of annelid relationships.)

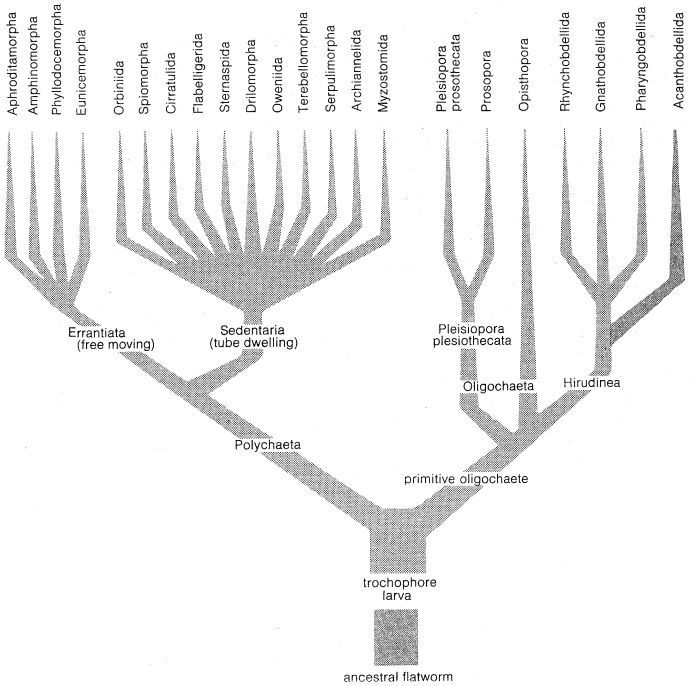


Figure 5: Possible origin of relationships among annelids.

The question of which polychaete order preceded the others remains unresolved. The Archiannelida were long considered to have been the earliest polychaete group because of their primitive condition; some members (*e.g.*, *Polygordius*), which lack setae and external segmentation and have simple nervous, muscular, and circulatory systems, are now considered, however, to be a specialized group. *Polygordius* species typically are small in size; they have cilia on their surface for locomotion, respire through the skin, and have internal fertilization. Finally, the larvae undergo nonpelagic development. The polychaetes appear therefore to have undergone radiative evolution, in which every character has been modified independently of the others. There is thus little basis for regarding any one order as ancestral to the others.

The evolution of oligochaetes from polychaetes may be related to the change from a marine to a freshwater habitat. One view is that oligochaetes evolved in marine swamps subjected to periodic drying; survival during dry periods would have been made possible by egg cocoons. Another hypothesis, contrary to the first, is that the primitive oligochaete was adapted to permanent freshwater conditions rather than to a terrestrial habitat. Some authorities consider the oligochaetes to have evolved from some members of the Eunicemorpha (Lumbrineridae) or Drilomorpha (Capitellidae), but this may result from a superficial resemblance in body form and thus may be of little evolutionary significance.

Reproductive structures provide not only the main criteria for understanding the course of evolution within the oligochaetes but also provide the basis for the classification of oligochaetes. Two principal paths of evolution apparently occurred in the oligochaetes—the Plesiopora plesiothecata branch and the Opisthopora branch. The Plesiopora plesiothecata branch, in which the vas deferens (*i.e.*, the tubule carrying sperm from the testes) opens outward on the segment immediately behind the segment containing the testes, evolved into two lines differentiated on the basis of whether the seminal receptacle, a storage cavity, opens in front of the testes (Plesiopora prosothecata line) or at the same segment or posterior to the testes (Prosopora line). In the second principal branch, the Opisthopora, the vas deferens opens a few segments behind the testes. Although the Haplotaxidae are placed in the Opisthopora branch, they are considered to be intermediate between the Opisthopora and Plesiopora plesiothecata and to be the transitional group between aquatic and terrestrial oligochaetes.

Significance of reproductive structures in oligochaete evolution

There is little doubt that the leeches evolved from the oligochaetes, since both groups have a clitellum, at least during the reproductive period, and are hermaphroditic. The Acanthobdellae are considered to be the link between the oligochaetes and leeches because they possess setae and walls between segments; the order contains only one known species, however. The three remaining orders of leeches evolved into two lines based on whether or not the animals have jaws.

Fossils. The fossil record of annelids is limited because they have almost no hard body parts. Tubes constructed by polychaetes and polychaete jaws are the most commonly encountered fossil specimens. Most fossil records of oligochaetes are doubtful, and fossil leeches are unknown. Some burrows, or tubes, have been interpreted as belonging to wormlike creatures from Precambrian strata (more than 620,000,000 years old). Fossils resembling the scale worm *Halosydna*, the sea mouse *Aphrodita*, *Nereis*-like forms, and calcareous tubes similar to present-day *Serpula* and *Spirorbis* species have been described. The shells of Paleozoic mollusks (more than 230,000,000 years old) are occasionally marked by U-shaped tubes similar to those made by the polychaete *Polydora*, a modern-day pest of oysters. The tough jaws of polychaetes, containing minute spiny black teeth known as scolecodonts, occur from the Cambrian Period (about 570,000,000 to 500,000,000 years ago) onward.

CLASSIFICATION

Distinguishing taxonomic features. Because most taxonomists involved with annelids specialize in one of the three classes, the taxonomy of the classes is somewhat unrelated. Classification of free-living and sedentary polychaetes relies almost exclusively on external characters, such as the shape of the head, number and nature of structures, such as appendages (including anal ones), parapodia, setae, and tube construction. Oligochaete classification into orders and families relies largely on internal structures, especially the arrangement and number of gonads, the position of the gonoducts, and particularly the location of the male pore. Setal characteristics are employed to a limited extent because they are generally uniform among species. Leech classification is based on the presence or absence of setae and the nature of the mouth, proboscis (feeding organ), jaws, suckers, eyes, and reproductive system.

Annotated classification. The following classification incorporates the views of several authorities.

PHYLUM ANNELIDA (segmented worms)
Body wall covered with a cuticle secreted by the epidermis and containing an outer circular and inner longitudinal muscle layer; chitinous (tough, complex carbohydrate material) setae usually present, secreted by follicular cells and arranged segmentally; head or prostomium preoral, with or without appendages; closed circulatory system, with blood often containing a respiratory pigment; coelom, of schizocoelomic origin, divided segmentally into compartments by walls, or septa; nervous system includes a dorsal, bilobed brain and a pair of connective nerves that encircle the digestive tract and unite to form a ventral nerve cord with 1 ganglion per segment.

Class Polychaeta (marine worms)
Paired lateral appendages, or parapodia, bearing chitinous setae; name of group refers to the many setae per segment; head with or without appendages; sexes generally separate with gametes discharged directly into the water, where fertili-

zation and development occur; the free-swimming larva called a trochophore; about 5,400 living species; free-moving and sedentary (tube-dwelling) forms.

Order Aphroditamorpha (scale worms)

Free moving; dorsally rounded, with flattened pairs of scales more or less alternating with the dorsal cirri (slender projections); head with 1 or 3 tentacles, 2 palpi (fleshy sensory projections), and 4 tentacular cirri used for feeding and respiration; projecting (protrusible) proboscis cylindrical in shape, with border of soft papillae (nipple-like projections) and 4 chitinous jaws; size, 0.5 to 25 cm; examples of genera: *Aphrodita* (sea mouse), *Halosydna* (common scale worm), *Arctonoe.*

Order Amphinomorpha

Free moving; prostomium with 1 to 5 antennae, 2 palps, and a caruncle (posterior ridge) deeply set into anterior segments; parapodia with 2 lobes and branchiae (gills); size, 0.5 to 35 cm; examples of genera: *Eurythoe* (fireworm), *Euphrosyne.*

Order Phyllodocemorpha

Free moving; a large group characterized by a protrusible proboscis that may or may not be armed with chitinous jaws, teeth, or papillae; prostomium with 1 to 5 antennae, with palpi, and with none to 3 pairs of eyes; parapodia well developed into 1 or 2 lobes usually bearing compound setae; size, 0.2 to over 1 m; examples of genera: *Anaitides, Syllis, Hesione, Nereis, Glycera* (bloodworm), *Nephtys.*

Order Eunicemorpha

Free moving; head with or without appendages and eyes; proboscis with dorsal maxillae (upper jaws) of 1 to many paired pieces, a ventral pair of mandibles (lower jaws) more or less fused along the median line, and a pair of embedded maxillary carriers; parapodia single-lobed, often with many acicula (needlelike structures); size, minute to 3 m; examples of genera: *Palolo* (palolo worm), *Eunice, Stauronereis, Lumbineris, Onuphis.*

Order Orbiniida

Sedentary; head pointed or rounded without appendages; proboscis eversible and unarmed; body divided into distinct thorax and abdomen; gills arise dorsally from thoracic region; size, minute to 40 cm; examples of genera: *Scoloplos, Paraonis.*

Order Spiomorpha

Sedentary; at least 2 long feeding tentacles adapted for grasping and arising from prostomium; body divided into 2 or 3 body regions; size, 0.5 to 25 cm; examples of genera: *Spio, Polydora, Chaetopterus* (parchment tube worm), *Magelona.*

Order Cirratulida

Sedentary; prostomium pointed and without appendages; 1 or more pairs of tentacular cirri arising from dorsal surface of anterior segments; gills, if present, long and slender, inserted above parapodia; size, minute to 20 cm; examples of genera: *Cirratulus, Cirriformia.*

Order Flabelligerida

Sedentary; setae of anterior segments directed forward to form a cephalic (head) cage; prostomium and peristomium retractile, with 2 palpi and retractile branchiae; size, 1 to 10 cm; examples of genera: *Flabelligera, Stylariodes.*

Order Sternaspida

Sedentary; body short, swollen at both ends; anterior setae short and thick; posterior end with ventral shield bearing radiating setae and anal branchiae; size, 3 cm; genera include *Sternaspis.*

Order Drilomorpha

Sedentary; head without appendages, but prostomium may form lobes or membranes anteriorly or laterally; body divided into 2 or 3 regions; paired branchiae may arise laterally from body segments; size, 0.5 to 25 cm; examples of genera: *Capitella, Arenicola* (lugworm), *Axiothella* (bamboo worm), *Ophelia.*

Order Oweniida

Sedentary; anterior end with or without divided lobed membrane; anterior segments long; dwelling tube mucoid, coated with sand or shell fragments; size, 0.2 to 10 cm; genera include *Owenia.*

Order Terebellomorpha

Sedentary; head concealed by filamentous tentacles; branchiae, simple or branched, arising from dorsal surface of anterior end; body divided into thorax and abdomen; tube of mucoid substance to which sediment adheres; size, 1 to 40 cm; examples of genera: *Amphicteis, Terebella, Pista, Thelepus.*

Order Serpulimorpha (the feather dusters)

Sedentary; head concealed with featherlike filamentous branchiae; body divided into thorax and abdomen; tube mucoid or calareous; size, minute to 50 cm; examples of genera: *Sabella, Eudistylia, Serpula, Hydroides.*

Order Archiannelida

Minute, primitive, with ciliated epidermis; prostomium small, with or without appendages; parapodia absent; septa reduced or absent; size, minute; genera include *Polygordius.*

Order Myzostomida

Body disk-shaped or oval without external segmentation; external or internal commensals or parasites of echinoderms, especially crinoids; size, minute to 1 cm; genera include *Myzostoma.*

Class Oligochaeta

Primarily freshwater or terrestrial with setae arising directly from body wall; name of group refers to the few setae per segment; head and body appendages generally lacking; hermaphroditic, with testes located anteriorly to ovaries; gonoduct system complex; seminal receptacle used to store sperm; reproduction by copulation, with fertilized eggs laid in a cocoon secreted by clitellum; development direct, without larval stages; about 3,200 living species.

Order Pleisiopora plesiothecata

Chiefly aquatic worms; male gonopores in segment immediately behind testes; seminal receptacle at or near segment containing testes; size, minute to 1–3 cm; examples of genera: *Nais, Tubifex* (sludge worm).

Order Pleisiopora prosothecata

Aquatic and terrestial; male gonopores in segment immediately behind the segment containing testes; seminal receptacle anterior to testes; size, minute to 5 cm; genera include *Enchytraeus.*

Order Prosopora

Aquatic and terrestrial; male gonopores on the segment containing the testes or, when 2 pairs of testes are present, in the more posterior segment; size, minute to 14 cm; examples of genera: *Lumbriculus, Branchiobdella.*

Order Opisthopora (earthworms and allies)

Male gonopores several segments behind segments containing testes; size, minute to over 3 m; examples of genera: *Haplotaxis, Eisenia, Lumbricus* (earthworm), *Megascolides.*

Class Hirudinea (leeches)

Primarily freshwater, but also terrestrial and marine forms; small sucker at anterior end, large sucker at posterior end; fixed number of body segments at 34; body cavity filled with connective tissue; hermaphroditic, with fertilized eggs laid in a cocoon secreted by clitellum; development direct without larval stages; about 300 living species.

Order Acanthobdellida

Primitive group; setae present on 5 anterior segments; no anterior sucker; parasitic on fish in Lake Baikal (U.S.S.R.); size, small; genera include *Acanthobdella.*

Order Rhynchobdellida

An eversible pharynx used to penetrate host tissue; jawless; distinct blood vessels contain colourless blood; freshwater or marine inhabitants; size, minute to 20 cm; examples of genera: *Glossisphonia, Piscicola, Pontobdella.*

Order Gnathobdellida

Pharynx with 3 toothed jaws, noneversible; red blood circulates through distinct vessels; terrestrial, in moist tropical regions; also in freshwater; size, minute to 20 cm; examples of genera: *Hirudo* (medicinal leech of man), *Haemopis.*

Order Pharyngobdellida

Pharynx without jaws and noneversible; carnivorous leeches that cannot suck or penetrate hosts; freshwater or terrestrial; size, minute to 5 cm; examples of genera: *Erpobdella, Odontobdella.*

Critical appraisal. There is not complete agreement on annelid classification, especially above the family level and among the polychaetes. The lack of a uniform system results not only from gaps in knowledge but also from both the tremendous diversity in structure and specialization in habitat, especially in the polychaetes.

Alternate classifications place the Oligochaeta and Hirudinea in a single class, called the Clitellata, because both groups possess a clitellum; the oligochaetes and leeches would thus be subclasses within the class Clitellata. In other systems, the Polychaeta and Oligochaeta are listed as subclasses of the class Chaetopoda because both groups have setae. Both the Archiannelida and Myzostomida, treated as polychaete orders in the classi-

fication above, are sometimes given class status, so that, under this system, there are five classes.

The class Polychaeta has also been divided into two subclasses or orders, the Errantiata (free-moving forms) and Sedentaria (sedentary, or tube-dwelling, forms), based on the mode of life. The subclasses, even if they are not given a definite position in polychaete classification, are convenient designations. If the two categories are considered to be subclasses, then the 14 orders are as indicated in the system outlined above; if Errantiata and Sedentaria are used as orders, however, only 70 polychaete families remain.

The oligochaetes may also be divided into two orders, the Microdrili and Megadrili, on the basis of aquatic or terrestrial habitat, respectively. If this system is employed, only the terrestrial members of order Opisthopora are included with the Megadrili, and the remainder are considered as Microdrili.

The placement of the families of leeches into the four orders outlined above is generally agreed upon by workers of the field.

BIBLIOGRAPHY. R.P. DALES, *Annelids* (1963), a semipopular work; M. FLORKIN and B.T. SCHEER (eds.), *Chemical Zoology*, vol. 4, *Annelida, Echiura, and Sipuncula* (1969), on physiology; "Annélides . . .," in A.P. GRASSE (ed.), *Traité de zoologie*, vol. 5 (1959), a comprehensive review in French; O. HARTMAN, *Literature of the Polychaetous Annelids*, vol. 1, *Bibliography* (1951), list of all polychaete literature to date, and vol. 2, *Catalogue of the Polychaetous Annelids of the World* (1959), list of all scientific names of polychaetes; M.S. LAVERACK, *The Physiology of Earthworms* (1963), a physiological treatment; J. STEPHENSON, *The Oligochaeta* (1930), a comprehensive treatment of oligochaetes; K.H. MANN, *Leeches (Hirudinea): Their Structure, Physiology, Ecology and Embryology* (1962) a semipopular work.

(D.J.R.)

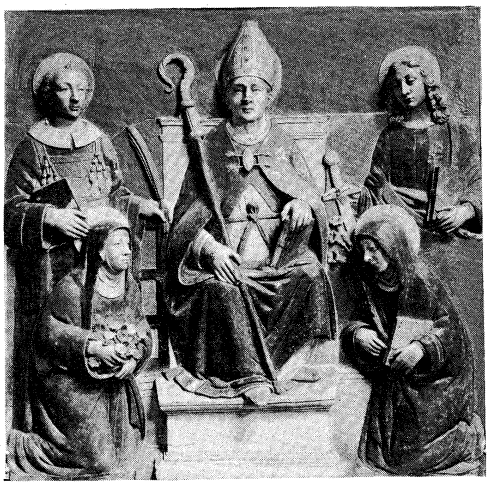
Anselm of Canterbury, Saint

The founder of Scholasticism, the philosophical school of thought that was to dominate the Middle Ages, Anselm of Canterbury was one of the most important thinkers between Augustine (354–430) and Thomas Aquinas (1225/26–1274). He is recognized today as the originator of the so-called ontological argument for the existence of God (based on the idea of an absolutely perfect being, the fact of the idea being in itself a demonstration of existence) and the satisfaction theory of the atonement, or redemption (based on the feudal theory of making satisfaction or recompense according to the status of a person against whom an offense has been committed, the infinite God being the offended party and finite man the offender). His works have been influential in the revival of the ontological arguments in the philosophy of René Descartes in the 17th century and the theology of Karl Barth in the 20th century. The accomplishments of Anselm also illustrate the international character of the medieval church and society (the *corpus Christianum*): the son of a Lombard (Italian) nobleman became an abbot in France and an archbishop in England.

He was born in 1033/34 in the Piedmont region of northwestern Italy. His birthplace, Aosta, was a town of strategic importance in Roman imperial and in medieval times, because it stood at the juncture of the Great and Little St. Bernard routes. His mother, Ermenberga, belonged to a noble Burgundian family and possessed considerable property. His father, Gondolfo, was a Lombard nobleman who intended that Anselm would make a career of politics and did not approve of Anselm's early decision to enter the monastic life. Anselm received an excellent classical education and was considered one of the better Latinists of his day. His early education impressed on him the need to be precise in his use of words, and his writings became known for their clarity.

Early monastic career

In 1057 Anselm left Aosta to enter the Benedictine monastery at Bec (located between Rouen and Lisieux in Normandy, France), because he wanted to study under the monastery's renowned prior, Lanfranc. While on his way to Bec, he learned that Lanfranc was in Rome, so he spent some time at Lyons, Cluny, and Avranches before



St. Anselm (centre), terra-cotta altarpiece by Luca della Robbia (1399/1400–1482). In the Museo Diocesano, Empoli, Italy.
Alinari—Art Reference Bureau

entering the monastery in 1060. In 1060 or 1061 he took his monastic vows. Because of Anselm's reputation for great intellectual ability and sincere piety, he was elected prior of the monastery after Lanfranc became abbot of Caen in 1063. In 1078 he became abbot of Bec.

In the previous year (1077), Anselm had written the *Monologium* ("Monologue") at the request of some of his fellow monks. A theological treatise, the *Monologium* was both apologetic and religious in intent. It attempted to demonstrate the existence and attributes of God by an appeal to reason alone rather than by the customary appeal to authorities favoured by earlier medieval thinkers. Moving from an analysis of the inequalities of various aspects of perfection, such as justice, wisdom, and power, Anselm argued for an absolute norm that is everywhere at all times, above both time and space, a norm that can be comprehended by the mind of man. Anselm asserted that that norm is God, the absolute, ultimate, and integrating standard of perfection.

Under Anselm, Bec became a centre of monastic learning and some theological questioning. Lanfranc had been a renowned theologian, but Anselm surpassed him. He continued his efforts to satisfactorily answer questions concerning the nature and existence of God. His *Proslogium* ("Addition," or "Sequence"), or "Address Concerning the Existence of God," established the ontological argument for the existence of God. In it he argued that even a fool has an idea of a being greater than which no other being can be conceived to exist; that such a being must really exist, for the very idea of such a being implies its existence.

The ontological argument

Anselm's ontological argument was challenged by a contemporary monk, Gaunilo of Marmoutier, in the *Liber pro insipiente*, or "Book in Behalf of the Fool Who Says in His Heart There Is No God." Gaunilo denied that an idea of a being includes existence in the objective order and that a direct intuition of God necessarily includes God's existence. Anselm wrote in reply his *Liber apologeticus contra Gaunilonem* ("Book [of] Defense Against Gaunilo"), which was a repetition of the ontological argument of the *Proslogium*.

William the Conqueror, who had established Norman overlordship of England in 1066, was a benefactor of the monastery at Bec, and lands in both England and Normandy were granted to Bec. Anselm made three visits to England to view these lands. During one of those visits, while Anselm was founding a priory at Chester, William II Rufus, the son and successor of William the Conqueror, named him archbishop of Canterbury (March 1093). The see had been kept vacant since the death of Lanfranc in 1089, during which period the King had confiscated its revenues and pillaged its lands.

Anselm accepted the position somewhat reluctantly but with an intention of reforming the English Church. He

Appointment as archbishop of Canterbury

refused to be consecrated as archbishop until William restored the lands to Canterbury and acknowledged Urban II as the rightful pope against the antipope Clement III. In fear of death from an illness, William agreed to the conditions, and Anselm was consecrated December 4, 1093. When William recovered, however, he demanded from the new Archbishop a sum of money, which Anselm refused to pay lest it look like simony (payment for an ecclesiastical position). In response to Anselm's refusal, William refused to allow Anselm to go to Rome to receive the pallium—a mantle, the symbol of papal approval of his archepiscopal appointment—from Urban II, lest this be taken as an implied royal recognition of Urban. In claiming that the king had no right to interfere in what was essentially an ecclesiastical matter, Anselm became a major figure in the investiture controversy; *i.e.*, over the question as to whether a secular ruler (*e.g*, emperor or king) or the pope had the primary right to invest an ecclesiastical authority, such as a bishop, with the symbols of his office.

The controversy continued for two years. On March 11, 1095, the English bishops, at the Synod of Rockingham, sided with the King against Anselm. When the papal legate brought the pallium from Rome, Anselm refused to accept it from William, since it would then appear that he owed his spiritual and ecclesiastical authority to the king. William permitted Anselm to leave for Rome, but on his departure he seized the lands of Canterbury.

Anselm attended the Council of Bari (Italy) in 1098 and presented his grievances against the King to Urban II. He took an active part in the sessions, defending the doctrine of the *Filioque* ("and from the Son") clause in the Nicene-Constantinopolitan Creed against the Greek Church, which had been in schism with the Western Church since 1054. The *Filioque* clause, added to the Western version of the Creed, indicated that the Holy Spirit proceeded from the Father and Son. The Greek Church rejected the *Filioque* clause as a later addition. The Council also reapproved earlier decrees against investiture of ecclesiastics by lay officials.

The satisfaction theory of redemption

When Anselm left England, he had taken with him an incomplete manuscript of his work *Cur Deus homo?* ("Why Did God Become Man?"). After the Council of Bari, he withdrew to the village of Liberi, near Capua, and completed the manuscript in 1099. This work became the classic treatment of the satisfaction theory of redemption. According to this theory, which is based upon the feudal structure of society, finite man has committed a crime (sin) against infinite God. In feudal society, an offender was required to make recompense, or satisfaction, to the one offended according to that person's status. Thus, a crime against a king would require more satisfaction than a crime against a baron or a serf. According to this way of thinking, finite man, since he could never make satisfaction to the infinite God, could expect only eternal death. The instrument for bringing man back into a right relationship with God, therefore, had to be the God-man (Christ), by whose infinite merits man is purified in an act of cooperative re-creation. Anselm rejected the view that man, through his sin, owes a debt to the devil, and placed the essence of redemption in individual union with Christ in the Eucharist (Lord's Supper), to which the sacrament of Baptism (by which a person is incorporated into the church) opens the way.

After completing *Cur Deus homo?* Anselm attended a council at the Lateran (papal palace) in Rome at Easter 1099. One year later, William Rufus died in a hunting accident under suspicious circumstances, and his brother Henry I seized the English throne. In order to gain ecclesiastical support, he sought for and secured the backing of Anselm, who returned to England. Anselm soon broke with the King, however, when Henry insisted on his right to invest ecclesiastics with the spiritual symbols of their office. Three times the King sought an exemption, and each time the Pope refused. During this controversy, Anselm was in exile, from April 1103 to August 1106. At the Synod of Westminster (1107), the dispute was settled. The King renounced investiture of bishops and abbots with the ring and crosier (staff), the symbols of their office. He demanded, however, that they do homage to him prior to consecration. The Westminster Agreement was a model for the Concordat of Worms (1122), which settled for a time the lay-investiture controversy in the Holy Roman Empire.

Anselm spent his last two years in peace and died on April 21, 1109. In 1163, with new canons requiring approvals for canonization (official recognition of persons as saints), Archbishop Thomas Becket of Canterbury (1118?–1170) referred Anselm's cause to Rome. Anselm was probably canonized at this time, for the Canterbury records for 1170 make frequent mention of the pilgrimages to his new shrine in the cathedral. For several centuries he was venerated locally. Clement XI (pope from 1700 to 1721) declared Anselm a doctor (teacher) of the church in 1720. His feast day is April 21.

BIBLIOGRAPHY. Major bibliographies are U. CHEVALIER, *Répertoire des sources historiques du moyen âge*, vol. 1, pp. 256–259 (1905); for more recent periodical literature in English, French, German, and Italian, see F.W. BATESON (ed.), *Cambridge Bibliography of English Literature*, vol. 1, pp. 282–283 (1957); and F. VAN STEENBERGHEN, *Philosophie des Mittelalters* (1950).

For Anselm's life and thought, see M. RULE, *Life and Times of St. Anselm*, 2 vol. (1883), and Rule's edition of Eadmer's *Vita Anselmi*, "Rolls Series" (1884); R.W. CHURCH, *St. Anselm* (1870, many reprints); J. CLAYTON, *St. Anselm: A Critical Biography* (1933); J. MCINTYRE, *St. Anselm and His Critics: A Re-interpretation of the Cur Deus Homo* (1954); W.R.W. STEPHENS, "Saint Anselm," in *Dictionary of National Biography*, vol. 1, pp. 482–503 (1908); E. GILSON, *History of Christian Philosophy in the Middle Ages* (1955); F. COPLESTON, *History of Philosophy*, vol. 2 (1950); F. CAYRE, *Manual of Patrology and History of Theology* (1940); Z.N. BROOKE, *The English Church and the Papacy* (1931, reprinted 1968); R. FAIRWEATHER (ed. and trans.), *A Scholastic Miscellany: Anselm to Ockham* (1956); R.W. SOUTHERN, *St. Anselm and His Biographer* (1963).

(J.A.Ke.)

Anseriformes

The order Anseriformes includes the well-known ducks, geese, and swans (family Anatidae) and the little-known screamers (family Anhimidae). The Anatidae comprise about 142 species of medium to large birds, usually associated with aquatic or marine habitats and known collectively as waterfowl (in the United States) or wildfowl (in Europe). These birds are of interest to many people for varied reasons: to the romantic and the artist for their swift, massed flights and association with wild places; to the sport hunter as a difficult and worthwhile quarry; to the economist for exploitation as producers of meat and feathers; to the farmer as possible competitors for grass or as damagers of crops; to the medical man as a possible reservoir of arthropod-borne viruses. The conservationist views them as biological indicators of the richness of wetland habitats and, like those habitats, as objects for his urgent attention. To the zoologist they are a close-knit group in evolution, but one that has adapted into a wide variety of ecological niches.

GENERAL FEATURES

Members of the Anatidae vary greatly in size, the largest being some 60 times as heavy as the smallest. Most are rather chunky birds (see Figure 1), with lengths from just under 30 centimetres to slightly over 150 centimetres and weights from about 300 grams to more than 18 kilograms. The neck is medium to long. The bill also varies widely, as befits birds with diets as different as grass, fish, and minute plankton, but is of medium length, typically broad and somewhat flattened, with a rounded tip, often with a pronounced nail, giving a slight hook to the tip. The bill is often provided with fine lamellae, short parallel plates at the edges of the mandibles that aid in food handling and, in some species, in straining tiny food organisms from water or mud. Anseriform behaviour patterns are an intriguing mixture of the innate and the learned; some so fixed that they can be used as taxonomic characters, others so pliable that completely new traditions can evolve within a few generations.

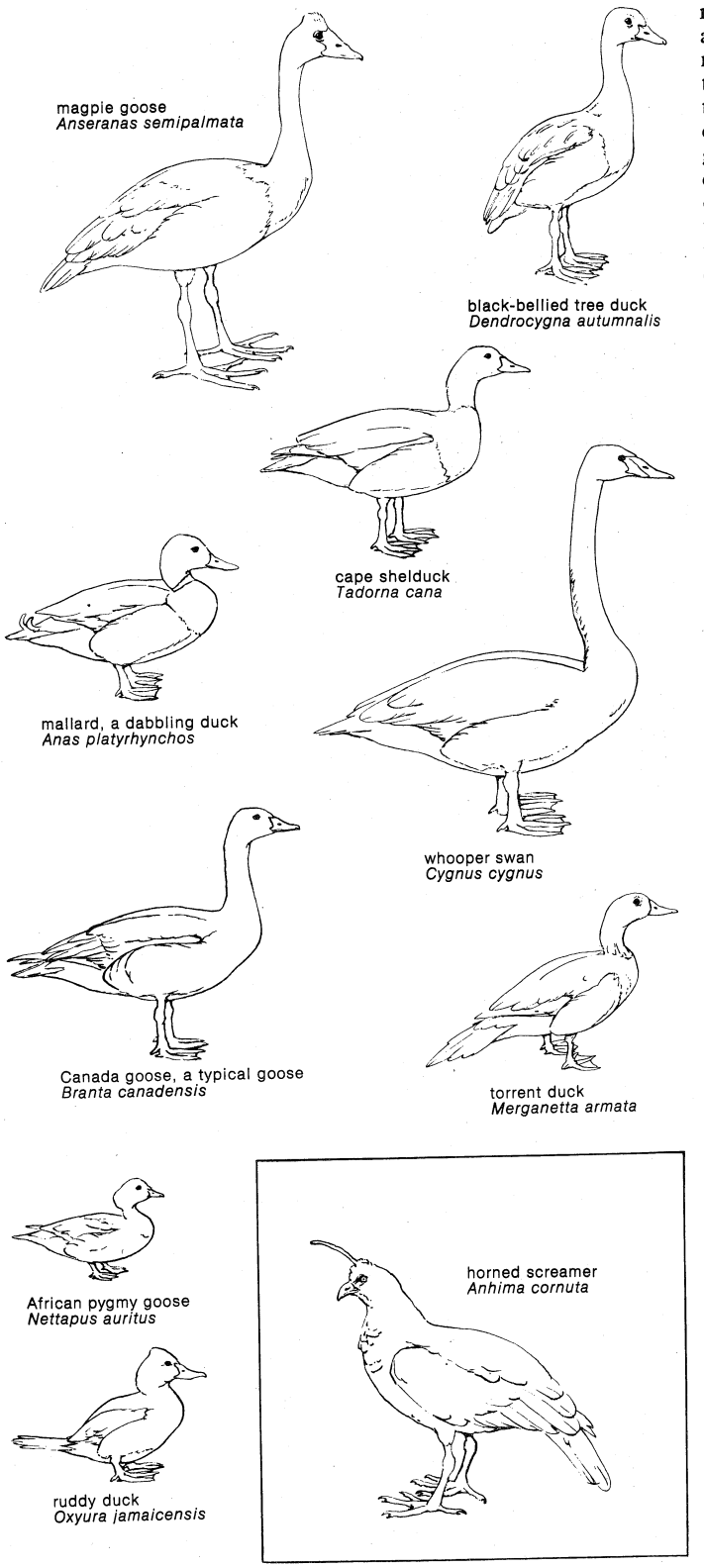


Figure 1: Comparative body plans of anseriform birds.
Drawing by D. Malick

The three species of screamers are quite different from waterfowl in general appearance. They are moderately long-legged birds about the size of a turkey, with lengths from about 70 to 95 centimetres and weights of 2–5 kilograms. The beak is rather chicken-like, with a pronounced downward hook. The feet are exceptionally large, the toes webbed only at their bases.

Anseriform birds breed on every major continent and island except the Antarctic and many undertake vast migrations over thousands of kilometres. Some species are widespread and number in the millions; other are represented by a few hundred confined to a single island. In the former category come the pintail (*Anas acuta*) and the mallard (*Anas platyrhynchos*), both found throughout the Northern Hemisphere; in the latter, the Hawaiian goose (*Branta scandvicensis*) and the Madagascar white-eye (*Aythya innotata*). Extinction has overcome four species within the last century, and island forms are very vulnerable. The Hawaiian goose was once down to fewer than 50 birds, the Laysan race of the mallard to seven individuals. Even the more common waterfowl, which teem in thousands at certain times and places, are not secure from drastic declines in numbers.

IMPORTANCE TO MAN

The large size of waterfowl has made them an inevitable quarry for man from prehistoric times. Not only do they provide protein but also large amounts of fat and feathers, especially down, much prized in the colder regions. Among the unusual uses of waterfowl parts may be mentioned the conversion of swan tracheae into children's whistles in Lapland, and the eating of the bill-knob of the king eider (*Somateria spectabilis*) as an aphrodisiac in Greenland. Wary and difficult to approach in their watery haunts, waterfowl exercised man's ingenuity to kill them before the advent of efficient weapons. The period of flightlessness was early discovered and exploited, the birds being driven into corrals of stone or netting. From the latter evolved the Dutch method of catching full-winged ducks by enticing them up large net-covered pipes leading from the secluded pond of a duck-decoy. The birds were seduced into the pipes by food or by their predator-following reaction to a trained dog. Many other ingenious traps were devised, from the clap-nets of the ancient Egyptians to the rocket-propelled nets of today's research workers.

Threatened species

Anseriform populations are now so threatened by the loss of their essential wetland habitats to drainage and technological developments that they are no longer able to withstand commercial exploitation. This is generally recognized in civilized countries, where market hunting is banned and efforts made to control sport hunting so that annual kill does not exceed annual production. As waterfowl migrations pay no heed to national boundaries, the need for international agreement on their conservation is paramount.

As the amount of human leisure increases so does the importance of nondestructive recreational use of waterfowl. A great flock of waterfowl is one of the remaining spectacles of nature and one that can be made accessible to crowds of people led through covered approaches to concealing blinds. The refuges created to hold reserves of stocks for shooting will, more and more, be developed with the spectator in mind. Such refuges must be numerous and large enough to give adequate feeding grounds. Extra food, however, can be provided as standing crops or distributed from stores. In these ways the occurrence of damage to agricultural interests in the neighbourhood can be avoided. Complaints of such damage are frequently exaggerated, for in many countries migratory waterfowl arrive after the harvests have been gathered and take only spilled or rejected grain and tubers, performing a useful cleaning up operation. Even when geese graze growing cereals in winter, the subsequent crop yield is not reduced. Conflicts can arise, but these can be resolved by changes in husbandry and by scaring techniques.

Another aesthetic exploitation of waterfowl has been possible because they readily adapt to captivity. Spending much time walking, swimming, or diving, they are less inconvenienced than most other birds by being pinioned (the tip of one wing removed) to prevent flight. They can then be kept in open paddocks instead of being confined in cages. Zoos and private aviculturists maintain waterfowl collections of varying size, the largest being that of the Wildfowl Trust, Slimbridge, England. Here no less than 131 of the 142 existing species have been maintained, and more than a hundred have bred. Such a re-

markably complete comparative collection has, of course, been useful for educational and research purposes. Breeding in captivity also enhances the possibility of restoring wild populations that have been reduced by overshooting or other excesses. The most notable case has been that of the Hawaiian goose. In 1950 fewer than 50 of these birds existed. Twenty years later, thanks to avicultural efforts at Slimbridge and on Hawaii itself, there were more than a thousand, several hundred of which had been liberated on the original habitat. For a different goal, but also effective, have been the efforts of hunting organizations to breed replacements for birds shot. Several species of geese and ducks have responded well to manipulation of habitat and the provision of nest sites. Artificially encouraged breeding may well become vital as the hitherto untouched vastnesses of northern marshes and tundras, where so many waterfowl breed, are opened up, destroyed, and polluted.

Domestication of waterfowl

In view of their ready adaptation to captivity, it is surprising that few waterfowl have been domesticated. The mallard was exploited 2,000 years ago in China, and 17 varieties have been developed, according to whether meat or egg production was important. The muscovy duck (*Cairina moschata*) was domesticated in Colombia and Peru before the arrival of the *conquistadores*. The greylag goose (*Anser anser*) has been domesticated for at least 4,000 years. Egyptian frescoes of that age already show changes in shape from the natural form, and eight main varieties are now known. The swan goose (*Anser cygnoides*) of eastern Asia has also been domesticated, with three varieties. Other species such as the Canada goose, mute swan, and Egyptian goose (*Alopochen aegyptiacus*) have been kept in semidomestication for ease of exploitation but without intensive breeding to change their forms. A remarkable form of exploitation has been that of the common eider. Its breeding colonies in the Arctic and sub-Arctic are protected and concentrated by the provision of nest sites and other techniques. The down with which the female lines her nest and covers the eggs is systematically collected during and after incubation, without disturbance or loss of productivity, for these birds are very tame.

NATURAL HISTORY

Life history. Young waterfowl are hatched with a complete covering of down and can take to the water as soon as they leave the nest, within 24 to 48 hours. From the first, they forage for themselves, but at least one parent remains with them, guarding, guiding and, initially, brooding them at night and during inclement weather.

Growth and molt

The downy plumage is retained for from two weeks (small ducks) to six weeks (large swans) and is then replaced gradually by the feathers of the juvenile plumage. The flight feathers, the last to develop, likewise vary in rapidity of growth, taking from five weeks to as many months. At fledging, young ducks are freed from parental care and must make their own way on migration. In species breeding in the far north, this begins soon after fledging, in early autumn. Young geese and swans remain with their parents during their first winter, migrating to and from the wintering grounds in their company. In most large species the juvenile plumage is retained through much of the first year of life. Ducks, however, begin to lose the juvenile body feathers almost at once. Some replace the juvenile plumage with an immature nonbreeding (or "basic") plumage, acquiring the first nuptial (or "alternate") plumage in the second autumn. Other species molt directly from juvenile to nuptial and are practically indistinguishable from adults in plumage and size, at the age of six months. Swans and geese do not reach full size until the end of their second year, and even at that age swans still retain some immature feathers. In migratory species social flocking and pair formation occur on the wintering grounds, followed by the return migration to the breeding grounds. In the high-Arctic regions the birds arrive ready to nest as soon as the snow cover goes and the water opens; further south the process is more leisurely. When the female duck begins to incubate, the male generally deserts her and joins forces with other males, often after making a quite extensive molt-migration, a movement to another area some distance from the breeding site. The nuptial plumage is lost, and a dull "eclipse" plumage, rather female-like, is assumed before the simultaneous molt of the flight feathers. The resulting flightless condition lasts three or four weeks, during which the birds skulk in thick cover or remain on large bodies of water. In female ducks this wing molt is delayed until after the young have fledged. The eclipse plumage is succeeded, in three to six months, by the next nuptial plumage. In the case of swans and geese, the male remains and molts with the female and family; immature birds and unsuccessful breeders may, however, make a molt-migration to a separate area.

The foregoing life cycles are typical of waterfowl breeding in north temperate or Arctic regions. In the tropics and south temperate regions, migrations are often absent or occur in response to less regular ecological factors such as the onset of the rainy season. Males and females tend to be similar in plumage and to lack eclipse plumages. The birds are ready to breed at any time that conditions are favourable. Undoubtedly there is a relationship between hormonal balance and the appearance, or absence, of eclipse or brightly coloured plumages, but the exact mechanism is not yet known.

The mature bird has a potentially long life but this is seldom achieved in the wild. In captivity, ducks have lived for twenty years, geese and swans for more than thirty. Wild populations suffer heavy losses (up to 70 percent) among first-year birds. The adults then experience annual mortalities from 10 percent to 50 percent, depending on the extent to which the population is shot and on natural factors. At the higher end of this scale the population is almost completely replaced within three years. With luck and cunning an occasional bird may survive for 15 or 20 years.

Behaviour. Maintenance of the plumage in an effective waterproof condition occupies much of the time not spent feeding or sleeping. The bill is used both to stimulate the oil gland (situated above the tail) and to spread the oil. Rubbing the chin and throat on oiled areas also helps the process. Preening occurs at the same time, the fine structure of the feathers being nibbled into the interlocking position necessary to prevent the entry of water. Rearrangement of the feathers involves preening, scratching with the feet, and a general body shake, produced by a muscular contraction sweeping from tail to neck. Various wing stretching movements settle the flight feathers. Bathing movements include dipping the head, beating the wings on the surface and, at high intensity, actual diving or somersaulting through the water. Sleep often follows such maintenance activities, the bill being turned and placed under the scapular (shoulder) feathers. Bathing is often a communal activity, and mutual preening is seen in several species.

Display behaviour

Social displays are one of the most interesting aspects of wildfowl behaviour, and many of the signal movements involved in the displays are clearly derived from maintenance activities. Thus preening dorsally, on the breast, and, especially, behind the wing can be seen in ritualized form in social situations. Likewise, the wing stretch and the general body shake occur in threat or sexual displays.

Most waterfowl are gregarious and have well-developed social integration signals. Many indicate the intention to fly by head-shaking or chin-lifting, so that the group takes off together. In swans and geese vocalization plays a part in this situation and also in group flight. Formal aggressive displays are important in gregarious species to prevent actual fighting and injury and to establish instead a stable dominance order. In threat, the behaviour often tends to increase the apparent body size, as when swans raise their wings and ruffle the scapular feathers. Wing flaps and flicking of the folded wings are common in geese. The chin may be lifted during aggressive display,

but more frequently the head is thrust forward, often with the bill open. The striking nature of these gestures is often emphasized by patches of colour, which may be concealed before the movement is made. Since these gestures may occur without associated feather patterns this suggests that the behaviour patterns evolved first.

The pair-forming displays are well developed and characteristic of the species, which is necessary if mating with closely related and sympatric (coexistent) species is to be avoided. Swans and geese (tribe Anserini) cement the pair bond by the "triumph ceremony," with mutual head waving and calling, typically when the male has driven off an intruder. Male sheldgeese (tribe Tadornini) have a puffing, strutting display. Their females incite them to attack other birds by sideways jabbing movements of the bill. Female incitement behaviour is found throughout the rest of the family, but the male pairing behaviour is more varied. It is particularly striking in the dabbling ducks (Anatini), where it is often social, a group of males displaying around a solitary female, who appears to do the mate selection. Ethologists have endeavoured to give objective names to the various postures and sequences (Figure 2). Thus "head-up-tail-up" involves a simultaneous upward jerk of head and tail and a lifting of the folded wings to display the speculum, a set of metallic coloured secondary flight feathers of the upper wing. The "grunt-whistle" involves throwing an arc of water at the female by a sideways flick of the bill, followed by a rearing up of the body, shaking of the head and tail, and, during the whole sequence, giving the call indicated by the behavioural term. The sea ducks (Mergini) also have elaborate male displays, bowing while producing dovelike coos; flagging the head from side to side; jerking the head back on the tail and kicking up a spurt of water (the "head-throw-kick" of the goldeneye, *Bucephala clangula*). Perhaps the most bizarre display is found in the stifftails (Oxyurini), such as the ruddy duck (*Oxyura jamaicensis*) of North America. The males cock their tails over their backs, inflate their lower necks, and beat on them with their bright blue bills, producing a chittering sound which terminates in a burp.

Displays that maintain the pair bond are less varied and tend to be more marked in the long-lived geese and swans, which pair for life, than in the short-lived ducks. Mutual preening and drinking displays are of this category, as are precopulatory displays, often ritualized feeding movements such as mutual head dipping, bill dipping, or head pumping. Following copulation, other mutual displays occur. The geese and swans indulge in calling with upstretched necks and lifted wings. More extraordinary is the "step-dance" of the whistling ducks (Dendrocygnini), in which the partners rear up side by side, treading water and each raising the outside wing.

The display patterns, whatever their origin, have been incorporated in the genetic code and appear ready-formed and complete. Although much wildfowl behaviour can thus be said to be innate, other activities are learned during the life of the individual. One special type of learning, imprinting, has been much studied in this group. The newly hatched, downy young exhibit a strong but unspecific following response to a large moving object, especially if it emits rhythmical sounds. While following such an object they become "imprinted" with the knowledge of its characteristics and thereafter treat it as the parent, even though it may be wildly unsuitable, a balloon for instance. The class of object is not necessarily treated as the sexual partner when the duck matures, but it does appear that male dabbling ducks learn the plumage patterns of the female from the ducks with which they are reared. The females, however, are apparently able to ignore such early experience, and to select the appropriate male when they mature; the male plumage patterns are probably sufficiently simple that their recognition is coded into the female's genetic inheritance.

Imprinting

Drawing by D. Malick based on Paul A. Johnsgard, *Handbook of Waterfowl Behaviour*
Copyright © 1965 by Cornell University. Used by permission of Cornell University Press.

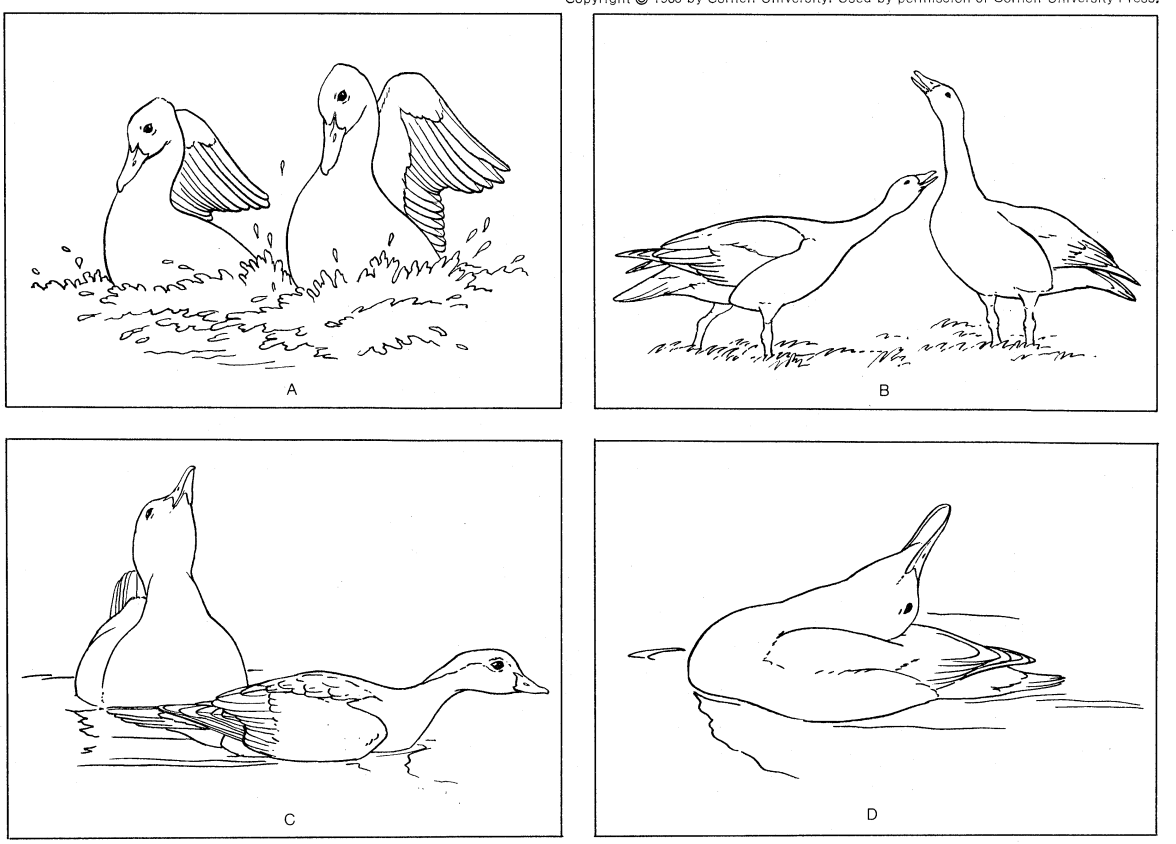


Figure 2: *Waterfowl behaviour.*
(A) Postcopulatory joint display of the fulvous tree duck (*Dendrocygna bicolor*).
(B) Triumph ceremony of the bar-headed goose (*Anser indicus*). (C) Female copulatory position of the wood duck (*Aix sponsa*). (D) Head throw display by male canvasback (*Aythya vallisneria*).

Many other aspects of waterfowl behaviour show a similar interweaving of innate and learned responses. Thus the young bird hatches with distinct preferences as regards the colour, size, and shape of objects at which it will peck. Within these limits, and those imposed by its bill structure and the habitat to which it is taken by its parent, the young bird can be quite adventurous in its feeding, gradually excluding unpalatable or unsuitable objects.

Feeding behaviour

Three main lines of feeding behaviour have evolved in the wildfowl—diving, dabbling, and grazing. Those that dive for food fall into two groups: inland species (Aythyini) favouring relatively shallow lakes up to six metres deep and predominantly feeding on plants such as pondweeds; and sea ducks (Mergini), using deeper marine waters and feeding on invertebrates and fish. Dabbling ducks feed on the surface or by upending to reach the food on the bottom, largely plant material. The grazing geese and swans take grasses and sedges on dry land or in marshes and grub for roots both on land and under the surface. Many of the geese have adapted to the various farm crops, new feeding traditions spreading rapidly as the birds of a group learn from one another.

Migratory behaviour is likewise partly innate, partly learned. Young ducks of migratory populations set off in the autumn in approximately the right direction, even if there are no old birds to guide them. Others show a fixed orientation, after capture and release, that has no apparent function and so has been termed "nonsense" orientation. In both cases the direction is innate, and so is the ability to determine it with reference to the sun's position (making allowance for its diurnal movement) or to the patterns of the stars. It appears, however, that the details of the sun movement and star patterns have to be learned, as do the characteristics of the home from which the bird starts and the winter quarters in which it ends. Like other migratory birds, waterfowl have the ability to return to the nesting area when displaced from it, but it is not clear how far they can be truly said to navigate; *i.e.*, to pinpoint their position in two coordinates, again perhaps with reference to sun or stars. In geese and swans, where the young travel with their parents, the possibility exists of their learning the chains of lakes, rivers, and other landmarks down which they fly and so of "map reading" on subsequent migrations. Certainly the migratory behaviour is highly modifiable in these longer-lived birds. New wintering resorts are adopted if they become favourable because of food supply or lack of disturbance. The migratory urge can be overcome completely in geese reared artificially. If the climate is not too severe and food and mates are available, such birds will remain where they grow up, even though left full-winged and free to come and go.

One form of waterfowl behaviour that seems to be largely innate is sound production. Vocalizations are rather simple and do not apparently offer the opportunity for individual elaboration found in songbirds. A range of honks, whistles, grunts, coos, and quacks are produced, mainly in the context of social contact and flock cohesion or in courtship display. The voices of the sexes are generally quite dissimilar. Some nonvocal sounds are also produced by the wings of some species, by inflatable air sacs of others.

Locomotion. Only the screamers (Anhimidae), with their broad wings and light bodies, are able freely to indulge in soaring flight. The magpie goose (*Anseranas semipalmata*) has somewhat similar wings, but typically the wildfowl have strong, rapid flight, the heavy body and relatively small wings giving a very high wing loading. The trumpeter swan (*Cygnus cygnus buccinator*) is probably the heaviest flying bird there is, weights up to 18 kilograms (38 pounds) having been claimed. Trumpeter swans beat their wings about three times a second, the smaller ducks twice as fast. Geese tend to fly in long extended lines, often adopting a V formation. Cruising air speeds of most waterfowl appear to be around 65 kilometres (40 miles) per hour; when pressed they can certainly reach 100 kph. On migration most flocks fly at altitudes of between 300 and 600 metres. Uncommonly they may be seen around 3,000 metres, and the bar-

Flight speeds and altitudes

Drawing by D. Malick

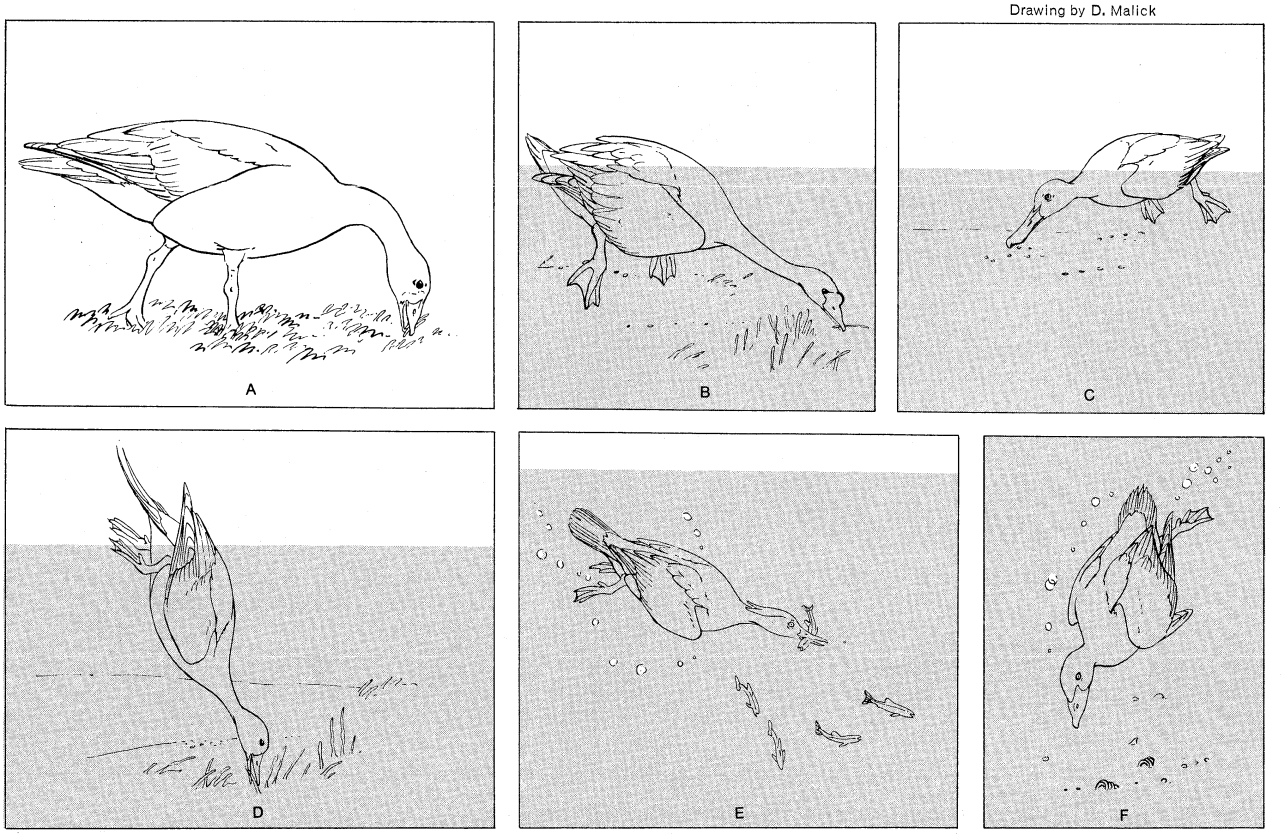


Figure 3: *Feeding characteristics of Anseriformes.*
(A) Grazing; (B) cropping underwater vegetation; (C) filter feeding; (D) dabbling; (E) diving for fish; (F) diving for bottom invertebrates.

headed goose (*Anser indicus*), breeding in Tibet and wintering in India, must fly at 6,000 metres to get through the Himalayan passes.

Swans and the heavier geese require a running start into the wind when taking off from water or land. The Aythyini and the Mergini also skitter along the surface before becoming airborne. Two of the steamer ducks (*Tachyeres*) have lost the power of flight and when in a hurry can only churn along the surface like paddle steamers. Small geese and most ducks can launch themselves directly into the air, pushing with the legs and giving a tremendous thrust with the wings, which may strike the water. In landing, the wings and spread feet are used for braking and guidance. Height is often lost rapidly by alternate sideslips, a process carried to extremes in the downward "whiffling" of geese, in which the birds may actually roll over on their backs, still keeping the head straight and level. The Anatini are quite manoeuvrable near the ground, back-flapping their way vertically down for the last few feet. The Aythyini land faster and normally only on water.

Walking on land is well developed in the longer-legged geese and in gooselike species. The "goose-step," with exaggeratedly lifted feet, is well shown by the spur-winged goose (*Plectropterus gambensis*). Others walk more straightforwardly and when pressed, if flightless for instance, can outrun a pursuing man. In the ducks, whose short legs are situated laterally and posteriorly, the gait is at best a waddle. Most of the Aythyini and Mergini have legs so far back that the body is carried almost upright when out of the water. This development reaches its extreme in the Oxyurini, which rarely leave the water on foot.

The magpie goose, with long, nearly web-free anterior toes and a long posterior hallux, is able to perch on twigs at the tops of trees. But it is very much of an exception. A number of other ducks, especially the hole nesters, can perch on branches or scramble up rough surfaces.

Underwater locomotion

It is for swimming that most of the waterfowl are especially adapted, with their waterproof feathering, fat-insulated body, and powerful legs with webbed feet. The feet paddle alternately in slow swimming, the whole leg being used when the bird is moving fast. All waterfowl will dive if pressed, and about two-fifths use diving as their normal feeding procedure. These submerge by arching the body and propelling themselves forward with both feet so that entry is in a smooth arc. The whistling ducks, however, take a jumping "header," clearing the water like coots, while those consummate divers, the stifftails, simply slip under. Diving in fast-running streams requires great power and adaptations, such as the streamlined form and the long, stiff tail of the torrent ducks (*Merganetta*). Once a diving duck is under water, the legs are sculled together. In some species the wings are opened and used as steering rather than as propelling devices.

Long-tailed, or old squaw, ducks (*Clangula hyemalis*) have been caught in fishing nets as deep as 55 metres, but this is exceptional and most species do not dive much below 6 metres. They normally remain below for less than 30 seconds, occasionally up to 90 seconds, but are physiologically capable of much longer submergence.

Reproductive behaviour. Copulation in waterfowl typically occurs in the water, the pair being in isolation. Indeed, the avoidance of disturbance at this time may be a major function of territorial behaviour. Colonial nesting is not normal, but high densities may occur when certain areas are unusually safe and suitable, islands for instance. More than 200 gadwall (*Anas strepera*) nests per acre have been reported in such circumstances. More usually the nests are scattered, the distance between them being determined in part by the strength of the territorial activity. Some swans, shelducks, and sheldgeese are highly aggressive and defend a territory around the nest. In most species the situation is less clear-cut. The pairs of duck remain within a "home range," incorporating water, a loafing spot, nesting cover, and food, but do not vigorously defend its borders. Rather, the territory proper seems to be an area about the female, moving as she does. A male intruding on a pair gives rise to the "three-bird chase" which is a characteristic of prenesting behaviour. Geese likewise defend an area around the female but also the immediate vicinity of the nest.

Nest building

Nest site selection is done mainly by the female, although the male may accompany her in her search. Nests built over water among marsh vegetation are found in the primitive members of the Anatidae and the protection thus afforded from predators may have tipped evolution toward a more wholly aquatic existence. Many species nest on dry land, however, often, in the Anatini, several hundred yards from water. Some species, especially the shelducks (Tadornini), use crevices and holes for nest sites. They do not construct their own but take over what is available, such as rabbit burrows. Most hole nesters, however, utilize holes in trees, again pre-existing ones, and will quite readily adopt nesting boxes where trees ancient enough to have holes are absent. Old tree nests of other species are sometimes used. A successful species like the mallard is adaptable in choice of nest site and may be found in any of the above situations.

The building of the nest is generally the work of the female, though the male assists in some species. It is a general rule that waterfowl are behaviourally incapable of carrying material in the bill, and even in the rare instance of courtship feeding shown by the red-crested pochard (*Netta rufina*), the female has to swim to the male to take his "gift." In most waterfowl, therefore, the nest is constructed of materials on the site within reach of the beak. They are pulled in and dropped sideways over the shoulder, forming a pile as the bird rotates. With the exception of species that share incubation, such as the magpie goose, the whistling ducks, and white-backed duck (*Thalassornis leuconotus*), the nest is progressively lined with down preened from the female's breast. This down is pulled over the eggs when the female leaves, keeping them warm and, in open sites, protected from the sight of predators.

Egg laying usually occurs early in the morning and occupies but a few minutes, though the female remains on the nest for up to an hour. Generally one egg per day is laid, with occasional gaps. The greatest spread recorded is in the velvet or white-winged scoter (*Melanitta fusca*), which takes 15 days to lay its usual clutch of nine eggs. Sperm has been shown to remain viable in a female mallard for a similar period, so it is probable that copulation prior to egg laying is sufficient to fertilize the whole clutch. Clutch size is variable within a species. The specific averages range from three to twelve eggs. There do not appear to be definite relations between clutch size and taxonomic, ecological, or geographical groupings, and its significance in anseriforms is still controversial. There is, however, a tendency for larger clutches to be laid by species that devote little parental care to their offspring. Egg size varies from 3 to 15 percent of the body weight in different species, the relatively largest being produced by the Oxyurini. In all cases the eggshells are immaculate white, green, or brownish. Surface finish ranges from very smooth to dull and large-pored.

The laying of eggs in nests other than their own is of widespread occurrence among waterfowl, and the very large clutches recorded, up to 22 in the mallard, are usually the work of more than one female. In the Aythyini and the Oxyurini such nest-parasitism is so common that it influences the breeding biology of a population. Only one species, the black-headed duck (*Heteronetta atricapilla*) of South America, is an obligate nest-parasite, always laying in the nests of other species.

With few exceptions, only the female incubates, the average period ranging from 22 to 39 days. Usually she leaves the nest periodically to feed, covering the eggs with down. In some species, like the common eider (*Somateria mollissima*), she apparently fasts during the four-week period. Her resulting constipation may be the cause of the extremely evil-smelling feces that she splashes over the clutch if suddenly flushed from it. This

was once thought to deter predators, but experiments have not confirmed this. As incubation proceeds, aided by both the high temperature (around 101° F) and the high humidity maintained in the nest, the female becomes increasingly reluctant to leave the nest at the approach of a predator. If flushed late in incubation or with hatched young, she will indulge in injury-feigning, flapping or dragging her wings in an uncoordinated fashion which serves to distract predators.

Adaptations for short breeding seasons

The length of the incubation period shows little relation to taxonomic position or body size, but both it and the subsequent fledging period are correlated with the latitude at which breeding takes place. The continuous daylight of the Arctic summer enables the young to feed around the clock, but this does not sufficiently offset the fact that there is only a short interval between the melting of the snows and the onset of autumnal bad weather. Selection in the course of evolution has reduced the breeding sequence as far as possible. Thus the small high-Arctic races of the Canada goose (*Branta canadensis*) have so reduced the period from nest initiation to fledging that they can breed in areas having but two snow-free months. Larger races nesting in California take twice as long. Waterfowl are not double brooded in the wild, even when there is time enough. Most will renest, however, if the first clutch is destroyed. The second clutch is generally smaller but may have a higher fledging success in that hatching occurs in the less uncertain weather of late summer. Aviculturists regularly take advantage of this capacity, incubating the first clutch under bantam hens.

Since incubation does not start until the clutch is complete, hatching is generally well synchronized and may be complete within six hours. Another six hours is required for the down to dry, and young birds usually remain in the nest overnight. While in the nest they are brooded by the female and become imprinted with her visual and auditory characteristics. The downy young have sufficient internal food reserves to last two or three days if inclement weather prevents feeding as soon as they leave the nest.

The nesting behaviour of screamers is somewhat different from that of waterfowl. A substantial nest is built on marshy ground or in shallow water, sometimes reaching a height of 30 centimetres or more. The clutch varies from two to six eggs, depending on the species. Young screamers are covered with yellow down and leave the nest when a few days old, following their parents in the manner of young waterfowl.

Knowledge of waterfowl breeding, although far from complete, has come a long way from medieval times when barnacle geese (*Branta leucopsis*) were thought to hatch from barnacles on rocks because no one had then found their high-Arctic nests.

Ecology. Most waterfowl are associated with fresh water, at least during the breeding season. Relatively few inhabit the areas of early igneous rocks, which have deep lakes and fast streams of low productivity. More preferred are glaciated areas rich in shallow, productive lakes and marshes: the "pothole" country of North America, the silted deltas of rivers, the coastal marshes. An area with many small waters will support more ducks than a single large lake, because the extent of land-water margin is important. The longer shoreline produces diverse food organisms and provides isolation for breeding pairs, cover from predators, and shelter from winds. Paradoxically, periodic droughts are advantageous to duck production. The small lakes dry out and their bottoms become colonized with plants. When reflooding occurs, a rich diet of seeds and of invertebrate animals feeding on the decaying vegetation becomes available. Long-term or permanent droughts, such as occasioned by climatic changes or, increasingly, by man's drainage operations, can be disastrous. The widening acreage of artificial water reservoirs only partly offsets the loss, for they are normally large, deep waters, of little use to waterfowl except for roosting.

Migration

The ability to migrate greatly increases the ecological flexibility of waterfowl. They can exploit the summer resources of the northern tundra without having to evolve the encumbering adaptations needed to survive the winter there; they can move great distances to take advantage of local rainfalls in arid countries; they can retire to secluded areas when flightless and then seek the rich harvests of man-exposed environments when fully equipped to face the dangers. There are many patterns of migration, some regular and in fixed, reciprocal directions, others varying with changes in weather. The distances covered also vary widely. The longest waterfowl migrations are probably those of the blue-winged teal (*Anas discors*), which nests up to 60° N in North America and winters down to beyond 30° S, a distance of over 9,600 kilometres (6,000 miles). In the Old World the northern shoveler (*Anas clypeata*) has a similar migration distance of up to about 11,000 km. The northern pintail and the garganey (*Anas querquedula*), which breed in the Siberian tundra and taiga and winter in the tropical swamps of Senegal and Chad, are even more remarkable in their adaptability.

The general timing of regular long-distance migrations has been built into an internal hormonal cycle in the course of evolution. These "annual clocks" are kept in synchrony by changes in day length, but the immediate factors initiating the stages of migration are climatic. A mass southward movement occurs when air pressure systems produce a favourable wind flow; overcast conditions or a warm spell may temporarily halt the birds; bright, cold conditions move them on.

Causes of mortality

Sudden adverse weather may cause considerable mortality, especially where a long sea crossing is involved, as from Greenland to Britain. But the effects of weather on waterfowl populations can be most drastic at the time of breeding. Late frosts and spring floods are disastrous, especially when there is no time for renesting. Consequently, high-Arctic breeders may have near-complete reproductive failures. The brant geese (*Branta bernicla*) of the population wintering in western Europe only produce good crops of young in about half the seasons.

Parasites of many kinds inhabit waterfowl, generally in a state of tolerance, gaining the upper hand only when the bird is stressed in some way, as by food shortage. This is also true for fungal infections, particularly by *Aspergillus*, a common cause of death in captivity and, exceptionally, in the wild. Bacterial diseases take their toll and sometimes cause spectacular die-offs. Thousands of ducks have succumbed to botulism when shallow, brackish waters dry out in warm countries. *Clostridium botulinum* then multiplies, producing toxins that cause a fatal paralysis if ingested. Fowl cholera (*Pasteurella*) sometimes causes epidemics. More frequently nowadays mass mortality is occasioned by pollution, especially oil spills, pesticides, and other poisons. A peculiarly insidious danger is lead poisoning caused by the ingestion of lead pellets that accumulate in the mud of waters frequently shot over.

The direct kill by man undoubtedly makes him a major predator of wildfowl, and this must be increasingly controlled in the future. Adult ducks, other than nesting females, suffer relatively little natural predation, avian or mammalian, unless they are sick or weakened by hunger. The main predatory impact comes on the eggs and downy young. Crows and gulls, mink, raccoons, coyotes, foxes, ground squirrels, snakes, snapping turtles, bullfrogs, pike, and carp all take some toll. In normal circumstances such predation does not unbalance population dynamics, which are, after all, geared to a high rate of loss in the early stages. Predator-control measures have seldom produced any comparable increase in wildfowl production.

Beneficial relations exist with other birds and animals. Thus Aythyini and eiders (*Somateria* species) actually prefer to nest in association with the smaller gulls (*Larus* species). The latter tend to drive off predators and, by increasing the number of available eggs, share the pressure of egg predators. The muskrat clears water openings in reed beds and builds "houses" that are utilized by ducks for standing and nesting. Beaver dams create very

attractive duck habitats. And, of course, man himself is not wholly destructive but creates some new areas for waterfowl and provides them with food, both unintentionally and on purpose.

FORM AND FUNCTION

Anatomy. The characteristic features of the family Anatidae are the skin-covered, lamellate bill and thick, fleshy tongue. The Anhimidae have a short non-lamellate bill, slightly hooked, like that of a domestic hen, and a horny tongue. Their bill is adapted for tearing aquatic plants. Among the Anatidae the basic bill has undergone a wide adaptive radiation. The geese have evolved strong, deep bills with hard, sharp lamellae. In some, such as the red-breasted goose (*Branta ruficollis*), the bill is short and slight, used only for grazing; in others, such as the snow goose (*Anser caerulescens*), it is long and heavy enough to deal with and dig for roots and tubers. The massive digging bill reaches maximum development in the magpie goose. The little pygmy geese (*Nettapus*) are so-called for their gooselike bills, but they actually feed on lotus seeds and water vegetation and do not graze or root for food. The European widgeon (*Anas penelope*), on the other hand, does graze extensively, but its bill differs little from the typical *Anas* bill, flattened with platelike lamellae. This is used for sifting particles out of mud or picking up food items from the lake bottom as the bird upends. The sieving bill is yet further developed in the shovelers and the pink-eared duck (*Malacorhynchus membranaceus*), the lamellae becoming extremely fine, enabling particles as small as diatoms to be taken from the surface film. The blue duck (*Hymenolaimus malacorhynchos*) has a rounded, expanded tip to the bill, which probably protects it when poking around sharp pebbles. The Aythyini have fewer lamellae and a narrower bill than the Anatini. In the mergansers (*Mergus* species) the lamellae have become toothlike projections in the long narrow bill, ideal for holding fish; additionally the tongue has two dorsal rows of barbs. None of the Anatidae have developed pickaxe bills or ones for crushing hard food items, such as mollusks. When such are taken they are broken up in the gizzard.

Variety of bill modifications

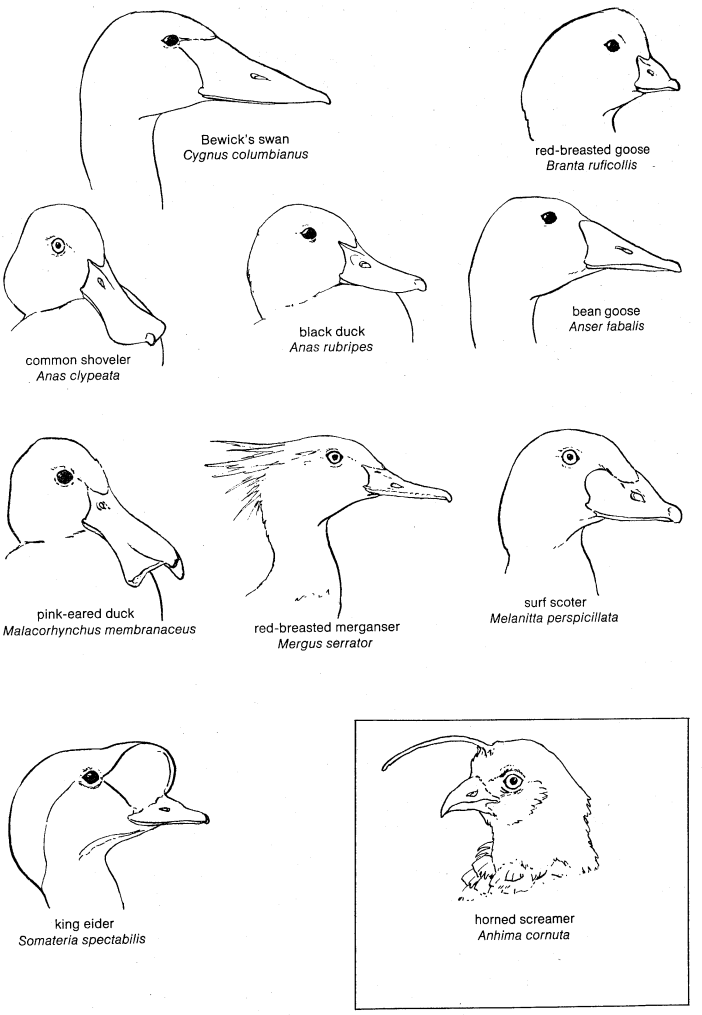


Figure 4: Heads of anseriform birds.

The legs and feet vary according to whether the main mode of progression is walking (geese) or swimming (divers). The latter have flattened tarsi to reduce water resistance on the forward stroke, the hind toe being lobed. Webs are fully developed between the front three digits, except in the magpie goose (primitively) and the Hawaiian goose, where it is believed that the greater flexibility of the toes makes for easier walking on broken lava beds. The claws on the digits are particularly sharp in forms (such as the Cairinini) that normally nest in tree holes.

The wings are not of unusual structure but vary widely in shape, from the broad expanse of the slow and manoeuvrable fliers to the narrow sweeps of the fast movers. Some of the Dendrocygnini have modified vanes on the outer primaries that produce a whirring noise in flight. The soughing creak of the wings of the mute swan (*Cygnus olor*) is well known. In a few cases the wing is so reduced in size that the bird is flightless, as is the Auckland Island race of the brown teal (*Anas aucklandica*). Many waterfowl have horny knobs on the carpal joint; in some there is a distinctly projecting spur as in the screamers, spur-winged goose, and torrent duck. In the first two, the spurs are actually used effectively in striking an adversary; in the last they may be of more use in progressing over slippery rocks in a raging stream.

Specializations of the windpipe

A structure that has received much attention, largely because of its taxonomic value, is the anatid trachea. This may be straight (geese and ducks) or looped in various ways (magpie geese and swans), the most elaborate being that of the trumpeter swan (*Cygnus cygnus buccinator*). Here it first enters the sternum, flexing twice into horizontal and vertical bony pockets, then emerging again to coil around and back into the lungs. The analogy to an orchestral wind instrument is obvious and a comparable sound is produced. Less obvious is the relation of shape to noise in the asymmetrical chamber (bulla) found at the base of the male trachea in most ducks. The male's whistle certainly differs from the female's quack, but it is difficult to account for the variation in bulla shapes—round and solidly ossified (Anatini), angular with membranous "windows" (Aythyini), or complexly membranous (Mergini).

The anseriform birds share with the curassows, tinamous, and ratites the possession of an intromittent male organ. This is not homologous with the mammalian penis but is a vascularized sac everted from the wall of the cloaca. It is protruded by muscular action and retracted by an elastic ligament. This gives it a spiral form, and sperm pass from cloacal papillae along a groove to the tip. This organ doubtless facilitates copulation on the water. Its presence is an aid to research and aviculture, since it permits the investigator readily to sex live birds that lack plumage dimorphism. The size of the organ also allows an estimate of the bird's age. That is, however, more safely done by determining the depth of the bursa of Fabricius, a glandular organ opening into the cloaca and diminishing with maturity.

Physiology. There are interesting adaptations in the respiratory physiology of diving ducks. The accumulation of carbon dioxide does not stimulate breathing as it does in land mammals. When the bird is under water, oxygen is drawn from stored oxyhemoglobin and oxymyoglobin in blood and muscles, instead of from the respiratory system. The air sacs are compressed, reducing buoyancy. Glycogen is broken down by anaerobic and aerobic means, and the blood flow is modified so that the central nervous system receives the main supply.

Rather little is known about the physiology of digestion

in waterfowl. It appears that, despite being almost entirely grazers, geese lack the ability to digest cellulose, either by the secretion of cellulases or by the symbiotic activities of micro-organisms in the gut or its cecae (branches). Such inefficient digestion is probably correlated with the extraordinarily rapid passage of ingested material, which begins to appear in the droppings within half an hour and can be completely eliminated in two hours.

Salt excretion is achieved by the paired supraorbital nasal glands situated on indentations on the skull over the eyes. Races living in marine or brackish environments have these glands enlarged, quite modifying the profile of the head. In a botulism infection, it is probable that the toxins of *Clostridium botulinum* cause these glands to malfunction, resulting in the characteristic salt imbalance and paralysis.

EVOLUTION AND PALEONTOLOGY

Anseriform fragments have been recognized from as far back as the Eocene, and certainly by the late Oligocene birds assignable to the present-day tribes Anserini, Anatini, and (possibly) Oxyurini existed. The Miocene added Dendrocygnini and Aythyini, the Pliocene Tadornini, Mergini, and possibly Cairinini. By the Pliocene several species osteologically indistinguishable from modern species were present. Fossils from the Pleistocene are much more extensive and amount to more than 90 anseriforms. Paleornithologists tend to be conservative in allocating new specific names to Pleistocene anseriforms, since the observed differences correspond to those between modern races, indicating temporal as well as geographical continuity. Nevertheless, 32 of the Pleistocene species are reckoned to be extinct.

Despite the number of fossil species, the fossil record is inadequate to throw light on the origin and evolution of the Anseriformes. The greatest number and variety of their fossils have been reported from the Palaearctic Region, but this does not necessarily mean that the group originated there. Indeed, a number of other considerations point to a tropical origin, possibly in the Southern Hemisphere. At present, the greatest number of endemic and monotypic genera and the more generalized species are to be found there. Be that as it may, the Anseriformes soon spread and were to be found over all their present range quite early in their history.

Main lines of anseriform evolution

It is probable that the ancestral forms were long-legged unspecialized birds with features related to the galliform curassows (Cracidae), to the ciconiiform flamingos (Phoenicopteridae), and to the anseriform screamers and magpie goose. While the screamers have probably advanced but little, the Anatidae have diversified rapidly and extensively. Two main lines may be distinguished. One, typified by the swans and geese, became large-bodied, long-lived, slow-breeding birds, with prolonged pair and family bonds and associated parental care of the young. The other line, characterized by most of the ducks, has favoured small-bodied, short-lived birds having but brief pair bonds and minimal care of the young, necessarily correlated with prolific production of young. Superimposed on these general trends were adaptive radiations; from omnivorous to specialized food habits; from a largely terrestrial to an almost fully aquatic existence. An adaptation for the latter, the double molt, enabled the males to develop striking nuptial plumage, which prevented interspecific hybridization when many species were sympatric. Migratory habits strengthened this requirement and by the same token, gene flow within such species was rapid. Racial differences developed only in nonmigratory or otherwise isolated species, or in those migratory forms that maintained family groupings when away from the breeding grounds, and so developed learned, traditional attachments to the area where they were born.

Despite the wide variety of present-day Anatidae, the family is basically genetically homogeneous. This is shown by the readiness with which hybrids occur, at least in the forcing conditions of captivity. Over four hundred kinds of hybrids are known. Many of the interspecific crosses are fertile, and some remarkable intergeneric and intertribal hybrids are known. The evidence from serological investigations and from the physico-chemical examination of egg-white proteins and egg-shell structure similarly shows the family to be close-knit, only the magpie goose deserving classificatory isolation. Likewise the eggshell structure and nitrogen content throw some light on evolutionary relationships. Feather pigments pose a number of fascinating problems, and recently the composition of feathers has been analyzed to throw light on the geographical origins of goose populations.

CLASSIFICATION

Distinguishing taxonomic features. Broad classification divisions are based on skeletal and other internal structures. Finer divisions are decided on plumage type and patterns, including those of the downy young, and on behavioural features such as the length of pair bond and on displays associated with pair formation and maintenance.

Annotated classification. The classification here follows that of Jean Delacour (1954) as modified by Paul Johnsgard (1965, 1968).

ORDER ANSERIFORMES

Birds with skull desmognathous (maxillopalatines fused) and holorhinal (posterior border of nostrils rounded); one pair of posterior sternal notches; four digits on each foot; wing diastataxic (gap between secondaries four and five); 11 primaries, the outer one reduced; oil gland tufted; eggs unmarked; young nidifugous, with thick down.

Family Anhimidae (screamers)

Pleistocene-Recent. Short, curved, nonlamellate bill; ribs without uncinate process; feet unwebbed, hallux long; plumage uniformly distributed; extensive air cells between skin and body. South America; length 70–95 centimetres, weight 2–5 kilograms; two genera, three species.

Family Anatidae (ducks, geese and swans)

Eocene-Recent. Skin-covered bill, lamellate edges, tipped with horny nail; occipital fontanelles generally present; hallux generally reduced and raised; tongue thick, bordered with spiny processes; primary wing feathers generally molted simultaneously; male copulatory organ present; body length about 30–150 centimetres, weight 300 grams–18 kilograms; worldwide; about 37 genera, 142 species.

Subfamily Eonessinae. Fossil. Known only from middle Eocene of Utah, one species.

Subfamily Romainvilliinae. Fossil. Known only from lower Oligocene of France. One species.

Subfamily Anseranatinae (magpie goose). Rear of head slopes upward and forward; metasternum three-lobed; circular foramen in precoracoid bone; short descending process on lachrymal; anterior toes half webbed, long hallux not raised; long trachea coiled beneath skin; gradual molt of primaries; feeds young from bill; shares features with Anhimidae and with Anserinae. Northern Australia; one genus, one species.

Subfamily Anserinae (swans, typical geese, and whistling ducks). Legs short to medium; neck medium to long. Anterior toes webbed, hallux short, elevated, tarsal scales reticulate. Sexes much alike in plumage, voice, and behaviour. Long pair bond and extended family life. Single annual molt. Vegetarian. Wings usually moderately long, pointed; tail short. All species swim capably, many primarily aquatic.

Tribe Dendrocygnini (whistling ducks). Whistling voice. Relatively long-legged, with upright stance. Enclosed bony orbit (eye socket). Strongly patterned adult plumage; downy young with light stripe below eyes and around nape. Mature at one year. Mutual postcopulatory step-dance. Widely distributed in tropics and subtropics. Two genera, nine species.

Tribe Anserini (swans and geese). Simple trachea and syrinx. Mature in second to fourth year. Behaviour includes "triumph" ceremony and precopulatory head dipping. Breeding distribution Holarctic and north temperate, except for one tropical goose and two species of south temperate swans. Five genera, 21 species.

Tribe Stictonettini (freckled duck). Ducklike with short legs but many anserine features. Australia. One monospecific genus.

Subfamily Anatinae (ducks and aberrant geese). Tarsal scales scutellate (overlapping plates). Male trachea with asymmetrical enlargement (bulla); sexes differ in voice, and also usually in plumage and behaviour; brief pair bond and

family life. Double annual molt; males often alternate a cryptic "eclipse" plumage with a conspicuous nuptial one; downy young strongly patterned. Female "inciting" display present.

Tribe Tadornini (sheldgeese and shelducks). Iridescent green wing speculum and, generally, white wing coverts; sexes alike in majority; mature in second year. Pugnacious. Downy young black and white. Worldwide except North and Central America. Five genera, 15 species (one extinct).

Tribe Tachyerini (steamer ducks). Ponderous birds with short wings; plumage mottled with no metallic colours, sexes alike. Closely related to Tadornini, but differ in display behaviour. Western and southern South America. One genus, three species.

Tribe Cairinini (perching ducks). Heterogeneous group; perching habit; metallic wing plumage; small species sexually dimorphic in plumage, larger ones in size; hole nesters; downy young sharp-clawed and stiff-tailed; pair bond usually short-lived or barely developed. Primarily tropical and subtropical; some temperate. Nine genera, 13 species.

Tribe Anatini (dabbling ducks). Short-legged; lobed hallux. Tracheal bulla round and ossified; mature in first year; iridescent wing specula; female plumage with disruptive patterning; female inciting behaviour very ritualized; female has distinctive decrescendo call; mutual head pumping as precopulatory display; complicated group displays in males. Downy young with eye and cheek stripes. Most feed on water surface or by upending. Worldwide. Five genera, 40 species.

Tribe Aythyini (pochards and scaup). Fresh-water divers, mainly vegetable feeders, with large feet set well back, and strongly lobed hallux. Tracheal bulla angular with membranous "windows"; sexual dimorphism in plumage, but male nuptial plumage not strikingly patterned; mature at one year. Females without decrescendo call; courtship behaviour simple; downy young lack facial stripe. Worldwide distribution. Three genera, 16 species (one extinct).

Tribe Mergini (sea ducks and mergansers or sawbills). Mostly salt-water divers, feeding on animals. Bulla structures often elaborate; sexual dimorphism in plumage, voice and behaviour; male courtship patterns very varied; maturity requires at least two years. Typically Holarctic, one species in South America. Eight genera, 20 species (two extinct).

Tribe Oxyurini (stiff-tailed ducks). Fresh-water divers with legs so far to rear that walking is difficult; lack tracheal bullae, but sound production by inflatable sacs in neck; sexual dimorphism in dense, shiny plumage, male patterns not elaborate; long, stiff tail feathers; elaborate male displays; large eggs and precocial young. Mainly Southern Hemisphere. Three genera, eight species.

Subfamily Paranyrocinae. Fossil. Known only from lower Miocene of South Dakota. One species.

Critical appraisal. There is general agreement on the family and subfamily divisions of the order and at the generic level, where most workers agree on the lumping of species into large genera. Some controversy exists at tribal level. Some workers would include the Tachyerini within the Tadornini; extract the eider ducks from the Mergini and erect for them the tribe of Somaterini; reorganize the admitted "ragbag" of the Cairinini; eliminate Stictonettini and return the freckled duck (*Stictonetta naevosa*) to the Anatini. Indeed it is on the positioning of certain "difficult" species that most discussion revolves. Certain species whose taxonomic positions are in dispute are best thought of as linking species, the result of evolutionary lines that diverged early from near the bases of two major phyletic lines. Examples of such species are the white-backed duck (*Thalassornis leuconotus*), which seems intermediate between the Oxyurini and Dendrocygnini; the Cape Barren goose (*Cereoposis novaehollandiae*), placed in the Tadornini or the Anserini; the ringed teal (*Callonetta leucophrys*, sometimes placed in the genus *Anas*), in the Anatini or the Cairinini; the pink-headed duck (*Rhodonessa caryophyllacea*, extinct since about 1940), in the Anatini or the Aythyini; the coscoroba swan (*Coscoroba coscoroba*), in the Dendrocygnini or the Anserini; and the marbled duck (*Marmaronetta angustirostris*, sometimes placed in *Anas*), intermediate between the Anatini and Aythyini. The study, in recent years, of new taxonomic characters such as behaviour patterns and biochemistry has substantially improved our understanding of relationships within the Anseriformes.

BIBLIOGRAPHY. G.L. ATKINSON-WILLES (ed.), *Wildfowl in Great Britain* (1963), a regional study of distribution and populations; K.M. BAUER and U.N. GLUTZ VON BLOTZHEIM, *Handbuch der Vögel Mitteleuropas*, vol. 2 and 3, *Anseriformes* (1968–69), much wider coverage than the name implies, the most complete source book on the biology of Holarctic waterfowl; H.J. BOYD, G.V.T. MATTHEWS, and M.A. OGILVIE (eds.), *Wildfowl*, formerly *Annual Reports of the Wildfowl Trust* (1950–70), 21 volumes, containing articles on all aspects of the biology of wildfowl; J. DELACOUR, *The Waterfowl of the World*, 4 vol. (1954–64), monographic treatment of all species with coloured plates by Peter Scott; H.J. FRITH, *Waterfowl in Australia* (1967), the only book devoted to waterfowl of the Southern Hemisphere; H.A. HOCHBAUM, *The Canvasback on a Prairie Marsh*, 2nd ed. (1959), a classic species monograph, *Travels and Traditions of Waterfowl* (1955), a review of waterfowl migration and orientation; P.A. JOHNSGARD, *Handbook of Waterfowl Behaviour* (1965), an inventory of behaviour patterns, especially social and sexual, *Waterfowl: Their Biology and Natural History* (1968), contains photographs, many coloured, of nearly every species; F.H. KORTRIGHT, *The Ducks, Geese and Swans of North America* (1942), a thorough treatment, with excellent illustrations; J.P. LINDUSKA (ed.), *Waterfowl Tomorrow* (1964), authoritative articles on all aspects of waterfowl conservation, management, and research; J.G. MILLAIS, *The Natural History of the British Surface-feeding Ducks* (1902), an early classic, still of value, *British Diving Ducks* (1913), two-volume work with much useful information; J.C. PHILLIPS, *The Natural History of the Ducks*, 4 vol. (1922–26), a classic monograph covering many aspects of waterfowl biology; P. SCOTT, *A Coloured Key to the Wildfowl of the World*, rev. ed. (1968), contains coloured plates of all species of the Anatidae, with details of distribution and aids to identification.

(G.V.T.M.)

Antarctica

Lying almost concentrically about the South Pole, Antarctica—the name of which means "opposite to the Arctic"—is the southernmost continent, a circumstance that has had momentous consequences for all aspects of its character. It is the fifth largest continent, covering about 5,500,000 square miles (14,200,000 square kilometres), and would be essentially circular except for the outflaring Antarctic Peninsula, which reaches toward the southern tip of South America (some 600 miles away), and for two principal embayments (see ROSS SEA; WEDDELL SEA). These deep embayments of the southernmost Pacific and Atlantic oceans make the continent somewhat pear shaped, dividing it into two unequal-sized parts. The larger has been called Greater Antarctica but is generally known as East Antarctica because most of its lies in east longitudes. The smaller, or Lesser, Antarctica, wholly in west longitudes, is generally called West Antarctica. East and West Antarctica are separated by the Transantarctic Mountains. Whereas East Antarctica consists largely of a high, ice-covered plateau, West Antarctica consists of an archipelago of mountainous islands covered and bonded together by ice.

The overlying continental ice sheet

The continent is almost wholly overlain by a continental ice sheet, containing approximately 7,000,000 cubic miles of ice and representing about 90 percent of the world's ice. The average thickness is about 6,500 feet. Many of the deep embayments of the Ross and Weddell seas are covered by ice shelves, or ice sheets floating on the sea. These shelves—the Ross Ice Shelf (*q.v.*) and the Ronne and Filchner ice shelves—together with other shelves around the continental margins, constitute about 10 percent of the area of Antarctic ice. Around the Antarctic coast, shelves, glaciers, and ice sheets continually "calve," or discharge, icebergs into the seas.

Because of this vast ice, the continent supports only a primitive indigenous population of cold-adapted land plants and animals. The surrounding sea is as rich in life as the land is barren. With the decline of whaling and sealing, the only economic base in the past, Antarctica now principally exports the results of scientific investigations that lead to a better understanding of the total world environment. The present scale of scientific investigation of Antarctica began with the International Geophysical Year (IGY) in 1957–58. Although early explorations were nationalistic leading to territorial claims,

modern ones have come under the international aegis of the Antarctic Treaty. This treaty, which was an unprecedented landmark in diplomacy when it was signed in 1959 by 12 nations, preserves the continent for nonmilitary scientific pursuits and prohibits future claims based on activities during the 30-year treaty term.

Antarctica, the most remote and inaccessible continent, is no longer the most unknown area in the world. Nearly all mountain regions have been aerially photographed, if not topographically mapped. Most have been visited by teams of geologists, geophysicists, glaciologists, and biologists. The 1,900-mile-long chain of the Transantarctic Mountains is completely surveyed, and maps at various scales now exist for most exposed mountain areas. Many hidden ranges and peaks are known from geophysical soundings of the Antarctic ice sheets. Probably within the near future, systematic aerial studies using newly developed radio-echo sounding instruments will make the ice-buried terrains almost as well mapped as the exposed ones.

The ice-choked and stormy seas around Antarctica long hindered exploration by wooden-hulled ships. No lands break the relentless force of the prevailing west winds as they race clockwise around the continent, dragging westerly ocean currents along beneath. The southernmost parts of the Atlantic, Pacific, and Indian oceans converge into a cold, oceanic water mass with singularly unique biologic and physical characteristics. Early penetration of this Antarctic (or Southern) Ocean, as it has been called, in the search for fur seals led in 1820 to the discovery of the continent. Icebreakers and aircraft now make access relatively easy, although still not without hazard in stormy conditions. A few tourists have now visited Antarctica, and it seems likely that scenic resources have greater potential for near future economic development than do mineral and biologic resources.

Scope of the Antarctic regions

The term Antarctic regions refers to all areas—oceanic, island, and continental—lying in the cold Antarctic climatic zone south of the Antarctic Convergence, an important and seasonally little variable oceanic boundary where warm subtropical waters meet and mix with cold polar waters. For legal purposes of the Antarctic Treaty, the arbitrary boundary of 60° south latitude is used. The familiar map boundaries of the continent known as Antarctica, defined as the South Polar land-mass and all its nonfloating grounded ice, are subject to change with future changes of climate. The continent was ice free during most of its lengthy geologic history, and there is no reason to believe it will not become so again in the probably distant future. When it does, the inhabitants will know a continent (East Antarctica) and nearby island archipelago (West Antarctica) of about the size of Australia and Indonesia today. Effects of the melting ice will be world-wide, for sea level will rise possibly as much as 150 or 200 feet, and global oceanic and atmospheric circulation patterns will change drastically.

This article is divided into the following sections:

I. Physical characteristics
  Geologic record
  Physical geography
II. Flora and fauna
  Vegetation
  Animal life
III. Natural resources and resource development
  Mineral resources
  Nonmineral resources
IV. Territorial evolution and political status
  Historical background
  IGY and the Antarctic Treaty

I. Physical characteristics

GEOLOGIC RECORD

Antarctica and continental drift. The geological evolution of Antarctica has followed a course generally similar to that of the other southern continents. The earliest chapters in Antarctica's fragmentary record extend far back, perhaps as much as 3,000,000,000 years, into early Precambrian time. Similarity in patterns of crustal and biologic evolution in the southern continents extends up until some 150,000,000 years ago, and evolutionary courses began to diverge conspicuously by about 70,000,000 years ago, or the early Cenozoic Era. Plant and animal migration routes that had apparently freely interconnected all southern continents were largely cut off by the beginning of the Cenozoic. Antarctica became isolated at a time when land mammals diversified and flourished elsewhere, populating all other continents of the world. The growth of great ice sheets early in the Cenozoic, perhaps as long ago as 40,000,000 to 50,000,000 years, was a profound event the effects of which were felt worldwide.

Now bathed by polar ice, Antarctica has abundant fossil evidence that its climate and terrain once supported far more populous flora and fauna than today's few seedless plants and primitive insects. Remains of a luxuriant extinct flora and newly discovered fossils of Mesozoic reptiles and amphibians, in particular, compare so closely to those of other southern continents that many geologists postulate former contiguity of these lands in a single, giant continent called Gondwanaland (see CONTINENTAL DRIFT). Continental stratigraphic evidence and the dating of sea floors indicate that, if such a supercontinent existed, some parts of it probably broke up in Late Mesozoic or Early Cenozoic times.

Supercontinent of Gondwanaland

Modern theory ties mobile zones to the interaction and jostling of immense crustal plates. Modern plate boundaries may be far different from ancient ones presumably marked by old fold belts. Ancient Antarctic mobile belts, such as followed by today's Transantarctic Mountains, terminate at continental margins abruptly, as if sliced off, and seemingly reappear again in other lands across young ocean basins. Much work now concentrates on attempting to match intercontinentally the detailed structure of opposed coasts, such as between Antarctica and Australia, in an effort to learn whether they had been actually connected before the latest cycle of crustal spreading from intervening midoceanic ridges. These ridges are now the principal sites of earthquake activity in Antarctic regions.

Structural framework. Most of the Antarctic geologic record lies hidden beneath vast regions of snow and ice making up more than 95 percent of the surface terrain. No one yet knows what important segments of the record lie concealed in buried ranges such as the newly found Gamburtsev Mountains, the topography of which has been mapped only by seismic reflections through the great East Antarctic Ice Sheet. The extraordinarily thick cover, the extremely difficult working conditions, and the great expense of mounting expeditions into remote areas have long held geologic knowledge of Antarctica far behind that of other continents. Great advances through the 1960s by geologists of many Antarctic Treaty nations, however, have yielded geologic maps of at least reconnaissance scale for nearly all exposed mountain areas.

From results mainly of British expeditions early in this century, the concept arose that Antarctica is made up of two main structural provinces—a stable mass called the Precambrian Shield of East Antarctica, between about 3,000,000,000 and perhaps 500,000,000 years in age, and a 50,000,000–200,000,000-year-old mobile belt of West Antarctica, separated by a lengthy fault-block belt, or horst, of the Transantarctic Mountains. East and West Antarctica have come to be known respectively as the Gondwana and Andean provinces, indicating general affinities of each sector with other regions—that is, the east seems to have affinity with the Gondwana region of peninsular India, and the west seems to represent a southerly continuation of the South American Andes. As new expeditions study and restudy each range in ever-increasing detail, concepts of the geological structure are continually modified. Antarctica's structural record is now known to be more complex than that implied in the past by the simple two-fold continental subdivision.

Crustal thickness and mobility

The average thickness of the Earth's crust for both East and West Antarctica approximates that of other continents. Although it has been postulated that West Antarctica might be an oceanic island archipelago if the ice were to melt, its crustal thickness of about 20 miles indicates an absence of oceanic structure. This

thickness is similar to that of coastal parts of other continents. The crust thickens sharply along the Transantarctic Mountains front, possibly a deep crustal fault system, and averages about 25 miles thick in East Antarctica. No earthquakes have occurred in historic time along these or other faults in Antarctica, now the quietest and most aseismic of all continents.

The ancient crust of Antarctica must have been highly mobile and the configuration of the continent many hundreds of millions of years ago in the Precambrian far different from today's. Ancient basins, some of a type called geosynclines, were filled with a variety of sedimentary and volcanic debris eroded from primeval lands. During mountain-building episodes these materials were complexly deformed and recrystallized deep within the crust to form, particularly in East Antarctica, great crystalline-rock complexes. At the surface, rocks were uplifted and mountains were carved by erosion as sediments filled new basins, and new folds of the Earth's crust were formed. Again and again this cycle was repeated during the evolution of Antarctica. Mobility ceased about 400,000,000 years ago in the Transantarctic Mountains. Between that time, in the Devonian Period, and about 160,000,000 years ago, in the Jurassic Period, a series of mainly quartzose sediments was laid down in ancient lakes and shallow seas in the sites of former mountain chains that had been carved away by erosion. Called the Beacon Sandstone, this formation of platform sediments contains a rich record of extinct Antarctic life, including freshwater fish fossils in Devonian age rocks; ancient, temperate forests of trees known as *Glossopteris* in coal deposits of Permian- and Triassic-age rocks (about 190,000,000 to 280,000,000 years ago); and large reptiles, such as *Lystrosaurus*, and amphibians in Triassic-age rocks. Tillites—rocks deposited by ancient glaciers—underlie Permian-age coal beds in many places in Antarctica as they do in other southern, including now tropical, continents. Widespread volcanism of the Kirkpatrick Basalt (Mt. Kirkpatrick) and igneous intrusions of the Ferrar Dolerite terminated Beacon sedimentation in Jurassic time about 160,000,000 years ago.

Beacon Sandstone

Mountain glaciers, which formed on such highlands as the Sentinel Range perhaps as early as about 50,000,000 years ago, advanced down valleys to calve into the sea. Fringing ice shelves were built and later became grounded as glaciation intensified. Local ice caps developed, covering West Antarctic island groups as well as mountain ranges of East Antarctica. The ice caps eventually coalesced into great ice sheets that tied together West and East Antarctica into the single continent that is known today. The continent has apparently been continuously glaciated since the first glaciers appeared.

Development of modern ice sheets

PHYSICAL GEOGRAPHY

Physiography and relief. There are two faces of the present-day continent of Antarctica. One, seen visually, consists of the exposed rock and ice-surface terrain. The other, seen only indirectly by seismic or other remote-sensing technique, consists of the ice-buried bedrock surface. Both evolved through long and slow geological processes. Glacially sculptured landforms predominate today, as they must have some 300,000,000 years ago, in an earlier period of continental glaciation of all of Gondwanaland. Through most of its history, however, Antarctica was affected by more normal processes of erosion and deposition as are other continents today.

General elevation

Antarctica, with an average height estimated at between 7,000 and 8,000 feet, is the world's highest continent. (Asia, the next, averages about 3,000 feet.) Without its ice, however, it would probably average little more than about 1,500 feet. It would then consist of a far smaller continent, East Antarctica, and a nearby island archipelago. A vast lowland plain between 90° E and 150° E (today's Polar and Wilkes subglacial basins) would be fringed by the ranges of the Transantarctic Mountains and of the Gamburtsev Mountains, 6,500 to 13,000 feet high. The rest might be a hilly to mountainous terrain. Relief in general would be great, with elevations ranging from 16,864 feet (5,140 metres) at Vinson Massif in the Sentinel Range, the highest point in Antarctica, to more than 8,200 feet below sea level in an adjoining marine trough to the west (today's Bentley Subglacial Trench). Areas now named lands, as most of Ellsworth Land and Marie Byrd Land, would be beneath the sea.

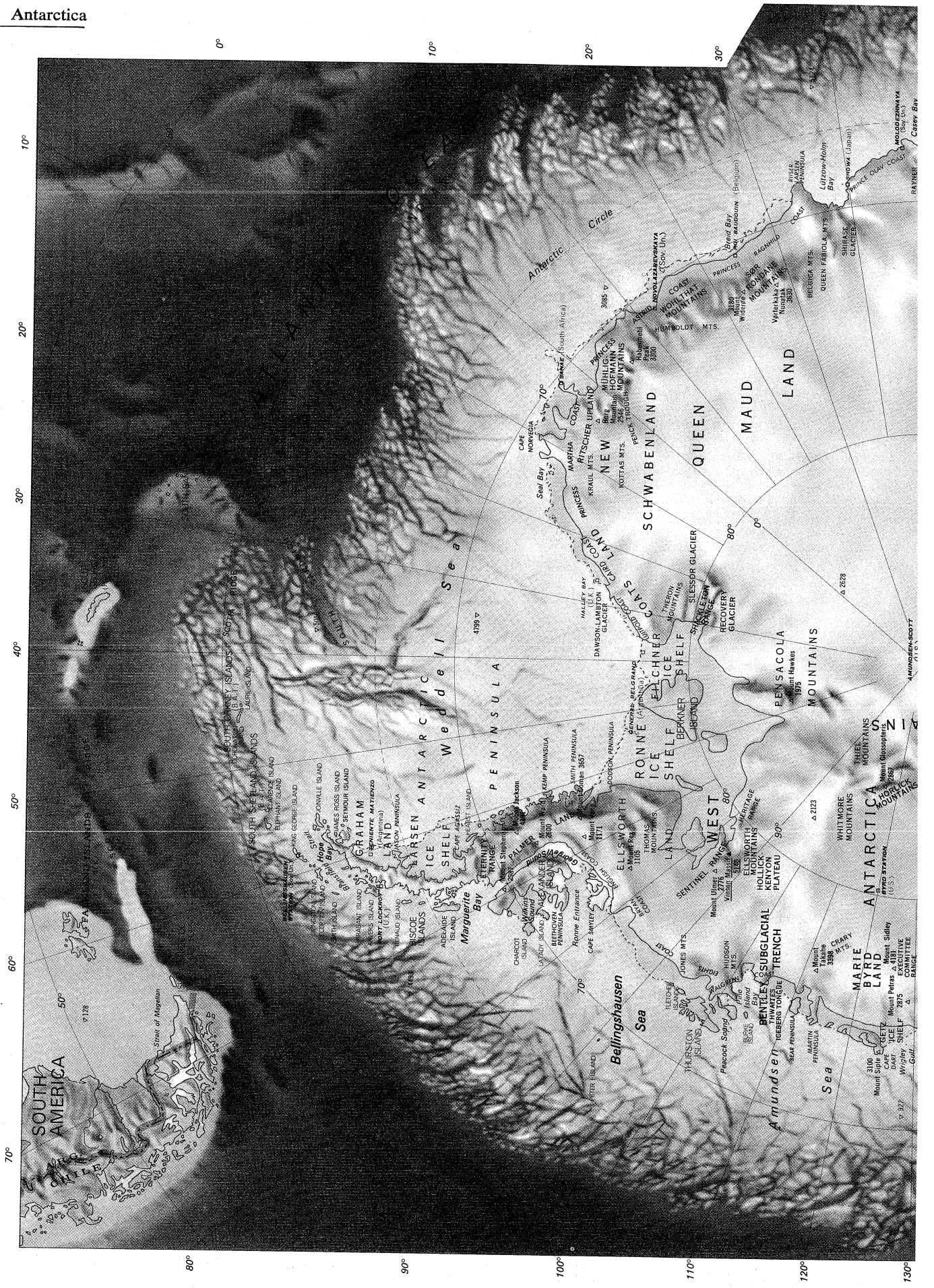


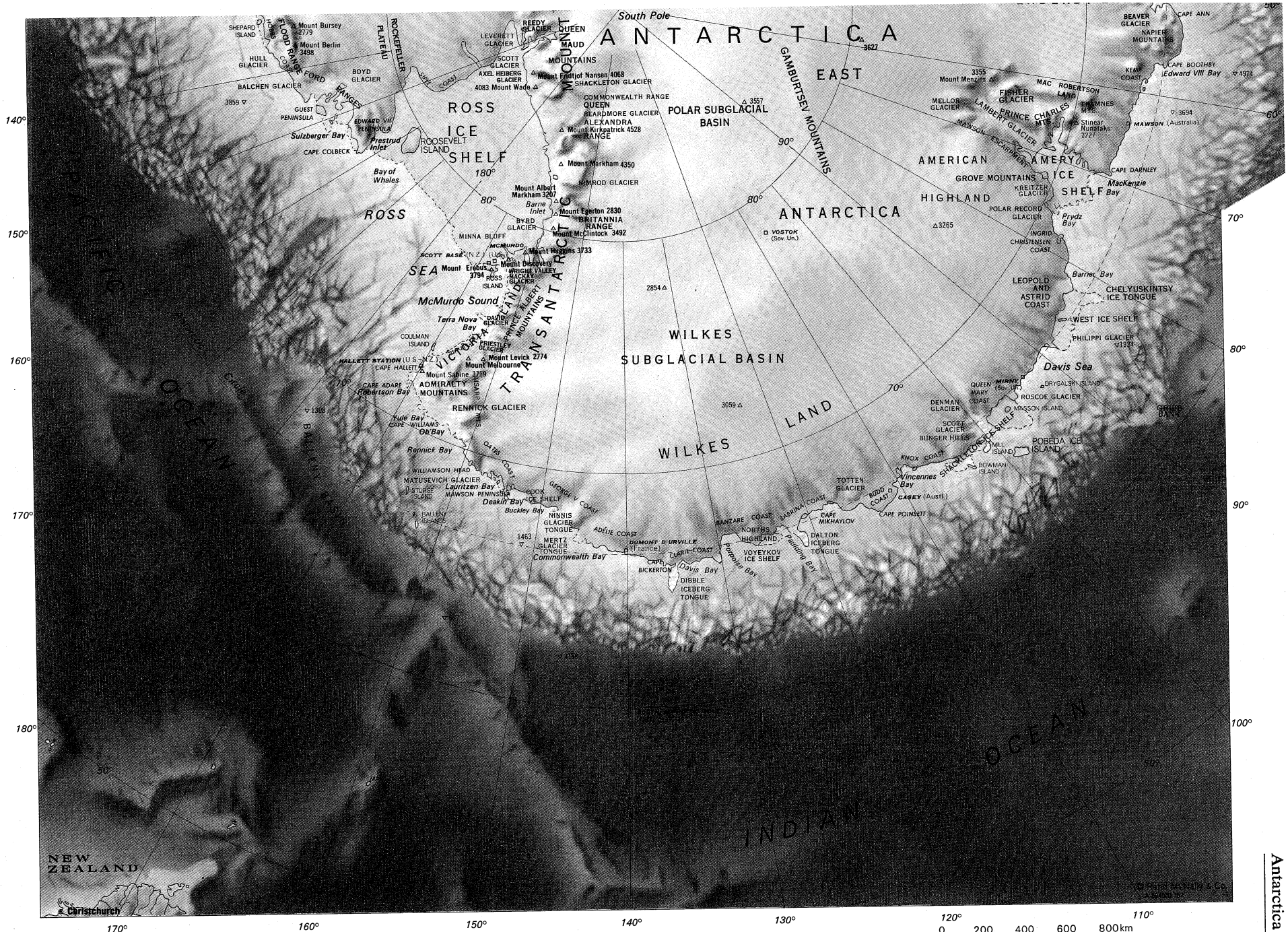
Wright Valley. View west up a McMurdo dry valley to Onyx River, Lake Vanda, and the polar plateau beyond.
By courtesy of the U.S. Navy

Because of the ice, the physiography today little reflects the buried bedrock-surface topography. The vast ice sheets of East Antarctica reach heights of 11,500 feet or more in two main centres, one at about 80° S of western Wilkes Land, the other in a belt stretching southeastward from about 75° S in central Queen Maud Land. Most of this ice drains to the sea around East Antarctic mountains and deglaciated ice-free coastal areas known as oases to Russian investigators. Some, such as in the Beardmore, Nimrod, and Reedy glaciers, pours as great valley outlet streams through breaks in the dam of the Transantarctic Mountains to feed the Ross Ice Shelf. A few outlet streams of polar ice—the Mackay, David, and Priestley glaciers—calve into Ross Sea. General lowering of ice-sheet levels caused some former outlet glaciers to almost completely recede and vanish from their valleys, producing such spectacular "dry valleys" as the Taylor, Wright, and Victoria valleys near McMurdo Sound. Smaller ice centres developed in West Antarctica and form 6,500-feet-high regions in Marie Byrd Land and in a belt between the Ellsworth and Thiel mountains. The latter belt forms a divide for ice moving eastward to Ronne Ice Shelf and westward to Ross Ice Shelf.

Ice-scarred volcanoes dot western Ellsworth Land and Marie Byrd Land. Volcanic landforms in places are interpreted as having formed by hot lavas contacting the glacier ice. Volcanoes, some still steaming, provide spectacular and scenic landforms at many places and include Mt. Sidley and others of the Executive Committee Range in Marie Byrd Land; Mts. Erebus, Terror, and Discovery near McMurdo Sound; and Mt. Melbourne in northern Victoria Land.

Effects of glacial erosion and deposition dominate


SOUTH AMERICA
Strait of Magellan
ARG.
CHILE
ANTARCTICA
Antarctic Circle
Weddell Sea
Bellingshausen Sea
Amundsen Sea
ANTARCTIC PENINSULA
GRAHAM LAND
LARSEN ICE SHELF
Marguerite Bay
BISCOE ISLANDS
ADELAIDE ISLAND
SOUTH SHETLAND ISLANDS
SOUTH ORKNEY ISLANDS
SCOTIA RIDGE
ELEPHANT ISLAND
CLARENCE ISLAND
KING GEORGE ISLAND
NOVOLAZAREVSKAYA (Sov. Un.)
WOHLTHAT MOUNTAINS
HUMBOLDT MTS.
PRINCESS RAGNHILD COAST
MÜHLIG-HOFMANN MOUNTAINS
PENCK TROUGH
RITSCHER UPLAND
MARTHA COAST
KRAUL MTS.
KOTTAS MTS.
NEW SCHWABENLAND
QUEEN MAUD LAND
CAPE NORVEGIA
Seal Bay
COATS LAND
CAIRD COAST
HALLEY BAY (U.K.)
DAWSON-LAMBTON GLACIER
THERON MOUNTAINS
SLESSOR GLACIER
SHACKLETON RANGE
RECOVERY GLACIER
ELCHNER ICE SHELF
BERKNER ISLAND
RONNE ICE SHELF
GENERE BELGRANO (Argentina)
GENERE BELGRANO (Italia)
PENSACOLA MOUNTAINS
Mount Hawkes 1975
THIEL MOUNTAINS
Mount Glossopteris
HORLICK MOUNTAINS
WEST ANTARCTICA
ELLSWORTH LAND
THOMAS MOUNTAINS
ELLSWORTH MOUNTAINS
HOLLICK-KENYON PLATEAU
SENTINEL RANGE
Mount Vinson 5140
HERITAGE RANGE
WHITMORE MOUNTAINS
BYRD STATION (U.S.)
MARIE BYRD LAND
Mount Sidley
EXECUTIVE COMMITTEE RANGE
Mount Petras
Mount Takahe 3398
CRARY MTS.
BENTLEY SUBGLACIAL TRENCH
THWAITES ICEBERG TONGUE
BEAR PENINSULA
MARTIN PENINSULA
Peacock Sound
HUDSON MTS.
JONES MTS.
THURSTON ISLAND
FLETCHER ISLAND
PETER I ISLAND
CHARCOT ISLAND
LATADY ISLAND
BEETHOVEN PENINSULA
Wilkins Sound
ALEXANDER ISLAND
PALMER LAND
ETERNITY RANGE
Mount Stephenson
Mount Jackson
KEMP PENINSULA
SMITH PENINSULA
DODSON PENINSULA
Mount Rea 1105
Mount Lanyon 3657
Mount Wilson 2800
Mount Gardner 2800
Ronne Entrance
CAPE SMYLEY
BRYAN COAST
ENDERBY LAND
HEARST LAND
Mount Ulmer 2776
Mount Martin 3180
Mount Crary 3100
Mount Siple
CAPE DART
GETZ ICE SHELF
Wrigley Gulf
Lützow-Holm Bay
Casey Bay
PRINCE OLAV COAST
MOLODEZHNAYA (Sov. Un.)
SHOWA (Japan)
RAYNER
RIISER-LARSEN PENINSULA
BELGICA MTS.
QUEEN FABIOLA MTS.
SHIRASE GLACIER
Vestfjella Nunatak
Wilderne Nunatak
Habermehl Peak 3300
Breid Bay
Mt. BAUDOUIN (Belgium)
SANAE (South Africa)
SOUTH GEORGIA ISLAND
SOUTH SANDWICH ISLANDS
SOUTH ORKNEY ISLANDS
SOUTH SHETLAND ISLANDS
ELEPHANT ISLAND
CLARENCE ISLAND
KING GEORGE ISLAND
JONVILLE ISLAND
SEYMOUR ISLAND
JAMES ROSS ISLAND
GRAHAM LAND
GENIVIEVE MATHIEZO (Argentina)
LARSEN PENINSULA
Hope Bay
Bransfield Strait
BRABANT ISLAND
ROSS ISLAND
SEYMOUR ISLAND
PORT LOCKROY (U.K.)
EDWARD ISLAND
BISCOE ISLANDS
ADELAIDE ISLAND
Marguerite Bay
ALEXANDER ISLAND
PALMER LAND
ETERNITY RANGE
Mount Stephenson
Mount Jackson
KEMP PENINSULA
SMITH PENINSULA
DODSON PENINSULA
Mount Rea 1105
Mount Lanyon 3657
Mount Wilson 2800
Mount Gardner 2800
Wilkins Sound
BEETHOVEN PENINSULA
LATADY ISLAND
CHARCOT ISLAND
Ronne Entrance
CAPE SMYLEY
BRYAN COAST
ENDERBY LAND
HEARST LAND
ELLSWORTH LAND
THOMAS MOUNTAINS
ELLSWORTH MOUNTAINS
HOLLICK-KENYON PLATEAU
SENTINEL RANGE
Mount Vinson 5140
Mount Ulmer 2776
Mount Martin 3180
HERITAGE RANGE
WHITMORE MOUNTAINS
BYRD STATION (U.S.)
MARIE BYRD LAND
Mount Sidley
EXECUTIVE COMMITTEE RANGE
Mount Petras
Mount Takahe 3398
CRARY MTS.
BENTLEY SUBGLACIAL TRENCH
THWAITES ICEBERG TONGUE
BEAR PENINSULA
MARTIN PENINSULA
Peacock Sound
HUDSON MTS.
JONES MTS.
THURSTON ISLAND
FLETCHER ISLAND
PETER I ISLAND
Bellingshausen Sea
Amundsen Sea
CAPE DART
GETZ ICE SHELF
Wrigley Gulf
Mount Siple
Lützow-Holm Bay
Casey Bay
PRINCE OLAV COAST
MOLODEZHNAYA (Sov. Un.)
SHOWA (Japan)
RAYNER
RIISER-LARSEN PENINSULA
BELGICA MTS.
QUEEN FABIOLA MTS.
SHIRASE GLACIER
Breid Bay
Mt. BAUDOUIN (Belgium)
SANAE (South Africa)
Vestfjella Nunatak
Wilderne Nunatak
Habermehl Peak 3300
NOVOLAZAREVSKAYA (Sov. Un.)
WOHLTHAT MOUNTAINS
HUMBOLDT MTS.
PRINCESS RAGNHILD COAST
MÜHLIG-HOFMANN MOUNTAINS
PENCK TROUGH
RITSCHER UPLAND
MARTHA COAST
KRAUL MTS.
KOTTAS MTS.
NEW SCHWABENLAND
QUEEN MAUD LAND
CAPE NORVEGIA
Seal Bay
COATS LAND
CAIRD COAST
HALLEY BAY (U.K.)
DAWSON-LAMBTON GLACIER
THERON MOUNTAINS
SLESSOR GLACIER
SHACKLETON RANGE
RECOVERY GLACIER
ELCHNER ICE SHELF
BERKNER ISLAND
RONNE ICE SHELF
GENERE BELGRANO (Argentina)
GENERE BELGRANO (Italia)
PENSACOLA MOUNTAINS
Mount Hawkes 1975
THIEL MOUNTAINS
Mount Glossopteris
HORLICK MOUNTAINS
WEST ANTARCTICA
Antarctic Circle
0°
10°
20°
30°
40°
50°
60°
70°
80°
90°
100°
110°
120°
130°



ANTARCTICA Elevations and depths in metres

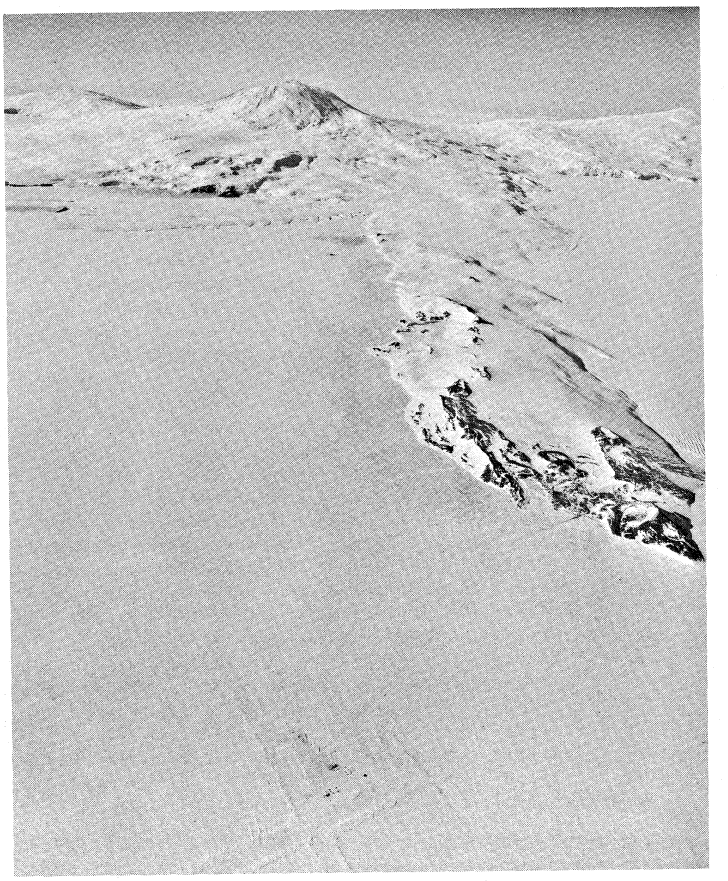
MAP INDEX

Physical features and points of interest

Adare, Cape....71·17s 170·14e
Adelaide Island.67·15s 68·30w
Adélie Coast....67·00s 139·00e
Admiralty Mountains.....71·45s 168·30e
Agassiz, Cape..68·29s 62·56w
Albert Markham, Mount, *mountain*.....81·23s 158·12e
Alexander Island.....71·00s 70·00w
American Highland.....72·30s 78·00e
Amery Ice Shelf.....69·30s 72·00e
Amundsen Bay.66·55s 50·00e
Amundsen-Scott, *scientific station*.....90·00s 0·00
Amundsen Sea.72·30s 112·00w
Ann, Cape.....66·10s 51·22e
Antarctic Peninsula.....69·30s 65·00w
Anvers Island..64·33s 63·35w
Atlantic Ocean..0·00 25·00w
Axel Heiberg Glacier.....85·25s 163·00w
Balchen Glacier.....76·46s 145·30w
Balleny Islands..66·35s 162·50e
Banzare Coast..67·00s 126·00e
Barne Inlet.....80·15s 160·15e
Barrier Bay.....67·45s 81·10e
Beardmore Glacier.....83·45s 171·00e
Bear Peninsula.74,30s 110·45w
Beaver Glacier..83·24s 169·30e
Beethoven Peninsula.....71·40s 73·45w
Belgica Mountains.....72·35s 31·10e
Bellingshausen Sea.....71·00s 85·00w
Bellingshausen Station, *scientific station*.....62·12s 58·58w
Bentley Subglacial Trench, *subsurface feature*.....78·00s 110·00w
Berkner Island..79·30s 49·30w
Berlin, Mount, *mountain*.....76·03s 135·52w
Bickerton, Cape.....66·20s 136·56e
Biscoe Islands..66·00s 66·30w
Boothby, Cape..66·34s 57·16e
Borg Mountain.....72·42s 3·30w
Bowman Island.65·17s 103·08e
Boyd Glacier....77·14s 145·25w
Brabant Island..64·15s 62·20w
Bransfield Strait.....63·00s 59·00w
Breid Bay.....70·15s 24·15e
Britannia Range, *mountain range*.....80·00s 158·00e
Bryan Coast....73·45s 82·00w
Buckley Bay....68·16s 148·12e
Budd Coast.....66·30s 113·00e
Bunger Hills....66·17s 100·47e
Burke Island....73·15s 104·35w
Bursey, Mount, *mountain*.....76·00s 132·40w
Byrd, *scientific station*.....80·00s 120·00w
Byrd Glacier....80·15s 160·20e
Caird Coast.....76·00s 24·30e
Casey, *scientific station*.....66·15s 110·35e
Casey Bay.....67·20s 48·00e
Charcot Island..69·45s 75·15w
Chelyuskintsy Ice Tongue....66·20s 82·00e
Clarie Coast....66·30s 133·00e
Coats Land, *physical region*.77·00s 28·00w
Colbeck, Cape..77·06s 157·48w
Coman, Mount, *mountain*.....74·02s 65·04w
Commonwealth Bay.....66·54s 142·40e
Commonwealth Range, *mountain range*.....84·15s 172·20e
Cook Ice Shelf..68·40s 152·30e
Coulman Island.73·27s 169·40e
Coronation Island.....60·37s 45·35w
Crary Mountains.....76·48s 117·40w
Dalton Iceberg Tongue.....66·15s 121·30e
Darnley, Cape..67·43s 69·30e
Dart, Cape.....73·06s 126·20w
David Glacier...75·19s 162·00e
Davis Bay.....66·08s 134·05e
Davis Sea.....66·00s 92·00e
Dawson-Lambton Glacier.....76·15s 27·30w
Deakin Bay.....68·23s 150·10e
Deception Island.....62·57s 60·38w
Denman Glacier.....66·45s 99·25e
Dibble Iceberg Tongue.....65·40s 135·10e
Discovery, Mount, *mountain*.....78·23s 165·00e
Dodson Peninsula.....75·46s 62·50w
Drake Passage..58·00s 70·00w
Drygalski Island.65·45s 92·30e
Dumont d'Urville, *scientific station*.....66·35s 140·00e
East Antarctica, *physical region*.80·00s 80·00e
Edward VII Peninsula.....77·40s 155·00w
Edward VIII Bay.66·50s 57·00e
Egerton, Mount, *mountain*.....80·50s 158·50e
Eights Coast....73·30s 93·00w
Elephant Island.61·10s 55·14w
Ellsworth Land, *physical region*.75·30s 80·00w
Ellsworth Mountains.....79·00s 85·00w
Enderby Land, *physical region*.67·30s 53·00e
English Coast...73·45s 73·00w
Erebus, Mount, *mountain*.....77·32s 167·09e
Eternity Range, *mountain range*.....69·46s 64·34w
Executive Committee Range, *mountain range*.....76·50s 126·00w
Filchner Ice Shelf.....79·00s 40·00w
Fisher Glacier...73·15s 66·00e
Fletcher Islands.72·40s 94·10w
Ford Ranges, *mountains*.....77·00s 145·00w
Framnes Mountains.....67·50s 62·35e
Fridtjof Nansen, Mount, *mountain*.....85·21s 167·33w
Gamburtsev Mountains, *subsurface feature*.....82·00s 78·00e
General Belgrano, *scientific station*.....77·45s 38·15e
George V Coast.68·30s 147·30e
George VI Sound.....71·00s 68·00w
Getz Ice Shelf...75·00s 129·00w
Glossopteris, Mount, *mountain*.....84·44s 113·51w
Graham Land, *physical region*.66·00s 63·30w
Grove Mountains.....72·53s 74·53e
Guest Peninsula.....76·18s 148·00w
Habermehl Peak.....71·49s 6·38e
Hallett, Cape....72·19s 170·18e
Hallett Station, *scientific station*.....72·18s 170·18e
Halley Bay.....75·27s 26·42w
Hawkes, Mount, *mountain*.....83·56s 55·45w
Hearst Island...69·25s 62·10w
Heritage Range, *mountain range*.....79·30s 84·00w
Hobbs Coast....74·45s 131·00w
Hollick Kenyon Plateau.....79·00s 97·00w
Hope Bay.....63·23s 57·00w
Horlick Mountains.....85·23s 121·00w
Hudson Mountains.....74·32s 99·20w
Huggins, Mount, *mountain*.....78·17s 162·28e
Hull Glacier.....75·05s 137·15w
Humboldt Mountains.....71·45s 11·30e
Indian Ocean...10·00s 70·00e
Ingrid Christensen Coast.....69·30s 76·00e
Jackson Mount, *mountain*.....71·23s 63·22w
James Ross Island.....64·15s 57·45w
Jason Peninsula.....66·10s 61·00w
Joinville Island..63·15s 55·45w
Jones Mountains.....73·32s 94·00w
Kemp Coast....67·10s 58·00e
Kemp Peninsula.....73·08s 60·15w
King George Island.....62·00s 58·15w
Kirkpatrick, Mount, *mountain*.....84·20s 166·19e
Knox Coast.....66·30s 105·00e
Kottas Mountains.....74·20s 12·00w
Kraul Mountains.....73·10s 14·10w
Kreitzer Glacier.70·25s 72·30e
Lambert Glacier.71·00s 70·00e
Larsen Ice Shelf.68·30s 62·30w
Latady Island...70·45s 74·35w
Laurie Island....60·44s 44·37w
Lauritzen Bay...69·05s 156·50e
Leopold And Astrid Coast...67·10s 84·10e
Leverett Glacier.85,30s 150·00w
Levick, Mount, *mountain*.....74·08s 163·12e
Livingston Island.....62·36s 60·30w
Luitpold Coast..78·30s 32·00w
Lützow-Holm Bay.....69·10s 37·30e
McClintock, Mount, *mountain*.....80·13s 157·26e
Mackay Glacier.76·54s 162·00e
MacKenzie Bay.68·20s 71·15e
McMurdo, *scientific station*.....77·50s 166·25e
McMurdo Sound.....77·30s 165·00e
Mac. Robertson Land, *coast*....68·10s 65·00e
Marguerite Bay.68·30s 68·30w
Marie Byrd Land, *physical region*.....80·00s 120·00w
Markham, Mount, *mountain*.....82·51s 161·21e
Martin Peninsula.....74·25s 114·10w
Masson Island..66·08s 96·34e
Matusevich Glacier.....69·11s 157·20e
Mawson Escarpment....73·05s 68·10e
Mawson Peninsula.....68·35s 154·11e
Melbourne, Mount, *mountain*.....74·21s 162·47e
Mellor Glacier...73·30s 66·30e
Menzies, Mount, *mountain*.....73·30s 61·50e
Mertz Glacier Tongue, *glacier*.....67·40s 144·45e
Mikhaylov, Cape.....66·51s 118·33e
Mill Island.....65·30s 100·40e
Minna Bluff, *peninsula*.....78·32s 166·30e
Mirny, *scientific station*.....66·33s 93·00e
Molodezhnaya, *scientific station*.....67·35s 46·35e
Mühlig-Hofmann Mountains.....72·00s 5·20e
Napier Mountains.....66·30s 53·40e
New Schwabenland, *physical region*.72·30s 1·00e
Nimrod Glacier..82·27s 161·00e
Ninnis Glacier Tongue, *glacier*.....68·12s 147·12e
Norths Highland.....66·40s 126·00e
Norvegia, Cape.71·25s 12·18w
Novolazarevskaya, *scientific station*.....70·45s 11·50e
Oates Coast....70·00s 160·00e
Ob' Bay.....70·35s 163·22e
Pacific Ocean...10·00s 150·00w
Palmer Land, *physical region*.71·30s 65·00w
Paulding Bay...66·35s 123·00e
Peacock Sound.72·55s 100·00w
Penck Trough, *valley*.....73·00s 2·45w
Pensacola Mountains.....83·45s 55·00w
Peter I Island...68·47s 90·35w
Petras, Mount, *mountain*.....75·52s 128·38w
Philippi Glacier.66·45s 88·20e
Pine Island Bay.74·50s 102·05w
Pobeda Ice Island.....64·30s 97·00e
Poinsett, Cape..65·42s 113·18e
Polar Record Glacier.....69·49s 75·30e
Polar Subglacial Basin *subsurface feature*.....85·00s 110·00e
Porpoise Bay...66·30s 128·30e
Port Lockroy, *scientific station*.....64·48s 63·30w
Prestrud Inlet...78·18s 156·00w
Priestley Glacier.....74·20s 163·22e
Prince Albert Mountains.....76·00s 161·30e
Prince Charles Mountains.....72·00s 67·00e
Prince Olav Coast.....68·30s 42·30e
Princess Astrid Coast.....70·45s 12·30e
Princess Martha Coast.....72·00s 7·30w
Princess Ragnhild Coast.....70·15s 27·30e
Prydz Bay.....69·00s 76·00e
Queen Alexandra Range, *mountain range*.....84·00s 168·00e
Queen Fabiola Mountains.....71·30s 35·40e
Queen Mary Coast.....67·00s 96·00e
Queen Maud Land, *physical region*.....72·30s 12·00e
Queen Maud Mountains.....86·00s 160·00w
Rayner Glacier..67·40s 48·30e
Recovery Glacier.....81·10s 28·00w
Reedy Glacier...85·30s 134·00w
Renaud Island..65·40s 66·00w
Rennick Bay....70·18s 161·45e
Rennick Glacier.70·30s 161·45e
Rex, Mount, *mountain*.....74·57s 76·00w
Riiser-Larsen Peninsula.....68·55s 34·00e
Ritscher Upland.....73·20s 9·30w
Robertson Bay..71·25s 170·00e
Rockefeller Plateau.....80·00s 135·00w
Roi Baudouin, *scientific station*.....70·25s 24·20e
Ronne Entrance, *inlet*.....72·30s 74·00w
Ronne Ice Shelf.78·30s 61·00w
Roosevelt Island.....79·30s 162·00w
Roscoe Glacier..66·30s 95·20e
Ross Ice Shelf..81·30s 175·00w
Ross Island.....77·30s 168·00e
Ross Sea.....76·00s 175·00w
Sabine, Mount, *mountain*.....71·55s 169·33e
Sabrina Coast...67·00s 119·30e
Sanae, *scientific station*.....70·30s 2·30w
Scotia Ridge, *undersea ridge*.57·00s 27·00w
Scotia Sea.....56·00s 40·00w
Scott Base, *scientific station*.....77·50s 166·25e
Scott Glacier....66·15s 100·05e
Scott Glacier....85·45s 153·00w
Seal Bay.....71·40s 12·25w
Sentinel Range, *mountain range*.....78·10s 85·30w

MAP INDEX (continued)

Seymour Island.....64·17s 56·45w
Shackleton Glacier.....84·35s 176·15w
Shackleton Ice Shelf.....66·00s 100·00e
Shackleton Range, *mountain range*.....80·40s 26·00w
Shepard Island..74·25s 132·30w
Shirase Glacier.70·10s 38·35e
Showa, *scientific station*.....69·00s 39·35e
Sidley, Mount, *mountain*.....77·02s 126·00w
Siple, Mount, *mountain*.....73·15s 126·06w
Siple Coast.....82·00s 153·00w
Slessor Glacier..79·50s 28·30w
Smith Island....63·00s 62·30w
Smith Peninsula.....74·25s 61·15w
Smyley, Cape...72·26s 78·10w
Sør Rondane Mountains.....72·00s 25·00e
South Magnetic Pole.....66·40s 140·10e
South Orkney Islands.....60·35s 45·30w
South Pole.....90·00s 0·00
South Shetland Islands.....62·00s 60·00w
Stephenson, Mount, *mountain*.....69·49s 69·43w
Stinear Nunataks, *peaks*.....69·42s 64·40e
Sturge Island...67·27s 164·18e
Sulzberger Bay.77·00s 152·00w
Takahe, Mount, *mountain*.....76·16s 112·14w
Teniente Matienzo, *scientific station*.....64·58s 60·02w
Terra Nova Bay.74·45s 164·30e
Theron Mountains.....79·05s 28·15w
Thiel Mountains.....85·15s 91·00w
Thomas Mountains.....75·32s 70·57w
Thurston Island.72·20s 99·00w
Thwaites Iceberg Tongue, *ice tongue*.....74·45s 106·30w
Totten Glacier..66·45s 116·10e
Transantarctic Mountains.....85·00s 175·00w
Ulmer, Mount, *mountain*.....77·35s 86·09w
Usarp Mountains.....71·10s 160·00e
Vang, Mount, *mountain*.....73·56s 68·39w
Victoria Land, *physical region*.75·00s 163·00e
Vincennes Bay..66·30s 109·30e
Vinson Massif, *mountain*.....78·35s 85·25w
Vørterkaka Nunatak, *mountain*.....72·20s 27·29e
Vostok, *scientific station*.....78·30s 106·50e
Voyeykov Ice Shelf.....66·20s 124·38e
Wade, Mount, *mountain*.....84·51s 174·15w
Walgreen Coast.75·15s 105·00w
Ward, Mount, *mountain*.....71·55s 66·00w
Weddell Sea....72·00s 45·00w
West Antarctica, *physical region*.79·00s 100·00w
West Ice Shelf..67·00s 85·00e
Whales, Bay of..78·30s 164·20w
White Island....66·44s 48·35e
Whitmore Mountains....82·35es 104·30w
Widerøe, Mount, *mountain*.....72·08s 23·30e
Wilkes Land, *physical region*.69·00s 120·00e
Wilkes Subglacial Basin, *subsurface feature*.....75·00s 130·00e
Wilkins Sound..70·15s 73·00w
Williams, Cape..70·29s 164·05e
Williamson Head, *headland*.....69·09s 157·49e
Wohlthat Mountains.....71·35s 12·20e
Wright Valley...77·33s 161·30e
Wrigley Gulf....74·00s 129·00w
Yule Bay.....70·44s 166·40e



Ross Island seen across frozen McMurdo Sound. McMurdo Station is on the tip of the island at right with its aircraft runway in the foreground. The volcanic Mt. Erebus is in the upper centre.
By courtesy of the U.S. Navy

everywhere in Antarctica, and erosional effects of running water are relatively minor. Yet, on warm summer days, rare and short-lived streams of glacial melt water do locally exist. The evanescent Onyx River, for example, flows from Lower Wright Glacier terminus to empty into the nondrained basin of Lake Vanda near McMurdo Sound. The origin of Lake Vanda and other highly saline lakes of the McMurdo Sound dry valleys is controversial. High salt concentrations in the lakes has been explained as resulting from (1) volcanic or hot-spring activity, (2) evaporite leaching from local bedrock, (3) wind blowing of marine salts from the nearby sea, (4) evaporational concentration from a trapped marine basin, or (5) chemical weathering of bedrock and soil. Recent studies tend to support the last hypothesis. Water is doubtless an important agent of weathering, or rock disintegration, by frost wedging in many ice-free areas. Even at 85° S, where free-air temperatures never rise above 32° F (0° C), temperatures on sun-heated rock surfaces have been measured at more than 59° F (15° C). Wind-drift snow quickly melts on such surfaces, and the water trickles into nearby cracks where, out of the sun, it freezes, expands, and breaks the rock further. Surfaces of angular rock fragments on many mountains may thereby have been produced. The importance of weathering and soil formation, long believed to be processes incapable of operating in the cold and dry Antarctic environment, is slowly being recognized. The Antarctic studies, indeed, may have important bearing in interpreting future work in terrains believed to be similar on Mars.

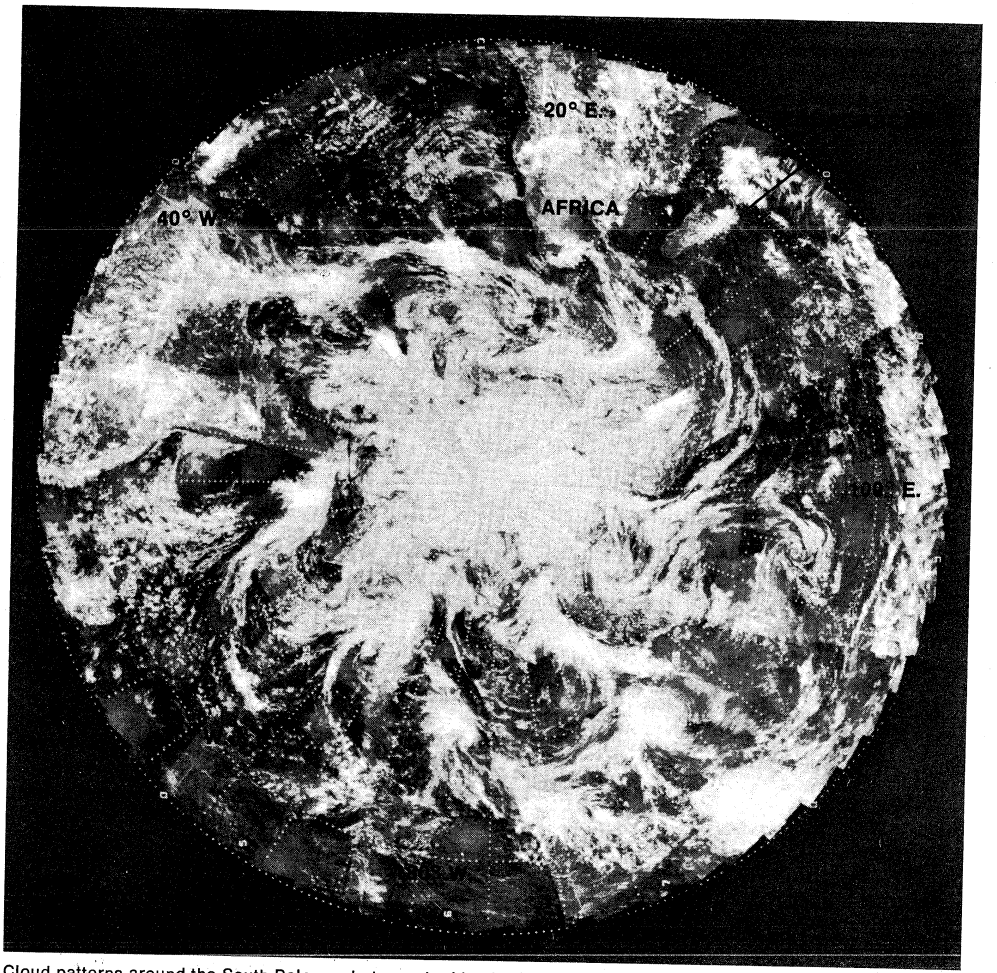
Causes of salinity in lakes

Climate and meteorology. The unique weather and climate of Antarctica provide the basis for its familiar appellations—Home of the Blizzard and White Desert. These are well earned. By far the coldest continent, Antarctica has winter temperatures that range from −126.9° F (−88.3° C), the world's lowest recorded temerature, measured at Vostok Station on August 24, 1960, on the high inland ice sheet, to −76° F (−60° C) near sea level. Temperatures vary greatly from place to place, but direct measurements in most places are generally available only for summertime. Only at fixed stations operated since the IGY have year-round measurements been made. Winter temperatures rarely reach as high as 52° F (11° C) on the northern Antarctic Peninsula, the warmest part of the continent, during winds of maritime air and under a cloud cover that restricts radiation loss. Mean temperatures of the coldest months are −4° to −22° F (−20° to −30° C) on the coast, and −40° to −94° F (−40° to −70° C) in the interior, the coldest period on the polar plateau being usually in late August just before return of the sun. Whereas midsummer temperatures may reach as high as 59° F (15° C) on the Antarctic Peninsula, those elsewhere are usually much lower, ranging from a mean of about 32° F (0° C) on the coast to between −4° and −31° F (−20° and −35° C) in the interior. These temperatures are far lower than those of the Arctic, where monthly means range only from about 32° F (0° C) in summer to −31° F (−35° C) in winter.

Temperatures

Wind chill, the cooling power of the atmosphere, is the major debilitating weather factor of Antarctic expeditions. Fierce winds characterize most coastal regions, particularly of East Antarctica, where cold, dense air flows down the steep slopes off interior highlands. Known as katabatic winds, they are a surface flow that may be smooth if of low velocity, but that may become greatly turbulent, sweeping high any loose snow, if a critical velocity is surpassed. This turbulent air may appear suddenly and is responsible for the brief and localized Antarctic "blizzards" during which no snow actually falls

Winds



Cloud patterns around the South Pole as photographed by the U.S. satellite, ESSA 3, January 28, 1967. At the top is the southern end of Africa and Madagascar Island.
By courtesy of the Commerce Department, National Oceanic and Atmospheric Administration, National Environmental Satellite Service, Washington, D.C.

and skies above are clear. During one winter at Mirny Station, gusts reached more than 110 miles per hour on seven occasions. At the Adélie Coast camp of Douglas Mawson's Australasian Antarctic Expedition (1911–14), velocities in the winter of 1912–13 averaged 40 miles per hour for 64 percent of the time. Gusts estimated at between 140 and 155 miles per hour on December 9, 1960, destroyed a Beaver aircraft at Mawson Station on the MacRobertson Land coast. Not all these winds are truly katabatic because many have components produced by factors other than slope, such as the passage of cyclonic disturbances.

Winds on the polar plateau are usually light, with monthly mean velocities at the South Pole ranging from about nine miles per hour in December (summer) to 17 miles per hour in June and July (winter). Winds on the polar plateau are of complex origin and are generated not only by downslope drainage but more commonly by other factors, such as the forces initiated by the Earth's rotation.

The Antarctic atmosphere, because of its low temperature, contains only about one-tenth the water vapour concentration found in temperate latitudes. This atmospheric water largely comes from ice-free regions of the southern oceans and is transported in the troposphere into Antarctica mostly in the 140° sector (80° E to 140° W) from Wilkes Land to Marie Byrd Land. Most of this water precipitates as snow along the continental margin. Rainfalls are almost unknown. Despite the tremendous volume of potential water stored as ice, Antarctica must be considered as one of the world's great deserts; the average precipitation (water equivalent) is only about two inches per year over the polar plateau, though considerably more, perhaps ten times as much, falls in the coastal belt. Lacking a heavy and protective water-vapour-rich atmospheric layer, which in other areas absorbs and reradiates to earth long-wave radiation, the Antarctic surface readily loses heat energy into space.

Solar and terrestrial radiation

Many factors determine Antarctica's climate, but the primary one is the geometry of the Sun–Earth relationship. The 23.5° axial tilt of the Earth to its annual plane of orbit, or ecliptic, about the Sun results in long winter nights and long summer days, alternating between both polar regions and causing seasonal variations in climate. On midwinter day, about June 21, the sun's rays reach to only 23.5° (not exact, due to refraction) from the South Pole along the latitude of 66.5° S, a line familiarly known as the Antarctic Circle. Although "night" theoretically is 6 months long at the geographic pole, one month of this actually is a twilight period. Only a few coastal fringes lie north of the Antarctic Circle. The amount of incoming solar radiation and thus heat depends additionally on the incident angle of the rays and therefore decreases inversely with latitude to reach a minimum at the geographic poles. These and other factors are essentially the same for both polar regions. The reason for their great climatic difference primarily lies in their reverse distributions of land and sea: the Arctic, an ocean surrounded by land; the Antarctic, a continent surrounded by ocean. The Arctic Ocean, a climate-ameliorating heat source, has no counterpart at the South Pole, the great elevation and perpetually reflective snow cover of which instead intensify its polar climate. Moreover, during Antarctic winters, freezing of the surrounding sea effectively more than doubles the size of the continent and removes the oceanic heat source to nearly 1,800 miles from the central polar plateau.

Outgoing terrestrial radiation greatly exceeds absorbed

incoming solar radiation. This loss results in strong surface cooling, giving rise to the characteristic Antarctic temperature inversions in which temperature increases from the surface upward to about 1,000 feet. About 90 percent of the loss is replaced by atmospheric heat from lower latitudes, and the remainder by latent heat of water-vapour condensation.

Great cyclonic storms circle Antarctica in endless west to east procession, exchanging atmospheric heat to the continent from sources in the southern Atlantic, Pacific, and Indian oceans. Moist maritime air interacting with cold polar air makes the Antarctic Ocean in the vicinity of the Polar Front one of the world's stormiest. Few storms bring snowfalls to interior regions. With few reporting stations, weather prediction has been exceedingly difficult but is now greatly aided by satellite photography.

Many current studies on the upper atmosphere of Antarctica are leading toward solution of problems worldwide in scope. Very-low-frequency radio waves, called "whistlers," which are generated by lightning at opposite, or conjugate, Northern Hemisphere points and received in Antarctica, are providing much information about the Earth's magnetosphere. Antarctica has made many contributions as well to knowledge of the ionosphere, geomagnetism, and solar cosmic radiation.

Glaciology and oceanography. Antarctica today provides the best available picture of the probable appearance 20,000 years ago of northern North America under the great Laurentide Ice Sheet. Causal factors leading to birth and development of these continental ice sheets and then to their decay and death are, nevertheless, still poorly understood. The factors are complexly interrelated. Moreover, once developed, ice sheets tend to form independent climatic patterns and thus to be self-perpetuating and eventually perhaps even self-destructing. Cold-air masses draining off Antarctic lands, for example, cool and freeze surrounding oceans in winter to form an ice pack, which reduces solar-energy input by increasing reflectivity and makes interior continental regions even more remote from open oceanic heat and moisture sources. The East Antarctic Ice Sheet has grown to such great elevation and extent that little atmospheric moisture now nourishes its central part.

Evolution of ice sheets

Hypotheses for continental glaciation fall into five main groups: (1) the topographic thesis, which maintains that, because temperature decreases with altitude, mountain uplift causes glaciation; (2) the land-displacement thesis, which asserts that polar climatic zones have always existed and that continents have drifted into and out of them; (3) the atmospheric-composition thesis, which maintains, first, that carbon dioxide and water vapour provide a "greenhouse" effect by selectively absorbing long-wave radiation and if decreased will cause cooling and, second, that atmospheric dust reflects or absorbs solar radiation, thereby cooling the Earth, and acts as nucleii for snow precipitation; (4) the planetary thesis, which affirms that perturbations in the Earth's orbital motions cause changes in the amount of received solar radiation; and (5) the cosmic thesis, which makes reference to movements of the solar system through cold regions of space or to variation in solar-energy output. Strong arguments can be marshalled against any individual hypothesis, and the ultimate cause doubtless lies in the combined effect of several factors.

A recent hypothesis ties periodic Northern Hemisphere "ice ages," or glaciations, to periodic instability and surging of the Antarctic Ice Sheet. Instability is believed to occur when glacier buildup becomes so great as to cause pressure sufficient to melt the base of the ice sheet. Then, riding on water, the East Antarctic ice dome rapidly collapses and the outflowing ice forms extensive ice shelves reaching perhaps to the Antarctic Convergence, according to the hypothesis. The resulting increased reflectivity over such a large area would then cool the entire earth. Once collapsed, the ice-dome base would no longer be at the pressure melting point, and the ice shelves, no longer fed, would soon disappear by calving and melting. Consequent decreased reflectivity over the Antarctic Ocean would then result in global warming and thus an interglacial period. The process could repeat itself many times.

Antarctic ice budget

Today, the Antarctic Ice Sheet would seem to be approximately in a state of equilibrium, neither increasing nor decreasing significantly according to the best estimates. Snow precipitation is offset mainly by continental ice moving seaward—by three mechanisms, ice-shelf flow, ice-stream flow, and sheet flow. The greatest volume loss is by calving from shelves, particularly the Ross, Ronne, Filchner, and Amery ice shelves. Much loss may also occur by bottom melting from ice shelves, but its importance is controversial and some glaciologists believe ice shelves actually gain mass by bottom accretion (see ROSS ICE SHELF).

The Antarctic ice sheets are, in effect, settling tanks for atmospheric constituents. They contain a record of climates and ancient atmospheres, of global volcanic activity, and of cosmic-particle bombardment extending back for thousands of years. Upper layers of ice, for instance, record the effects of high-latitude pollution by industrial wastes and of fallout of radioactive strontium (strontium-90) from thermonuclear tests. This record in ice is now being studied in cores from deep holes drilled through the ice sheet.

The seas around Antarctica have often been likened to the moat around a fortress. The turbulent "Roaring Forties" and "Furious Fifties" lie in a circumpolar storm track and a westerly oceanic current zone known commonly as the West Wind Drift, or Circumpolar Current. The three major oceans, Atlantic, Pacific, and Indian, are bounded on the south by Antarctic shores. Warm, subtropical surface currents move southward in the western parts of these oceans and then turn eastward upon meeting the Circumpolar Current. The warm water meets and partly mixes with cold Antarctic water, known as Antarctic Surface Water, to form a mass with intermediate characteristics called Subantarctic Surface Water. Mixing occurs in a shallow but broad zone of approximately 10° latitude lying south of the Subtropical Convergence (at about 40° S) and north of the Antarctic Convergence (between about 50° and 60° S). The Subtropical Convergence generally defines the northern limits of a water mass having so many unique physical and biologic characteristics that it is commonly given a separate name, the Antarctic, or, sometimes, the Southern Ocean.

Ocean currents and convergences

The two convergences are well defined and important oceanic boundary zones that profoundly affect climates, marine life, bottom sedimentation, and ice-pack and iceberg drift. They are easily identified by rapid changes in temperature and salinity. Antarctic waters are less saline than tropical waters because of their lower temperatures and lesser evaporational concentration of dissolved salts. In southward crossing of the Subtropical Convergence zone and thus into the sub-Antarctic climatic belt, surface-water temperatures drop by as much as about 9° to 16° F (5° to 9° C). In crossing the Antarctic Convergence, from the sub-Antarctic into the Antarctic climatic zone, surface-water temperature drops further from about 14° to 7° F (8° to 4° C) in summer, or from about 5° to 2° F (3° to 1° C) in winter.

Whereas the pattern of surface currents, controlled largely by the Earth's rotation, winds, water-density differences, and the geometry of basins is relatively well understood; that of deeper water masses is more complex and less well-known. North-flowing Antarctic Surface Water sinks to about 3,000 feet beneath warmer Subantarctic Surface Water along the Antarctic Convergence to become the Subantarctic Intermediate Water. This water mass, as well as the cold Antarctic Bottom Water, spreads far north beyond the Equator to exchange with waters of the Northern Hemisphere. The Antarctic Bottom Current is identifiable in the Atlantic as far north as the Bermuda Rise. Currents near the continent result in a circumferential belt of surface-water divergence accompanied by upwelling of deeper water masses.

Two forms of floating ice masses build out around the continent: (1) glacier-fed semipermanent ice shelves, some of enormous size, such as the Ross Ice Shelf, and



U.S. Navy icebreakers cutting a path through floating sea ice near Little America.
By courtesy of the U.S. Navy

(2) an annually frozen and melted ice pack that in winter reaches to about 56° S in the Atlantic and 64° S in the Pacific. Antarctica has been called the pulsating continent because of the annual buildup and retreat of its secondary ice-fronted coastline. Pushed by winds and currents, the ice pack is in continual motion. This movement is westward in the coastal belt of the East Wind Drift at the continent edge, and eastward (further north) at the belt of West Wind Drift. Icebergs—calved fragments of glaciers and ice shelves—reach a northern limit at about the Subtropical Convergence. With an annual areal variation about six times as great as that for the Arctic ice pack, the Antarctic pack doubtless plays a far greater role in varying heat exchange between ocean and atmosphere and thus probably in altering global weather patterns. Long-term synoptic studies, now aided by satellite photography, are presently searching for long-period changes in the Antarctic ice-pack regimen and the possible relation of these changes to global climate changes. Drilling of sediments on the floors of the Antarctic continental shelves and adjoining oceanic regions by the Deep Sea Drilling Project ship, the "Glomar Challenger," over a period of several years beginning in 1972, will provide evidence for extending the known history of such changes back many millions of years to the time before Antarctic glaciers existed.

II. Flora and fauna

VEGETATION

The cold-desert climate of Antarctica supports only an impoverished community of cold-tolerant land plants that are capable of surviving lengthy winter periods of total or near-total darkness during which photosynthesis cannot take place. Growth must occur in short summer bursts lasting only a few days, a few weeks, or a month or two, depending upon such diverse factors as latitude, seasonal snowpacks, elevation, topographic orientation, wind, and moisture, in both the substrate and the atmosphere. Moisture is the most important single variable and is provided mainly by atmospheric water vapour and by local melt supplies from fallen snow, drift snow, and permafrost. Stream runoff is exceedingly rare. Extreme cold, high winds and aridity inhibit growth even in summer in most areas. There are, however, certain areas at high latitude and high elevation that have local "microclimates" formed by differential solar heating of dark surfaces, and these areas are able to support life. The importance of such microclimates was demonstrated by the second Byrd Antarctic Expedition (1933–35), which found that lichens in Marie Byrd Land grow preferentially on darker-coloured heat-absorbing rock.

Antarctic plants total about 800 species, of which 350 are lichens. Lichens, although slow growing, are particularly well adapted to Antarctic survival. They can endure lengthy high-stress periods in dormancy and almost instantly become photosynthetic when conditions improve. Bryophytes (mosses and liverworts), totalling about 100 species, predominate in maritime regions, but mosses can grow nearly everywhere that lichens grow. Liverworts are reported only from coastal and maritime regions. Numerous species of molds, yeasts, and other fungi, as well as fresh-water algae and bacteria, complete the listing of Antarctic plants. These forms are extremely widespread and are reported as far as 87° south latitude. In addition, Antarctic seas are highly productive in plankton plant life, particularly in near-shore, nutrient-rich zones of upwelling. Diatoms, a type of algae, are especially abundant.

Compared to more temperate regions, plants in Antarctica contribute little to processes of rock weathering and soil formation. Recent studies suggest that, under favourable moisture conditions, chemical interchange between mosses or lichens and their substrates may occur but at an exceedingly slow rate.

Today's barren Antarctic landscape little resembles ancient Paleozoic and Mesozoic ones with their far greater floral displays. Antarctic glaciation, probably beginning 50,000,000 years ago, forced the northward migration of all vascular plants (ferns, conifers, and flowering plants). Only nonwoody forms have again populated sub-Antarctic regions and have scarcely repenetrated the Antarctic zone.

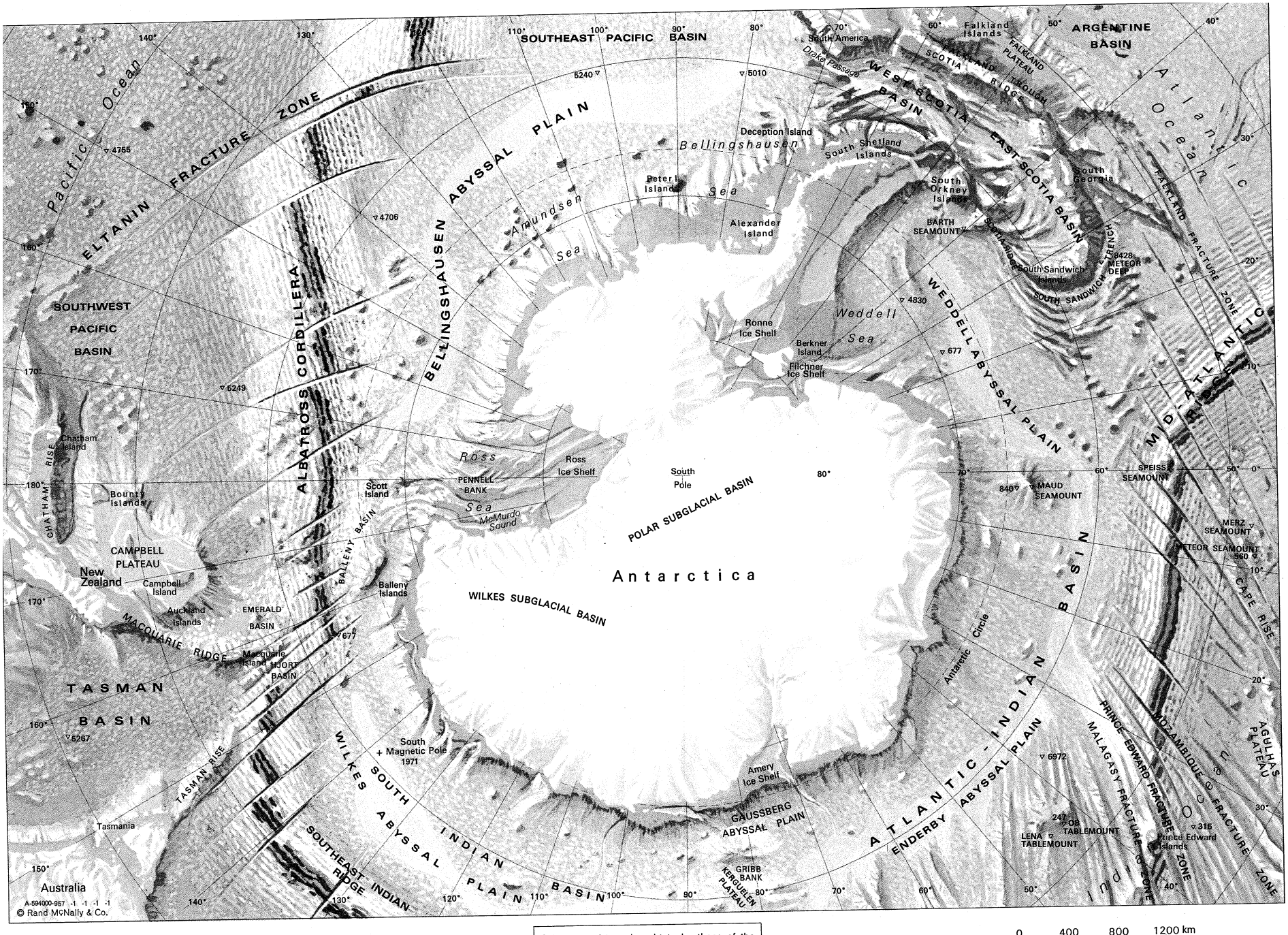
Plant life in the sub-Antarctic

In contrast to Antarctica, lying south of the Antarctic Convergence, the islands north of the Convergence in the sub-Antarctic botanical zone—including the South Georgia, Crozet, Kerguelen, and Macquarie islands—are characterized by an abundance of vascular plants of many species, at least 50 being identified on South Georgia alone. Thus, whereas plants reproducing by spores are characteristic of Antarctica, seed plants chiefly characterize sub-Antarctic regions.

Man has greatly influenced the natural ecosystem in many Antarctic and sub-Antarctic regions. A number of alien species of vascular plants near whaling stations have been introduced, and doubtless many alien microorganisms exist near all Antarctic stations. Alien herbivores, chiefly sheep and rabbits, have decimated plant communities on many sub-Antarctic islands. Rabbits have exterminated the native cabbage *Pringlea antiscorbutica* over wide areas on Kerguelen, and sheep have decimated tussock communities on South Georgia.

ANIMAL LIFE

Land fauna. The native land fauna is wholly invertebrate. Apparently climatically less tolerant and less easily dispersed, the fauna follows plant colonization of newly



A-594000-957 -1 -1 -1 -1
© Rand McNally & Co.

SOUTH POLAR REGION Depths in metres

Colours used are thought to be those of the various rocks and sediments on the sea floors. Differences in relief are shown by relief shading.

0 400 800 1200 km
0 400 800 1200 mi

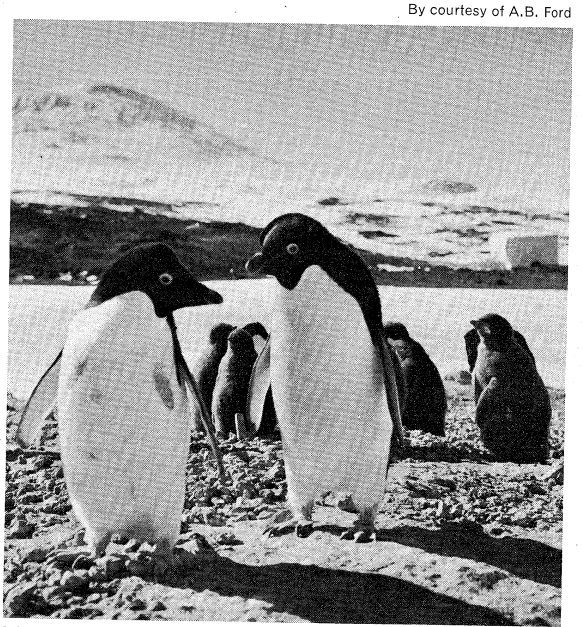
deglaciated regions and therefore is not as widely distributed. The Antarctic microfauna includes heliozans, rotifers, tardigrades, nematodes, and ciliate protozoans. The protozoans, in 29 species, dominate soil and freshwater communities. The terrestrial macrofauna consists entirely of arthropods, which number about 130 species. Of these, 54 are parasitic on birds and seals. The principal arthropod groups represented include Acarina, or mites (67 species); Collembola, or springtails (19); Mallophaga, or biting lice (37); Anoplura, or sucking lice (four); Siphonaptera, or fleas (one); and Diptera, or midges (two). Two species of beetles, probably alien, are known from islands near the Antarctic Peninsula. The dominant free-living forms, mites and springtails, live under stones and are associated with spore-reproducing plants.

Birds. The absence of mammalian land predators and the rich offshore food supply make Antarctic coasts a haven for immense seabird rookeries. A long geologic history of similar conditions has made Antarctica an ideal breeding and evolutionary centre for birds. One of the most characteristic, the penguin, evolved by about 40,000,000 years ago, as indicated by fossil remains in Seymour Island. About 45 species of birds live south of the Antarctic Convergence, but only three—the emperor penguin, Antarctic petrel, and South Polar (McCormick's) skua—breed exclusively on the continent. Two of the seven penguin species, the emperor and Adélie, are typically Antarctic and are distributed widely around the entire coastline. Gentoo and chinstrap penguins occupy northern Antarctic Peninsula coasts. Except for emperors, all of these, in addition to king, rockhopper, and macaroni penguins, are found widely in sub-Antarctic island regions. More than half the Antarctic breeding species of birds belong to the order Procellariiformes, including mainly petrels and a few albatrosses. The Antarctic and sub-Antarctic birds also include species of cormorants, pintails, gulls, terns, sheathbills, and pipit.

Penguins

Recent banding and recovery studies show that some Antarctic birds travel throughout the world. One giant petrel caught alive at Fremantle, Australia, had been banded as a nestling only two months previously in the South Orkney Islands. In that time span, it had learned to fly and had travelled no less than 10,000 miles, about halfway around the world. Rare sightings of skuas and petrels far in the continental interior, even near the South Pole, suggest that these powerful birds may occasionally cross the continent. Experiments show that Antarctic birds, including the flightless penguin, have strong homing instincts and excellent navigational capability, apparently the result of a highly developed sun-azimuth orientation system and biological clock mechanism even with the sun remaining continuously high. Adélie penguins released as far as 1,900 miles from their nests are known to have returned within a year.

By courtesy of A.B. Ford



Adélie penguins at Cape Royds rookery on Rose Island. In the background is Mt. Erebus.

Feeding habits of birds

Feeding habits vary widely from species to species. Most depend on the abundantly provisioned larder of the sea. The seabirds feed mainly on crustacea, fish, and squid, mostly at the surface or, in the case of cormorants and penguins, at depths of up to about 150 feet. Shore birds forage for mollusks, echinoderms, and littoral crustacea. Sheathbills, the southern black-backed gull, giant petrels, and skuas feed occasionally, as allowed, on other birds' unguarded eggs. The voracious skua and giant petrel are even known to attack the young or weak of other species, particularly penguins. Low aircraft overflights of rookeries are now prohibited because frightened adults leave chicks and nests undefended from predators.

Dependent upon seafood, most birds leave the continent each autumn and follow Antarctica's "secondary" coastline as the ice pack builds northward. The emperor penguins, however, are unique exceptions and remain behind as solitary guardians (other than man) of the continent through the long winter night. The emperors, once thought rare, are now estimated to number more than 1,000,000 in about 25 known colonies. These giant birds weigh from about 55 to 100 pounds and stand to nearly four feet in height. They congregate in close-packed hordes, females each autumn laying a single egg that their mates incubate by carrying on their broad feet beneath a warm fold of abdominal skin. Whereas other penguins build pebble nests for summer use, the emperors, without nests, have uniquely adapted to violent winter conditions in which nest maintenance would be impossible in the driving snows of Antarctic blizzards.

Sea life. The prolific zooplankton of Antarctic waters feed on the copious phytoplankton and, in turn, form the basic diet of whales, seals, fish, squid, and seabirds. The Antarctic waters, because of their upwelled nutrients, are more than seven times as productive as sub-Antarctic waters. The most important organism in the higher food chain is the small, shrimplike krill, *Euphausia superba,* only an inch or two in length when mature. But for their habit of congregating in vast, dense schools, they would have little food value for the large whales and seals. Their densities are great, however, and a whale, with built-in nets of baleen and hairlike fibres, can strain out meals of a ton or more in a few minutes. During the three to four months spent in Antarctic waters, the original population of baleen whales alone could consume an estimated 150,000,000 tons of krill. Animals on the sea bottom of the near-shore zone include the sessile hydrozoans, corals, sponges, and bryozoans, as well as the foraging, crablike pycnogonids and isopods, the annelid worm polychaeta, echinoids, starfish, and a variety of crustaceans and mollusks. Winter ice and anchor ice however, keep the sublittoral zone relatively barren to about 50 feet in depth.

The shrimplike krill

Fishes

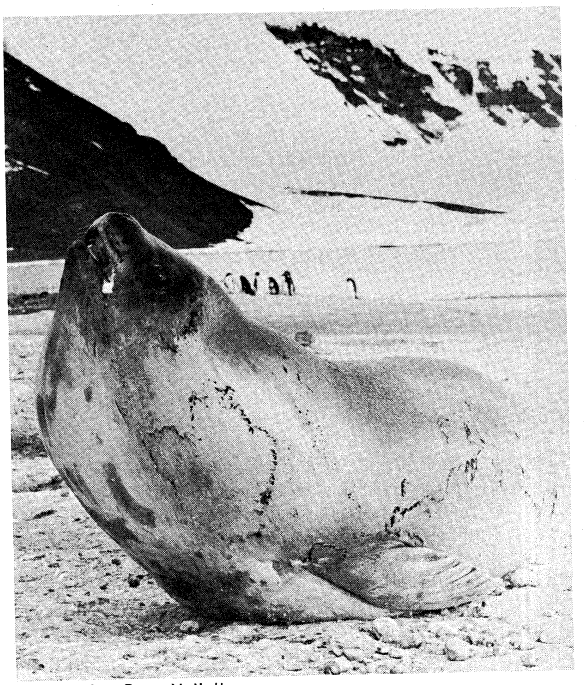
Of the nearly 20,000 kinds of modern fish, no more than about 100 are known from seas south of the Antarctic Convergence. Nearly three-quarters of the 90 or so sea-bottom species belong to the Nototheniiformes, the Antarctic perches. At sea bottom there are also the Zoarcidae, or eel-pouts; the Liparidae, or sea snails; the Macrouridae, or rat-tailed fishes; and the Gadidae, or cod-like fishes. Rare nonbony types in the Antarctic zone include hagfish and skates. Many species of deep-sea fish are known south of the Antarctic Convergence, but only three, a barracuda and two lantern fishes, seem to be confined to this zone. Antarctic fishes are well adapted to the cold waters; the bottom fish are highly endemic, 90 percent of the species being found nowhere else. This supports other biologic and geologic evidence that Antarctica has been isolated for a very long time.

Seals

Antarctic native mammals are all marine and include seals (pinnipeds) and porpoises, dolphins and whales (cetaceans). Only one otariid, or fur seal, breeds south of the Antarctic Convergence; four species of phocids, or true seals—the gregarious Weddell seal, the ubiquitous crab-

eater seal, the solitary and aggressively carnivorous leopard seal, and the rarely seen Ross seal—breed almost exclusively in the Antarctic zone, and another, the southern elephant seal, breeds near the Convergence at South Georgia, Kerguelen, and Macquarie islands. The sea lion, an otariid, is plentiful in the Falkland Islands but probably never ventures into the cold Antarctic waters. The fur seal and the elephant seal are now regenerating after near extinction from excessive hunting: the former may now number about 30,000 or more and the latter about

E.E. Gless



Ross seal at Cape Hallett.

700,000. Weddell seals are thought to number about 500,000, the crab-eater about 5,000,000 to 6,000,000, and the Ross seals about 50,000. Weddell seals are unique in being able to survive under fast ice, even in winter, by maintaining open breathing holes with their teeth. The leopard seal, armed with powerful jaws and huge canines, is one of the few predators on adult penguins. A number of mummified seal carcasses, chiefly crab-eaters, have been found at distances up to nearly 30 miles from the sea and elevations up to about 3,000 feet in the McMurdo dry valleys. Finding no food in such inland wanderings, the crab-eaters eventually died, and their leathery carcasses were preserved by the coldness and aridity of the climate.

Whales

Whales and their cetacean relatives, the porpoises and dolphins, range widely from the Arctic to the Antarctic, in all oceans and seas. A number of species range to, but generally not across, the Antarctic Convergence and so are considered only peripheral Antarctic types. Among the fish- and squid-eating toothed whales, or odontocetes, are a few peripheral Antarctic porpoises and dolphins and the pilot whale. More typical of the Antarctic are the killer whale, sperm whale, and rare bottle-nosed, or beaked, whale. Seven species of baleen, or whalebone, whales also inhabit Antarctic waters, subsisting on the plentiful krill; these include the southern right whale, humpback whale, and four kinds of rorqual—the blue whale, the fin whale, the sei whale, and the lesser rorqual, or minke. The pygmy right whale is endemic to Antarctic and sub-Antarctic waters. The killer whale, one of the most intelligent of marine animals, hunts in packs and feeds on larger animals, such as fish, penguins and other aquatic birds, seals, dolphins, and other whales. Despite the name, there is no authenticated account of attacks on humans near Antarctica. Excessive slaughter in the past has drastically decimated stocks of the larger whales, particularly the giant blue whales, the largest mammals on Earth. Near extinction, the blue whales are now protected by international agreement.

Alien mammals that now reside semipermanently in Antarctic and sub-Antarctic regions include sheep, rabbits, dogs, cats, rats, mice, and human beings. Effects on local ecosystems are great, from pollution of station areas by human wastes to erosion from overgrazing by sheep and to decimation of bird populations by dogs and cats and of whale and fur-seal stocks by humans. Even so, Antarctica remains by far the least contaminated land on Earth. Under the Antarctic Treaty, it is designated as a special conservation area, and many former activities of men are now prohibited in an attempt to preserve the natural ecological system of the unique environment.

Conservation

III. Natural resources and resource development

Antarctica, it has been suggested, may have become a continent for science because it was useful for nothing else. Certainly, the great success of the Antarctic Treaty and of the political experiment in international cooperation is in no small way attributable to the fact that exploitable resources have not been found. Articles of the treaty do not exclude economic activities, but neither do they set up jurisdictional procedures in the event that any are undertaken.

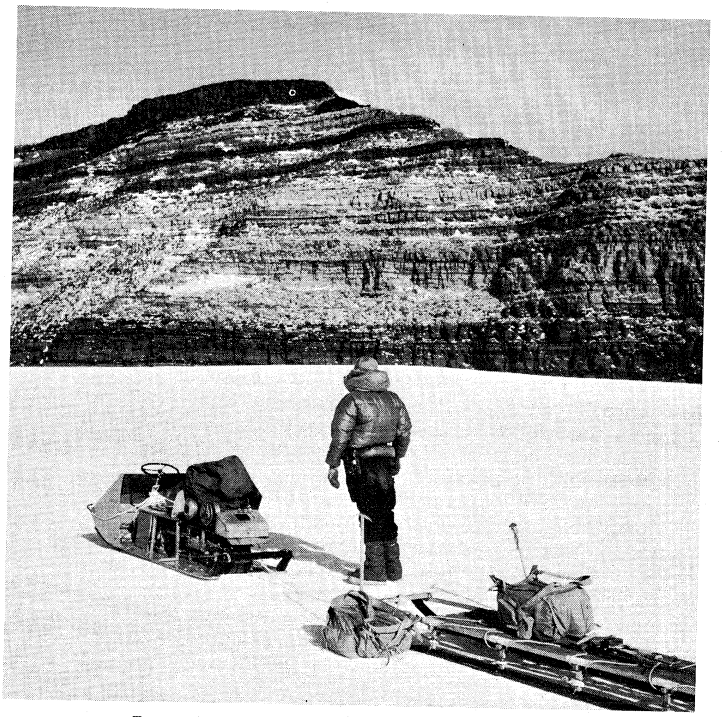
Many territorial claims existed before the signing of the Antarctic Treaty in 1959, and these claims clearly were not abrogated by signatory nations. Article IV states, in part:

> Nothing contained in the present Treaty shall be interpreted as a renunciation by any Contracting Party of previously asserted rights or claims to territorial sovereignty in Antarctica.

Multiple claims in some regions have never been resolved by international courts, and a number of countries, including the United States, recognize the validity of no claims in the absence of permanent habitation and settlements on the continent. Increasing economic pressures will force mineral and petroleum exploration into more and more remote regions as resources are gradually depleted in other lands. The stage will eventually come, though probably not in this generation and perhaps not even in the next, when market and technological conditions will make it economically feasible to carry the search to Antarctica and its continental shelves. The problems of rights of ownership and development will provide a severe test for the Antarctic Treaty in that event, but these political problems would have to be solved before any development could proceed.

Economics of resource development

Most early Antarctic expeditions through the 19th century were directly or indirectly of economic incentive. For some, it was the search for new trading routes; for others, it meant the opening of new fur-sealing grounds; still others saw a possibility of mineral riches. Natural-resource exploitation to date has centred in the sub-Antarctic and Antarctic seas, and virtually none has yet occurred on the continent. In a recent analysis of resource potentials, "Antarctic natural resources" were defined as "any natural materials or characteristics (in the Antarctic region) of significance to man." By this broad definition, the term includes not only biologic and mineral resources but also the land itself, water, ice, climate, and space for living and working, recreation, and storage. "Economic" resources are those that can be used or exported at a cost that is less than their value. Any attempted appraisal must therefore be continually re-evaluated in terms of current market values, logistical costs, and technological developments. Few known Antarctic resources have any economic importance in terms of present-day estimates of these factors. The factors are complexly interrelated and difficult to assess for the present, let alone the future. For example, technological advances that could allow development in Antarctica might instead allow development of marginally economic resources in other regions. Moreover, by the time it might become feasible for development of an Antarctic resource, such as petroleum, if indeed one exists, other suppliers for the market might be found, such as, in this case, fusion reactors or solar or geothermal energy, which would greatly change cost factors.



Exposed stratified rock of the Beacon Group in the Pensacola Mountains near the South Pole. Flat-lying Permian siltstone and coal measures are capped by diabase sill.
By courtesy of the U.S. Geological Survey

MINERAL RESOURCES

Antarctic cost levels. Antarctic mineral resources do exist, as does the technology to remove and transport them. The critical factor in the equation for determining economic feasibility is the cost of logistics, and this, at the moment, greatly exceeds the potential value of any known resource on the continent. Factors in the equation could conceivably change, but this does not seem likely, to any significant extent, in the foreseeable future.

Difficulties of transport

Transport of supplies and personnel is the principal cost factor. Although long-distance ocean shipping ordinarily is relatively inexpensive, in Antarctica icebreakers are generally needed to open a channel, by mid-December, to McMurdo Station, and, thus, freight costs to this station are estimated to run at least twice and perhaps ten times that for similar cargos over the same distances to nonpolar ports. Without allowing for icebreakers, delivered cost of fuel oil at McMurdo Station, for instance, is calculated to be about twice the cost at United States ports. Delivery by ski-equipped aircraft to the inland stations brings the cost to about ten times that at McMurdo. Figures such as these suggest that commercial operations might well be feasible in near-shore regions requiring little or no icebreaker assistance and that commercial activities inland are at present virtually precluded even for the most valuable conceivable resource.

Reports of mineral deposits. The geology of Antarctica is now known sufficiently well to allow rather certain prediction of the existence of a variety of mineral deposits, some probably large. The fact that none of significant size, besides coal, is now known to exist is merely the result of inadequate sampling. With the amount of ice-free terrain in Antarctica estimated at somewhere between 1 and 5 percent, the probability is practically nil that a potential ore body would be exposed. Moreover, whereas generations of prospectors have combed temperate and even Arctic mountains, Antarctic mountains have been visited mostly only by reconnaissance parties since the IGY.

Parallels between Antarctica and other continents

The high degree of certainty that mineral deposits do indeed exist is based on the close similarity between known geologic areas of Antarctica and those of mineral-rich provinces of South America, South Africa, and Australia. This comparison is valid whether or not the hypothesis of Gondwanaland breakup and continental drift proves to be correct. The gold-producing Witwatersrand beds of South Africa parallel the terrains of western Queen Maud Land. The copper-rich Andes of western South America parallel the young mountain belt of the Antarctic Peninsula. If petroleum prospecting in the Agulhas Basin off South Africa and the Otway Basin off southern Australia is successful, promising basins may lie in mirror-image position on Antarctic continental shelves near Coats Land and the Adélie and George V coasts. In any event, severe problems with icebergs, moving ice packs, and ice shelves would be encountered in any work in offshore Antarctic waters.

A great variety of metal minerals associated with ore deposits on other continents are also reported from Antarctica and include stibnite (antimony), chromite (chromium), chalcocite and chalcopyrite (copper), gold, galena (lead), molybdenite (molybdenum), cassiterite (tin), and sphalerite (zinc). In addition, there are indications of sedimentary phosphate. None of these deposits is in a location or of size and grade that approaches economic utility. The search for sizable ice-buried concentrations of the metallic minerals will be a difficult prospecting venture requiring costly geophysical and geochemical surveying and core drilling, and the great thickness of ice precludes even these techniques for most inland areas.

Antarctic coal

Antarctic coals have long been known. The deposits, belonging to one of the world's most extensive coalfields, range from semianthracite to anthracite and even to natural coke. The coals are generally high in ash, fixed carbon, and oxygen and low in hydrogen and sulfur. Most are too high in rank for possible utility in coke production. As a fuel supply, the coals cannot even compete with costly imported fuels at stations in Antarctica itself, let alone in the export market. Recent studies indicate locally high concentrations of ferromanganese nodules on the floor of the southeastern Pacific and Scotia Sea, particularly in a belt beneath the Antarctic Convergence. There is little potential, however, for early exploitation, at least before similar deposits on more accessible continental shelves throughout the world are first consumed. In summary, although the popular literature makes recurrent allegation of great "mineral riches" in Antarctica, there is clearly no present economic basis for such claims.

NONMINERAL RESOURCES

Biological resources. Resources of the sea first attracted men to Antarctica and provided the only basis for commercial activity in this region for many years. Commercial fur sealing began as early as about 1766 in the Falkland Islands and rapidly spread to all sub-Antarctic islands in the zeal to supply wealthy markets of Europe and China. Immense profits were made, but the toll was equally immense. Early accounts relate that millions of skins were taken from the Falklands within a few years after 1784. Ships of many nations participated in the slaughter, which was uncontrolled by international conservation agreements, as exist today, and thus soon led to near extinction of the formerly giant herds. By 1825, little profit was to be had in searching out the few remaining fur seals, and the seamen turned to elephant seals for their oil. A slight recovery of fur-seal herds in the 1870s led to increased hunting and soon to the complete disappearance of the animal. For nearly half a century none were seen, but a few were hidden on remote coasts. Regeneration has been slow, and only a few small colonies are known now on South Shetland, South Georgia, and a few other islands.

Whaling

The whaling industry, the only significant economic activity in Antarctic regions in modern times, dates back only a little more than 60 years. The blue, fin, and sei whales have been hunted most and now have been largely driven out of Antarctic waters. The blue, an endangered species, is protected, but hunting continues for the dwindling stocks of the others. The whaling industry in the Antarctic Ocean has in the last decade or so declined by at least 80 percent and by 95 percent in the Antarctic zone south of latitude 60° S. Control of the high seas is excluded from the Antarctic Treaty, and, thus, the out-

look is poor unless international conservation agreements can be reached.

The shrimplike krill, thriving in the nutrient-rich Antarctic waters, is a presently undeveloped but potential resource. It has been suggested that a krill-based industry might rival fish-meal production elsewhere as a source of protein but would probably require high market prices to be economic. Antarctic fishing has never been developed and seems to have little potential for the future.

Other resources. A rich imagination can see many possible uses of Antarctica and its materials. The continental ice sheet contains nearly 90 percent of the world's ice—a huge potential supply of nonsalt water—but any economic value is precluded by delivery costs. Antarctica has been proposed as a long-term deep-freeze storage site for grain and other foods, but calculations show that such usage cannot be economic, because of excessive shipping, handling, and investment costs. The Antarctic Treaty, at least until 1990, prevents the continent from being used as a site for radioactive-waste disposal and storage. Antarctica and its nearby islands could play an important role in wartime, particularly in the Scotia Sea region and Drake Passage, for control of interocean shipping. German commerce raiders, for example, in 1940–41, made considerable use of Kerguelen Island for this purpose. The Antarctic Treaty rules out military use, however, and with increasing capability in long-range aircraft, rocketry, and satellite surveillance and re-entry, any possible military importance of Antarctica will likely decrease.

Tourism

Antarctica contains abundant scenic resources, and these have been increasingly exploited within the last decade and seem to provide the greatest potential for future economic growth. The tourist industry began in a modest way in January and February 1958, with 9- and 12-day tours to the Antarctic Peninsula area arranged by the Argentine Naval Transport Command. Since January 1966, yearly tourist ships have plied Antarctic coastal waters, stopping here and there for visits at scientific stations and at penguin rookeries. Polar visionaries see an all-weather landing strip for wheeled jet aircraft at Marble Point near McMurdo Sound; one or more hotels nearby, perhaps in one of the McMurdo dry valleys and served by helicopter from the jet runway; and possibly even a centre for the alpine sports of skiing and mountaineering. With such facilities, they believe, greatly increased numbers of tourists could be brought to the continent, and costs would be decreased accordingly. Many other problems, though, would be raised. Facilities would have to include methods of sewage and waste disposal that would not affect highly sensitive local ecosystems. It would be important to ensure noninterference with Antarctic flora and fauna. Search and rescue facilities would be required. International agreements would have to be reached for legal jurisdiction and handling of civil and criminal cases that would inevitably arise.

IV. Territorial evolution and political status

HISTORICAL BACKGROUND

A great many nations, large and small, played important roles in the discovery and exploration of Antarctica. Who first saw the continent is clouded in controversy. The Russian expedition leader Fabian Gottlieb von Bellingshausen, the Englishman Edward Bransfield, and the American Nathaniel Palmer all claim first sightings in 1820. The first actual sighting most likely was by Bellingshausen.

Terra Australis

In about the year 650, however, long before European geographers of the Middle Ages and the Renaissance were to conjecture about the mythical Terra Australis, a New Zealand Polynesian war canoe, under command of one Ui-te-Rangiora, probably sailed at least as far south as the frozen ocean, as told by Maori legend. The legendary vast size of the continent shrank to nearly its present one when in 1772–75 the Englishman James Cook circumnavigated the globe in high southern latitude, proving that Terra Australis, if it existed at all, lay somewhere beyond the ice packs that he discovered between about 60° and 70° S.

The period from the 1760s to about 1900 was one dominated by exploitation of Antarctic and sub-Antarctic seas, particularly along Scotia Ridge. Sealing vessels of many nations, principally U.S. and British but including Argentine, Australian, South African, New Zealand, German, and Norwegian, participated in hunting that eventually led to near extinction of the southern fur seal. Many also hunted whales, but the less-profitable whaling industry climaxed following World War I after the decline of sealing. Among the few geographic and scientific expeditions that stand out during this period are those of Bellingshausen, commanding the Russian ships "Vostok" and "Mirny," in a circumnavigation of Antarctica in 1819–21; Bransfield, on a British expedition charting part of the Antarctic Peninsula in 1819–20; Dumont d'Urville, on a French expedition in 1837–40, when Adélie Land was discovered and claimed for France; Charles Wilkes, on a U.S. naval expedition in 1838–42 exploring a large section of the East Antarctic coast; and James Clark Ross, on a British expedition in 1839–43, which discovered the Ross Sea and Ross Ice Barrier (now called Ross Ice Shelf), as well as the coast of Victoria Land.

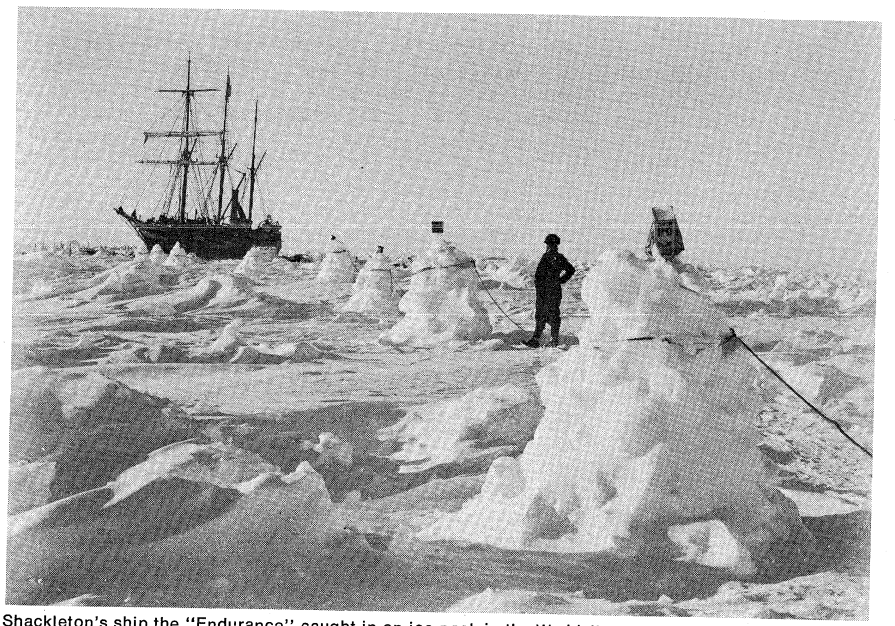
The "heroic era" of exploration. During the first two decades of the 1900s, commonly known as the "heroic era" of Antarctic exploration, great inroads were made not only into geographic but also scientific knowledge of the continent. The Englishmen Robert F. Scott and Ernest Henry Shackleton led three expeditions between 1901 and 1913, pioneering routes into the interior and making important geological, glaciological, and meteorological discoveries that provided a firm foundation for the scientific program of today. This era was preceded by two events that proved the feasibility of Antarctic overwintering: (1) the Belgian ship "Belgica," under command of Captain Adrien de Gerlache, became the first vessel to winter in Antarctic waters when, from March of 1898 to March of 1899, it was trapped and drifted in pack ice of the Bellingshausen Sea, and (2) a scientific party under Carsten E. Borchgrevink spent the next winter camped at Cape Adare, for the first planned overwintering on the continent.

The expeditions of Scott and Shackleton

Sledge probes deep into the interior were made by Scott on the British National Antarctic "Discovery" Expedition (1901–04) and by Shackleton on the British Antarctic "Nimrod" Expedition (1907–09) from base camps on Ross Island. New southing records were set by Scott, in company with Shackleton and E.A. Wilson, who reached 82°17′ S on Ross Ice Shelf December 30, 1902, and by Shackleton in a party of five, which reached 88°23′ S, a point within 97 miles of the pole, on January 9, 1909. The aerial age in Antarctica was presaged by Scott in 1902 on a captive balloon ascent for aerial reconnaissance, and the mechanical age by Shackleton in 1908 with the use of a motor car at Cape Royds, Ross Island. The successful experimental use of hardy Manchurian ponies and the pioneering of a route up the great Beardmore Glacier to the polar plateau by Shackleton paved the way for the epic sledging trip of Scott in 1911–12 to the South Pole.

Discovery of the Antarctic poles

National and personal prestige in attaining the Earth's poles, as well as territorial acquisition and scientific inquiry, provided strong motivation for polar exploration in the 1900s. The south magnetic pole, the point of vertical orientation of a magnetic dip needle, which was predicted by the German physicist Carl Friedrich Gauss to lie at 66° S, 146° E, inspired the unsuccessful quest, in about 1840, of the seafarers Wilkes, d'Urville, and Ross (Ross had earlier discovered the north magnetic pole). The point was later reached, on January 16, 1909, at 72°25′ S, 155°16′ E, on the high ice-plateau of Victoria Land by T.W.E. David and Douglas Mawson on a sledge journey from Cape Royds. The pole has migrated more than 550 miles since then to its present location on the Adélie Land coast. The South Pole of the Earth's rotation was the unattained goal of Shackleton in 1908–09 but was eventually reached on December 14, 1911, by Roald Amundsen of the Norwegian Antarctic Expedition of 1910–12 and, a month later, on January 17, 1912, by Scott of the British Antarctic "Terra Nova" Expedition of 1910–13. Whereas Amundsen's party of skiers and



Shackleton's ship the "Endurance" caught in an ice pack in the Weddell Sea off Coats Land during his Imperial Trans-Antarctic Expedition, 1914.
By courtesy of the Royal Geographical Society; photograph, Underwood and Underwood, New York

dog teams, using the Axel Heiberg Glacier route, arrived back at Framheim Station at Bay of Whales with little difficulty, Scott's man-hauling polar party—Scott, E.A. Wilson, H.R. Bowers, L.E.G. Oates, and Edgar Evans—using the Beardmore Glacier route, perished during a blizzard on the Ross Ice Shelf on the return march. (Two other related discoveries awaited IGY. The south geomagnetic pole, the theoretical pole of the Earth's magnetic field, on the East Antarctic Ice Sheet at 78°28′ S, 106°48′ E, was reached by a Soviet IGY tractor traverse on December 16, 1957. The pole of relative inaccessibility, the point most remote from all coasts, at 82°06′ S, 54°58′ E, was reached by a Soviet IGY tractor traverse on December 14, 1958.)

Trans-Antarctic expeditions

After Amundsen and Scott attained the South Pole, the idea that particularly haunted men's minds was that of an overland crossing of the continent. Conceived earlier by the Scotsman W.S. Bruce and the German Wilhelm Filchner to test the thought that a channel might exist connecting Ross and Weddell seas, a trans-Antarctic expedition was finally organized in 1914 by Shackleton. His ship, the "Endurance," was caught and crushed, however, in pack ice of the Weddell Sea, thus aborting one of the most ambitious polar expeditions hitherto planned. The idea lay dormant for several decades and came to fruition during IGY with the British Commonwealth Trans-Antarctic Expedition led by Vivian Fuchs. Using tracked vehicles and aided by aerial flights, the party left Shackleton Base on Filchner Ice Shelf on November 24, 1957, and by way of the South Pole reached the New Zealand Scott Base on Ross Island on March 2, 1958.

From World War I to IGY. The period between World Wars I and II marks the beginning of the mechanical, particularly the aerial, age of Antarctic exploration. Urgent wartime developments in aircraft, aerial cameras, radios, and motor transport were soon adapted for polar operation. By 1928, on November 16, the first heavier-than-air flight in Antarctica was made by Alaskan bush pilot C.B. Eielson and George Hubert Wilkins in a wheel-equipped Lockheed Vega monoplane. This pioneer flight was quickly followed by the better-equipped, aircraft-supported expeditions of the American naval officer Richard E. Byrd (1928–30, 1933–35, 1939–41, and 1946–47), which saw progressively increasing use of skiplanes and aerial photography. Byrd, on November 29, 1929, was first to fly over the South Pole (after having flown over the North Pole in 1926). His fourth expedition, called "Operation High Jump," in the summer of 1946–47 was the most massive sea and air exploratory assault hitherto attempted in Antarctica and involved 13 ships, including two seaplane tenders and an aircraft carrier and a total of 25 airplanes. Ship-based aircraft returned with 49,000 photographs that, together with those taken by land-based aircraft, covered about 60 percent of the Antarctic coast, nearly one-quarter of which had been previously unseen. Innovations by Byrd included the use of an autogiro in 1933–34 and six Sikorsky helicopters in 1946–47. Meanwhile, the courageous flight of Lincoln Ellsworth, an American, and Herbert Hollick-Kenyon, a Canadian pilot, across uncharted lands and icefields on the first aerial crossing of the continent from November 23 to December 5, 1935, clearly demonstrated the feasibility of aircraft landings and takeoffs, using unprepared surfaces, for inland exploration. These early aerial operations and the extensive use of ship-based seaplanes in Norwegian explorations of coastal Queen Maud Land during the 1930s were forerunners of today's extensive aerial programs.

The Byrd expeditions

National rivalries

The early discoveries led to a few controversies not only for territorial claims but also in geographical nomenclature. The struggle for national influence was especially acute in the slender peninsular land mass south of Scotia Sea that became known as O'Higgins Land (Tierra O'Higgins) to Chileans and San Martín Land (Tierra San Martín) to Argentines, after national heroes who helped in gaining independence from Spain. To Englishmen, it was known as Graham Land, after a former first lord of

Territorial Claims in Antarctica

country	sector	date of claim
Norway (Queen Maud Land)	20°W–45°E	January 14, 1939, by Norwegian Order in Council (also claimed Bouvet Island on January 23, 1928, and Peter I Island on May 1, 1931)
Australia (Australian Antarctic Territory)	45°E–136°E, 142°E–160°E	February 7, 1933, by British Order in Council; and June 13, 1933, by Australian Antarctic Territory Acceptance Act
France (Adélie Land)	136°E–142°E	April 1, 1938, by official decree (placed under Governor of Madagascar in 1924)
New Zealand (Ross Dependency)	160°E–150°W	July 30, 1923, by British Order in Council
Chile (Antarctic Peninsula)	90°W–53°W	November 6, 1940, by official decree
Great Britain (British Antarctic Territory)	80°W–20°W	July 21, 1908
Argentina (Antarctic Peninsula)	74°W–25°W	1942

the admiralty, and to Americans as Palmer Peninsula, after Nathaniel Palmer, the sealer alleged to have discovered the region in 1820. By international agreement, the region is now known simply as the Antarctic Peninsula, Graham Land its northern half and Palmer Land its southern half.

The first half of the 20th century is the colonial period in the history of Antarctica. Between 1908 and 1942, seven nations decreed sovereignty over pie-shaped sectors of the continent. Many nations—including the United States, the U.S.S.R., Japan, Sweden, Belgium, and Germany—carried out Antarctic exploration without lodging formal territorial claims, even though claims may have been announced by some of their exploratory parties. The United States government, for example, has never taken up the claims made in 1929 by Richard Byrd's expedition in the Ford Ranges of Marie Byrd Land (an area presently unclaimed), nor those made by Lincoln Ellsworth on aerial landings on November 23, 1935, in Ellsworth Land (an area now claimed by Chile) and on January 11, 1939, in the American Highland near Amery Ice Shelf of East Antarctica (an area now claimed by Australia). The German Antarctic Expedition of 1939 aerially photographed an extensive segment of Princess Astrid and Princess Martha coasts of western Queen Maud Land and, dropping metal swastikas over the region, claimed it for the Hitler government (the area is now claimed by Norway). Other claims were transferred, such as that made in 1841 by James Ross, who, after discovering and naming the coastal Ross Sea region after Queen Victoria, claimed it for the British crown; the area was later transferred to and is now claimed by New Zealand.

After the French claim of Adélie Land caused Americans to demand retaliatory action, the United States' official position was announced in 1924 by the Secretary of State Charles Evans Hughes:

> It is the opinion of this Department that the discovery of lands unknown to civilization, even when coupled with a formal taking of possession, does not support a valid claim of sovereignty, unless the discovery is followed by an actual settlement of the discovered country.

This policy has been reiterated many times since.

Few combative activities have marred the history of Antarctica. World War II brushed the continent lightly, only in that its nearby seas were used by Nazi commerce raiders. The threat of increased activity, however, prompted British warships to keep the northern Antarctic Peninsula under surveillance. On one visit to Deception Island in January 1943, it was discovered that Argentine visitors had been there the year before, leaving a brass cylinder with notice of claim to the peninsular region. The British obliterated the Argentine signs, hoisted the Union Jack, posted a notice of crown ownership, and returned the cylinder to the Argentine government. Reaction was swift. In London, growing concern that the territory might possibly be lost and that a pro-German Argentine government might control both sides of the vital Drake Passage linking Atlantic and Pacific sea routes resulted in a secret military plan, code-named "Operation Tabarin," to establish a base on Deception Island for closer watch. When the British returned to the island in February 1944, they found their earlier sign gone and an Argentine flag painted in its place. This they soon replaced with their own flag, and their base was established to back up the British claims to the region. Several other stations were built, and, with the conclusion of the war, the United Kingdom decided to maintain a continued presence in Antarctica.

The conflicting claims of Britain, Argentina, and Chile

Argentina and Chile both were stimulated to increase activities to back up their claims to the Antarctic Peninsula as a result of the British occupancy. (Chile had expressed a claim in 1940.) The Argentines had maintained a weather station in the South Orkney Islands continuously since 1903, and after 1947 they and the Chileans constructed bases at several sites. With the coming of the U.S. Ronne Antarctic Research Expedition (RARE) in 1947–48 to the old U.S. Antarctic Service East Base camp on Marguerite Bay, the peninsula protagonists—British, Argentine, and Chilean—became concerned that the United States might rejuvenate its claims. Any antagonism was soon overcome, and the Americans and British, for instance, joined forces for an arduous sledging journey down the east side of the peninsula. Violence flared on one occasion, in 1952, when a British party going ashore to establish a meteorological station at Hope Bay met with Argentine navy small-arms fire and retreated to their ship, the "John Biscoe." After a British government protest and an Argentine foreign-ministry apology for the mistake, the British shore party proceeded with construction.

Claims in the 1950s

By the mid-1950s, many nations had active Antarctic interests, some commercial, some scientific, and most political. In 1947–48, Australia had established stations on Heard and Macquarie islands and in 1954 built Mawson Station on the mainland coast of MacRobertson Land as a basis for its vast territorial claim. South Africans raised their flag over Prince Edward and Marion islands. France established permanent bases by 1953 in the Kerguelen and Crozet islands and surveyed much of the Adélie Land coast. In 1955, with icebreaker aid, Argentina established General Belgrano Station on the Filchner Ice Shelf. A profusion of British, Chilean, and Argentine bases had been built in such proximity to one another in the peninsula and nearby islands that their purpose seemed more for intelligence activities than for science. The international Norwegian–British–Swedish Expedition of 1949–52 carried out extensive explorations from Maudheim Base on the Queen Maud Land coast in the territory claimed in 1939 by Norway. The United States had shown little interest in Antarctica since the Ronne expedition and the U.S. naval "Operation Windmill," both in 1947–48 (the latter expedition was to obtain ground checks on the aerial photography of the previous season's "Operation Highjump"), but it continued its policy of nonrecognition of any claims. The Soviet Union had shown little interest, other than whaling, in Antarctica since Bellingshausen's pioneer voyage of 1819–21. On June 7, 1950, however, the Soviet government sent a memorandum to other interested governments intimating that it could not recognize any decisions on the regime for Antarctica taken without its participation. Such was the political climate on the continent during the organizational years for the coming International Geophysical Year of 1957–58.

IGY AND THE ANTARCTIC TREATY

The importance of coordinating polar science efforts was recognized in 1879 by the International Polar Commission meeting in Hamburg, Germany, and thus the 11 participating nations organized the First International Polar Year of 1882–83. Most work was planned for the better known Arctic, and, of the four geomagnetic and meteorologic stations scheduled for Antarctic regions, only the German station on South Georgia materialized. The decision was made at that time to organize similar programs every 50 years. In 1932–33 the Second International Polar Year took place, with 34 nations participating, but no expeditions were mounted to Antarctica.

The development of IGY. The idea for more frequent programs, every 25 years, was born at a small dinner party on April 5, 1950, at the Maryland home of the U.S. physicist James Alfred Van Allen, to take advantage of increasing technologic developments, interest in polar regions, and, not the least, the maximum sunspot activity expected in 1957–58. (The earlier, second polar year was a year of sunspot minimum.) The idea quickly germinated and grew: a formalized version was adopted by the International Council of Scientific Unions (ICSU), and in 1952 ICSU appointed a committee that was to become known as the Comité Spécial de l'Année Géophysique Internationale (CSAGI) to coordinate IGY planning. As plans widened to include the scientific study of the whole earth, more and more nations, eventually 67, showed interest in joining. Plans were laid for simultaneous observations, at all angles, of the sun, weather, the aurora, the magnetic field, the ionosphere, and cosmic rays. Whereas in the first polar year observations were confined to ground level and in the second to nearly

ICSU



Snow-cat traverse party near the Eufek Massif in Ronne Ice Shelf during the International Geophysical Year (IGY), 1958.
By courtesy of the National Academy of Sciences

33,000 feet by balloon, during IGY, satellites were to be launched by the United States and the U.S.S.R. for exploration of space. World Data Centres were established to collect all observations, making them freely available for analysis to scientists of any nation.

First Antarctic Conference

Two programs, outer space and Antarctica, were especially emphasized at an ICSU committee meeting in Rome in 1954. Antarctica was emphasized because very few geophysical studies had yet been made on the continent, because the south geomagnetic pole focusses auroral and cosmic ray activity in the Southern Hemisphere, and because on the eve of IGY almost half the continent had not yet even been seen by man. A meeting, the First Antarctic Conference, was then held in Paris in July 1955 to coordinate plans for expeditions, the advance parties of which were soon to set sail for the continent. Early tensions, due in part to overlapping political claims on the continent, were relaxed by the conference president's statement that overall aims were to be entirely scientific. Plans were laid for extensive explorations: 12 nations were to establish more than 50 overwintering stations on the continent and sub-Antarctic islands; the first regular aircraft flights to the continent were to be inaugurated (by the United States); massive tractor traverses were to be run in order to establish inland stations in West Antarctica (Byrd Station for the United States), at the south geomagnetic pole (Vostok Station for the U.S.S.R.), and the pole of relative inaccessibility (also for the U.S.S.R.); and an airlift by giant cargo aircraft was to be established in order to set up a station at the South Pole itself (Amundsen–Scott Station for the United States). Several major scientific programs were scheduled for Antarctica, dealing with the aurora and airglow, cosmic rays, geomagnetism, glaciology, gravity measurement, ionospheric physics, meteorology, oceanography, and seismology. Biology and geology were not primary studies of IGY.

Coastal bases were established in the summer of 1955–56 and inland stations the next summer for the official opening of IGY on July 1, 1957. For 18 months, until the end of IGY on December 31, 1958, a frenzy of activity not only in Antarctica but all over the world and in space resulted in a multitude of discoveries that revolutionized concepts of the Earth and its oceans, lands, glaciers, atmosphere, and gravitational and geomagnetic fields. Perhaps the greatest contribution was the political moratorium by the governments and the cooperative interchange between scientists of participating nations.

The Antarctic Treaty. With the ending of IGY the threat arose that the moratorium too would end, letting the carefully worked out Antarctic structure collapse into its pre-IGY chaos. In the fall of 1957, therefore, the U.S. Department of State reviewed its Antarctic policy and sounded out agreements with the 11 other governments with Antarctic interests. On May 2, 1958, Pres. Dwight D. Eisenhower issued identical notes to these governments proposing that a treaty be concluded to ensure a lasting free and peaceful status for the continent. Preparatory talks by the 12 governments were held in Washington, D.C., beginning in June 1958 and continuing for more than a year. A final conference on Antarctica convened in Washington on October 15, 1959. Agreement on the final draft was reached within six weeks of negotiations, and the Antarctic Treaty was signed on December 1, 1959. With final ratification by each of the 12 governments (Argentina, Australia, Belgium, Chile, France, Japan, New Zealand, Norway, South Africa, United Kingdom, United States, and U.S.S.R.), the treaty entered into full force on June 23, 1961. (After the original signing, Czechoslovakia, Poland, Denmark, The Netherlands, and Romania also acceded to the treaty.)

Washington conferences of 1958–59

The achievement of the Antarctic Treaty was an unprecedented landmark in political diplomacy. An entire continent was reserved for free and nonpolitical scientific investigation, as the preamble to the treaty outlines:

> . . . it is in the interest of all mankind that Antarctica shall continue forever to be used exclusively for peaceful purposes and shall not become the scene or object of international discord; . . . the establishment of a firm foundation for the continuation and development of such co-operation on the basis of freedom of scientific investigation in Antarctica as applied during the International Geophysical Year accords with the interests of science and the progress of all mankind; and . . . a treaty ensuring the use of Antarctica for peaceful purposes only and the continuance of international harmony in Antarctica will further the purposes and principles embodied in the Charter of the United Nations.

Contents of the Antarctic Treaty

Article I of the treaty provides for the peaceful use of Antarctica; Article II for international cooperation and freedom of scientific investigation; Article III for free exchange of plans, scientific results, and personnel; Article IV for the nonrenunciation of prior claim rights and for the prohibition of new claims and the citation of any activities during the treaty term as a basis for past or future claims; Article V for prohibition of nuclear explosions or waste disposal; Article VI for application of the treaty to all areas south of latitude 60° S, excluding the high seas, which come under international law; Article VII for open inspection of any nation's Antarctic operations by any other nation; Article XI for reference of disputes to the International Court of Justice if they can-

not be settled by peaceful negotiation or arbitration by involved parties; and Article XII for a review of the treaty after it has been in force for 30 years, if such a review is requested by any contracting party. An important provision of the treaty requires periodic meetings of representatives of signatory nations to take up occasional problems. Such meetings have agreed upon important measures for conservation of Antarctic flora and fauna and for the preservation of historic sites.

The post-treaty period. In order to continue and coordinate the international Antarctic scientific effort in the post-IGY period, ICSU in September 1957 organized the Special Committee on Antarctic Research, or SCAR. (In 1961 the word Scientific was substituted for Special.) The foundations for the committee were laid at its first meeting in The Hague, February 3, 1958. SCAR, a nonpolitical body, coordinates not only research activities in Antarctica itself but also, through ICSU, those Antarctic programs that relate to worldwide projects, such as the International Years of the Quiet Sun, the World Magnetic Survey, the Upper Mantle Project, the International Biological Programme, and the International Hydrological Decade. Member nations send representatives to periodic meetings of "working groups" for the various scientific disciplines. International scientific symposia are organized by SCAR for exchange of latest research results, on a timetable depending upon progress in the discipline. (For example, symposia on the geology of Antarctica were held in Cape Town in 1963 and in Oslo in 1970.) The great success of the political venture of the Antarctic Treaty depends in no small way on the achievements of SCAR and of the scientific and support teams in the field and laboratory.

SCAR

Scientific knowledge of Antarctica has increased steadily. Many important problems relating to knowledge of the entire Earth can be resolved only in the polar region. As pointed out by R.F. Scott's son Peter on an Antarctic visit in 1966, about half the topics of modern polar research could not even have been guessed at in his father's era only a half century earlier. Who at that time but a Jules Verne could have foreseen the advent in the 1960s of jet aircraft and turbine-powered helicopters and skiplanes and of data-recording machines powered by radioactive isotopes and polar-orbiting satellites that automatically collect meteorologic and upper atmosphere data across the continent and transmit it to a base collection station? The polar knowledge gained in the decade during and after IGY far outweighed that learned in the preceding millennia. The incredible advances in modern Antarctic science have only been possible by adapting to polar operation the great technologic advances in aircraft, oceanographic technique, and nuclear power (a nuclear power plant was installed at McMurdo Station in 1962 and a seawater distillation plant in 1965). Using ski-equipped aircraft, as well as turbine-powered helicopters, scientific and mapping teams have now explored nearly all major mountain regions.

The successful results of some scientific programs and the dead-ending of others force continual evaluation of scientific progress. As programs conclude or become automated and new programs develop, the logistic support must remain highly flexible to move into new areas. Many IGY stations have been closed and a few new stations opened. The U.S.S.R., long active only in East Antarctica, now also operates Bellingshausen Station in the Antarctic Peninsula and plans one or more additional stations in West Antarctica. The United States, operating mainly out of McMurdo Station (summer populations there average about 900 and winter about 200), maintains small year-around stations at the South Pole and in the Antarctic Peninsula and now uses former all-year stations, such as Byrd and Hallett, only for summer projects. Ten of the 12 signatory nations of the Antarctic Treaty still continue to maintain overwintering stations (Norway transferred its stations to South Africa, and Belgium closed its Roi Baudouin Base in 1967; both of these nations, however, continue sending summer research parties to Antarctica with the cooperation of other countries).

Conclusion. Antarctica, the Earth's coldest, remotest, and most unspoiled continent, provides a unique laboratory not only for natural science but also for political science. Antarctic and other IGY activities, the historic signing in 1959 of the Antarctic Treaty reserving an entire continent for nonpolitical scientific use, and the ensuing international community of polar-science efforts pointed the way for gradual thawing of the East–West Cold War of the 1950s and were direct forebears for a number of important international agreements in the 1960s on nuclear testing and the use of outer space.

Virtually all of Antarctica's surface terrain has now been seen, and most mountain regions have been aerially photographed, mapped, and visited by reconnaissance exploring parties of geologists, geophysicists, glaciologists, and biologists. A few areas are known in great detail. Least known is the bedrock terrain underlying the great ice sheets of the continent, and knowledge of this will be a principal goal in future decades of work. The principal export from Antarctica for many years, in any event, will probably be results of scientific investigations that add significantly to knowledge of the environmental system of the entire Earth.

BIBLIOGRAPHY

General: L.O. QUAM (ed.), *Research in the Antarctic* (1971); and the COMMITTEE ON POLAR RESEARCH, *Polar Research: A Survey* (1970), are the most current reviews of all fields of Antarctic research and exploration. Although a general review, R.E. PRIESTLEY and R.J. ADIE (eds.), *Antarctic Research* (1964), emphasizes British scientific results, particularly in the Antarctic Peninsula and Scotia Arc. The "Atlas of Antarctica," in Russian, 2 vol. (1966–69); numerous folios of the AMERICAN GEOGRAPHICAL SOCIETY, *Antarctic Map Folio Series* (1964 and continuing); and volumes of the AMERICAN GEOPHYSICAL UNION, *Antarctic Research Series* (1964 and continuing), provide modern maps and technical accounts of all phases of the research programs. Semitechnical to nontechnical reviews of current projects and exploration are in summary articles in the *Antarctic Journal of the United States* (bimonthly); a general nontechnical review is provided by H.G.R. KING, *The Antarctic* (1969). Factors for development of natural resources are appraised by N. POTTER in *Natural Resource Potentials of the Antarctic* (1969).

Geology and solid-earth geophysics: J.B. HADLEY (ed.), *Geology and Paleontology of the Antarctic* (1965); and J.C.F. TEDROW (ed.), *Antarctic Soils and Soil-Forming Processes* (1966), describe numerous topical studies, the former containing a comprehensive summary of Antarctic geology known to just after the IGY. R.J. ADIE (ed.), *Antarctic Geology* (1964) and *Antarctic Geology and Geophysics* (1972), are records of international symposia describing earth-science research by many nations. J.C. BEHRENDT and C.R. BENTLEY, *Magnetic and Gravity Maps of the Antarctic* (1968); and C. CRADDOCK and V.C. BUSHNELL (eds.), *Geologic Maps of Antarctica* (1970), provide modern maps and texts, the latter work covering most mountain regions at 1:1,000,000 scale.

Climate and Meteorology: M.J. RUBIN (ed.), *Studies in Antarctic Meteorology* (1966); S. ORVIG (ed.), *Climates of the Polar Regions* (1970); and W.S. WEYANT, *The Antarctic Atmosphere: Climatology of the Surface Environment* (1967), include topical studies and a few general review articles.

Glaciology and oceanography: M. MELLOR (ed.), *Antarctic Snow and Ice Studies* (1964); and A.P. CRARY (ed.), *Antarctic Snow and Ice Studies II* (1971), are collections mainly of topical studies of greatly varied scope; J.H. MERCER, *Glaciers of the Antarctic* (1967), provides a general review of Antarctica's glaciers. A.L. GORDON and R.D. GOLDBERG, *Circumpolar Characteristics of Antarctic Waters* (1970); and J.L. REID (ed.), *Antarctic Oceanology I* (1971), describe features of south polar water masses, their currents, and their interactions with subtropical and sub-Antarctic waters, as well as of the ocean floor and sediment carpet.

Biology: M.O. LEE (ed.), *Biology of the Antarctic Seas* (1964); G.A. LLANO (ed.), *Biology of the Antarctic Seas II* (1965); and G.A. LLANO and I.E. WALLEN (eds.), *Biology of the Antarctic Seas IV* (1971), are technical accounts of Antarctic sea life; a more popular summary is R.C. MURPHY, "The Oceanic Life of the Antarctic," *Scient. Am.*, 207:186–194 (1962). Technical accounts of terrestrial life are in J.L. GRESSITT (ed.), *Entomology of Antarctica* (1967); and S.W. GREENE *et al.*, *Terrestrial Life of Antarctica* (1967); a nontechnical review is provided by G.A. LLANO, "The Terrestrial Life of the Antarctic," *Scient. Am.*, 207:212–218 (1962). O.L. AUSTIN, JR. (ed.), *Antarctic Bird Studies* (1968), is a technical publica-

tion, whereas J. SPARKS and T. SOPER, *Penguins* (1967), comprehensively but nontechnically describes the characteristic flightless birds of sub-Antarctic and Antarctic coasts.

History: A complete list of all expeditions through the IGY is in B.B. ROBERTS, *Chronological List of Antarctic Expeditions* (1958). A thorough general account is in C.H. GRATTON, *The Southwest Pacific Since 1900* (1963); an account of early United States explorations is in K.J. BERTRAND, *Americans in Antarctica, 1775–1948* (1971). An interestingly written and thorough history to the IGY is W. SULLIVAN, *Quest for a Continent* (1957). Current explorations are reviewed each year in articles of the *Antarctic Journal of the United States.*

(A.B.Fo.)

Antennas and Wave Guides

Antennas are the components of radio, television, and radar systems that direct incoming and outgoing radio waves. One antenna frequently serves both purposes. Wave guides channel radio waves within the transmitting or receiving equipment itself and are frequently connected to the antenna. Antennas are usually metal and have a wide variety of configurations, from the mastlike devices used for radio and television broadcasting to the horns and large parabolic reflectors used to receive satellite signals and the radio waves generated in distant stars. Wave guides are often metal pipes or ducts.

Early antennas

The first antenna was devised by a German physicist, Heinrich Hertz, who about 1887 carried out a landmark experiment to test the theory of the British mathematician-physicist James Clerk Maxwell that light is only one example of a larger class of electromagnetic effects that could pass through air or empty space as a succession of waves. Hertz built a transmitter for such waves consisting of two flat, square metallic plates, each attached to a rod, with the rods in turn connected to metal spheres spaced close together. An induction coil connected to the spheres caused a spark to jump across the gap, producing oscillating currents in the rods. The reception of waves at a distant point was indicated by a spark jumping across a gap in a loop of wire. Hertz later improved his antenna by adding large, cylindrical reflectors.

The Italian physicist Guglielmo Marconi, the principal inventor of wireless (radio) telegraphy, constructed a number of antennas for both sending and receiving and discovered the importance of tall antenna structures in transmitting low-frequency signals. Frequency refers to the number of oscillations per second of an electromagnetic wave. In the early antennas built by Marconi and others, operating frequencies were usually determined by the antenna shape and size. Later the two functions were separated, and the work of determining the frequency was given to a new device, the oscillator, which generated the transmitted signal; the antenna could then be designed more effectively for transmission or reception at that particular frequency.

More powerful antennas were developed in the 1920s by combining a number of elements in a systematic array. Metal horn antennas were first constructed in the 1930s, following the development of wave guides to guide very high-frequency radio signals.

The horn antenna, actually the flared end of a wave guide, functions in much the same manner in directing high-frequency radio signals as a megaphone in directing the sound waves of the human voice.

TRANSMITTING ANTENNAS

Principles of antenna transmission

This section is concerned with transmitting antennas, that is, antennas for sending out a radio signal; most of the principles involved apply as well to receiving antennas.

The main characteristics of a transmitting antenna are its radiation pattern, its radiation efficiency, and its input impedance (see below).

The radiation pattern of a typical broadcast antenna is such that the antenna distributes the same relative energy in all horizontal directions and little or none in the vertical direction. On the other hand, a radar antenna concentrates most of its energy into a narrow beam that points in one direction, like a searchlight, that is, it is highly directive. Radiation patterns are often determined by setting up a receiver at a distant point, rotating the transmitting antenna and measuring the signal strength at the reception point for various angles of rotation. The relative signal strength is often displayed graphically in polar coordinates, as in Figure 1.

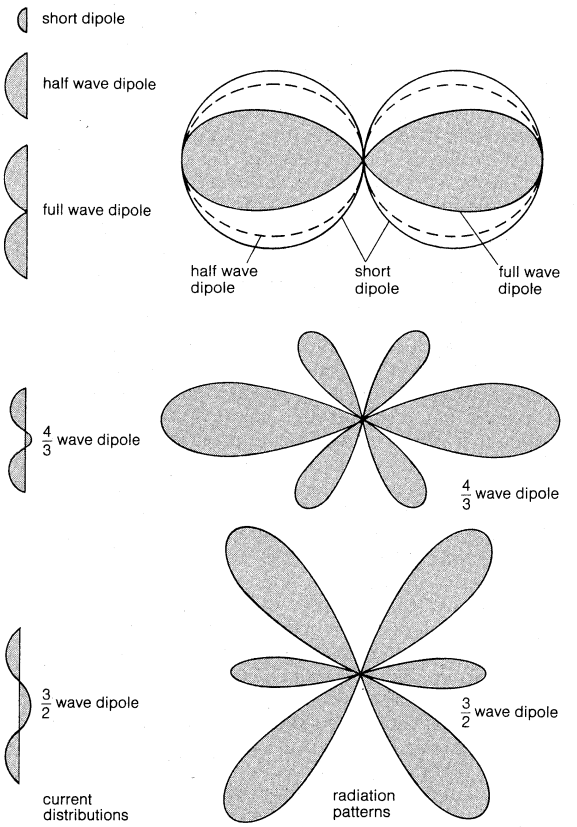


Figure 1: Radiation patterns of electric dipole antennas.
From W.L. Weeks, *Antenna Engineering* (© 1968); used with permission of McGraw-Hill Book Co.

The radiation efficiency of a transmitting antenna is defined as the ratio of the total power radiated from it to the total power fed into it. Though many high-frequency antennas radiate 95 percent or more of the power fed into them, this is not true of all antennas. If the dimensions of the antenna are small compared to the radiated wavelength or if it is constructed from materials that are poor conductors or inferior insulators, or if it is even located near such materials, a significant amount of power may be dissipated as heat.

The input impedance—*i.e.*, the opposition to the flow of alternating current—of an antenna determines how effectively it absorbs power from its power source, for example, a radio or TV transmitter. If the internal impedance of a transmitter is equal to the impedance of the antenna, the two are said to be matched, and maximum energy transfer takes place. When the antenna is small compared with the wavelength, and under certain other circumstances, the condition of being matched may be difficult to achieve, and only a small fraction of the available power may be transferred.

Both the radiation pattern and the input impedance vary with the frequency, or wavelength, at which the antenna is operating. The operating bandwidth of an antenna is the range of frequencies over which neither the radiation pattern nor the impedance varies greatly. Depending on antenna design, this bandwidth can be either small or large.

Single-wire types. There are a number of different types of transmitting antennas, each with its own characteristic advantages and disadvantages.

Dipole antenna. The simplest antenna is a straight wire or rod, divided at the centre to form a pair of terminals to which the source of power (transmitter) is connected. This is called a dipole antenna, and if the overall length of the antenna is approximately one-half wavelength at the operating frequency, it is called a half-wave

dipole. (The wavelength of a radio wave is the distance between crests or peaks of the wave as it travels through space or along an antenna wire.) If such an antenna is mounted vertically, it will radiate power uniformly in all horizontal directions. Radiation decreases vertically and in a direction parallel to the dipole (*i.e.*, straight up) it is essentially zero. If the overall length of a dipole is more than a wavelength, different patterns result (Figure 1).

Monopole antenna. For frequencies up to about 20 megahertz (20,000,000 hertz), and for many mobile radio applications, one-half of a dipole antenna is often mounted vertically above a conducting surface, such as the Earth or the metal roof of a car. This antenna is called a monopole.

If a monopole antenna is mounted horizontally and the length extended to several wavelengths, it is called a long-wire antenna. This type of antenna is useful for concentrating the radiated energy in a single direction.

Helical antenna. A wire or rod conductor can be shaped in the form of a helix to form a helical antenna. Its properties depend strongly on the ratio of helix diameter to wavelength. A wire monopole wound into a helix the diameter of which is very small compared to the wavelength performs much like a wire monopole of the same height; the advantage is that the input reactance is lower and it is self-resonant at a lower frequency.

For antennas of the type just described, the power from the transmitter is usually fed by a two-wire line or by a coaxial cable. A coupling device ensures optimum power transfer from the line to the antenna. Some typical feed arrangements are illustrated in Figure 2. Those depicted in 2F and 2G are for use with a coaxial cable.

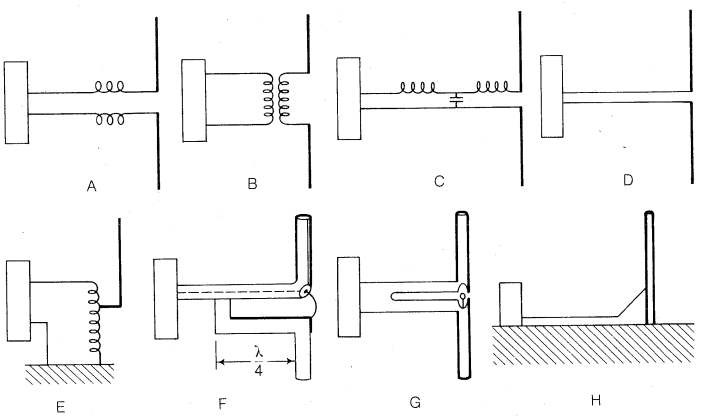


Figure 2: *Arrangements for feeding power to antenna.* (A) Inductor feed; (B) transformer feed; (C) network feed; (D) direct feed; (E) autotransformer feed; (F) and (G) coaxial feed using a balanced to unbalanced transformer (balun) with a quarter-wavelength (λ/4) stub; and (H) shunt feed.

The foregoing describes only single-wire antennas. To obtain certain desirable characteristics, such as greater directivity, a number of single-wire elements may be combined into an array or specific grouping.

Arrays. *Linear array.* The simplest type of array is one in which all of the elements are arranged in a single straight line called the line of the array. The orientation of the individual elements with respect to the line of the array is of major importance. If the elements are arranged end to end along the line (*i.e.*, collinearly), there is no signal radiated along the line of the array and the signal energy is broadcast symmetrically about the line of the array. On the other hand, the individual elements may be arranged perpendicular to the line of the array. In this case, the distribution of signal energy is not symmetric about the line of array; rather it depends on the element spacings and the relative phase of the excitations of the individual elements.

Planar array. Two or more linear arrays can be arranged side by side in the same plane to produce a planar array. Such an array is capable of transmitting equally well in both directions perpendicular to the plane of the array; if radiation is desired in one direction only, a doubling of the signal can be achieved by placing a reflecting screen or plate on one side.

Several planar arrays can be assembled, one behind the other, with proper spacing in between, to form a volume array that will still further concentrate the radiated energy in a single direction.

Electronically steerable array. In the arrays above, the characteristics are usually fixed by the way power is fed to each element. If the relative phase of the excitation is varied, the directional properties of the antenna are also varied. In the so-called electronically steerable array, the relative phase is varied electronically, permitting the direction of maximum radiation to be changed without physically altering the antenna elements. The steerable array is of value when large, bulky elements are involved and when rapid switching from one direction to another is desired, as in large radar installations.

A directional antenna normally radiates maximum energy in a single direction, with the radiated power falling off rapidly in other directions. In any practical antenna system, however, there may be other directions in which the radiation intensity peaks, although not as much as in the primary direction. These lesser peaks are called side lobes, and antenna engineering seeks to keep the energy in these side lobes as low as possible.

Reflectors and lenses. In many applications, such as radar, concentration of the radiated energy in a single direction may be accomplished by a large reflector or suitable lens. If the radiated energy emerges from the transmitter at the focus of a parabolic (dish-shaped) reflector, the energy will be focussed by the reflector much as the light in a searchlight and will give a highly directive beam. Variations of antennas of these kinds top the microwave relay towers, which carry telephone, television, and other communications media across the continents of the world. For some applications, the parabolic reflector may be replaced by plane reflectors if it is desired to change the direction of the radiated beam without otherwise changing its characteristics.

RECEIVING ANTENNAS

Similarities between transmitting and receiving antennas

The foregoing discussion has been concerned with transmitting antennas. Much of it applies equally to receiving antennas, which pick up energy from a radio wave and deliver it to a receiver. Receiving antennas do not respond equally to signals from every direction; their relative response depends on the direction of the incoming signal. For maximum reception, straight wire receiving antennas should be aligned parallel to the antenna transmitting the signal. A well-established principle in electronics holds that the radiation pattern of a structure used as a transmitting antenna is identical to its pattern when used as a receiving antenna. Consequently, by measuring the radiation pattern of an antenna, its receiving pattern can be accurately determined.

One measure of the effectiveness of a receiving antenna is termed its *effective area* and is a measure of the amount of signal power the antenna will intercept. There is a direct relationship between a receiving antenna's effective area and its directivity as a transmitter.

Whether the power delivered to the receiving equipment is sufficient will depend on the electrical noise levels in the receiving system. Competing noise may arise both from the receiving equipment and from outside noise sources. In conventional radio broadcasting, at frequencies below a few megahertz, for example, atmospheric noise due to thunderstorms will typically dominate all other sources of noise. For this reason, the design of receiving antennas for these low frequencies is usually not critical because incoming signals must be strong enough to override the high atmospheric noise. This is usually accomplished by using high transmitting powers. At frequencies above a few thousand megahertz, the receiver equipment noise often predominates, and receiving antennas are designed to introduce very little noise. One design technique is to make the directivity as high as possible; that is, to reduce or eliminate any extraneous pickup from directions other than the desired one.

At frequencies of the order of 100 megahertz, the external noise delivered in rural or suburban locations is typically of the same order of magnitude as the internal noise, except for bursts of man-made noise such as is sometimes generated by automobile ignition systems. Thus in the television and FM bands, for example, it is often worthwhile to elevate the receiving antenna to provide a line-of-sight path to the transmitting antenna and to increase the directivity.

Direction finding

One unique application of receiving antennas is their use to determine the direction of arrival of waves from a distant transmitter. Direction finding is accomplished by either comparing the phases of signals picked up by two antennas (the degree to which the two signals are in step) or by measurements of amplitude (the magnitude of the signal) from a rotatable antenna or a combination of both. The earliest direction-finding antennas were loop antennas. A loop antenna, when small compared to a wavelength, has a figure-eight radiation pattern with two rather sharp nulls (directions of little or no reception) between the two directions of strong reception. Thus, by rotating the antenna until a minimum signal is received, the direction of the transmitter can be determined. Any antenna having a sharp beam or a sharp null may be employed for the purpose of direction finding (see also NAVIGATION).

TRANSMISSION LINES

The electrical connection between the generator and the useful load, for example, between a radio transmitter and its antenna, is called a transmission line. Electrical energy travels along this line at speeds slightly less than that of light. The time delay associated with the specific length of a transmission line may or may not introduce important effects, depending on frequency; generally, the effects are insignificant at low frequency but very important at high frequency.

Figure 3A shows cross sections of conductors of common types of transmission lines; *a* is a coaxial line used extensively for cross-country telephone, radio, and television transmission; *b* is the parallel wire line (sometimes called twin lead) most familiar as a lead-in from FM and television antennas; *c* is a single wire over ground; *d* is a metal strip over ground (sometimes called microstrip when the ground plane is coated with an insulating material that supports the strip); and *e* is the shielded strip, or shielded microstrip. Lines *d* and *e* are frequently found in microwave circuits.

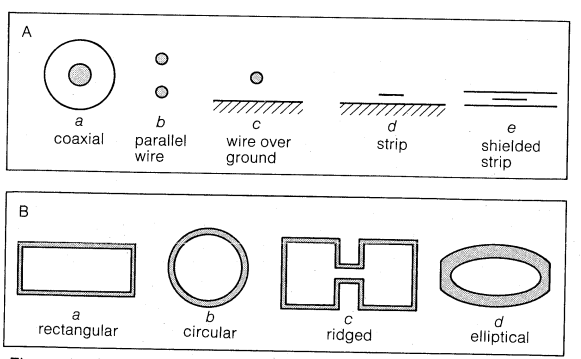


Figure 3: *Cross sections of typical devices for guiding radio waves.*
(A) Transmission lines; (B) wave guides.

WAVE GUIDES

Wave guides are used to guide high-frequency signals from the transmitter to the antenna or from the antenna to the receiver. They take many shapes and forms, but the most common are hollow metal tubes of rectangular or circular cross section (Figure 3B). The radio waves are confined to the interior of these tubes. One advantage of these wave guides is that they do not radiate any of the radio-frequency power passing through them, whereas some other types of transmission lines may radiate under certain conditions. In an ideal wave guide, the strength of the signal would not decrease as it travelled along the guide; in actual wave guides of convenient size, however, the power dissipated in the walls is relatively large.

The most commonly used wave-guide shape is a rectangle, with a height equal to about one-half its width. Circular wave guides are appropriate for such applications as radar in which two sections of wave guide must rotate with respect to one another. They are also of interest because of their ability to handle, simultaneously, tremendous numbers of voice, data, and television signals with small losses; for example, a two-inch diameter circular pipe of proper construction is capable of transmitting thousands of voice or television signals at millimetre wavelengths with very low losses.

There are other forms of wave guides for special purposes and applications. A longitudinal ridge in a rectangular or circular wave guide will permit operation at lower frequencies and over wider frequency bands. A closed metal pipe of any shape will guide waves. A dielectric (nonconductor) tube or rod, even without metal walls, will also guide waves; such structures are most familiar as "light pipes"; that is, plastic rods that will transmit light from one end to the other even though the rods are bent.

Radiation from wave guides

It is easy to alter a wave guide so that it will radiate. Any open-ended wave guide will radiate from the open end although some of the energy will be reflected back. The radiation can be improved by tapering the cross-sectional area upward in size; that is, by forming a horn at the end of the wave guide. If this is done smoothly, almost all of the energy will be radiated from the horn. Such wave-guide horn radiators are very effective and are members of a class of antennas known as aperture antennas. As with most other antennas, they can be used for either the transmission or reception of radio signals.

A wave guide will also radiate if slots or holes are cut in the walls. A number of such relatively small apertures may be arranged to form an array, usually a linear array if a single wave guide is involved. This type of radiating slot array finds application in situations that demand that the antenna be flush mounted to reduce wind resistance—for example, on a high-speed vehicle.

BIBLIOGRAPHY

Antennas: W.L. WEEKS, *Antenna Engineering* (1968), a relatively brief but general treatment of the whole field; J.D. KRAUS, *Antennas* (1950), an earlier standard text; HENRY JASIK (ed.), *Antenna Engineering Handbook* (1961), a detailed compendium; R.W.P. KING, *Theory of Linear Antennas* (1956), a detailed theoretical description of the properties of straight wire antennas; S.A. SCHELKUNOFF, *Advanced Antenna Theory* (1952), a discussion of the theory for a few antenna types presented by its originator; S. SILVER (ed.), *Microwave Antenna Theory and Design* (1949), an account of theory and practice developed during World War II for radar antennas.

Wave guides: G.C. SOUTHWORTH, *Principles and Applications of Wave-guide Transmission* (1950), a general account by one of the pioneers in the field; J.C. SLATER, *Microwave Electronics* (1950, reprinted 1969), a unified theoretical treatment; R.E. COLLIN, *Field Theory of Guided Waves* (1960), a detailed theoretical treatment of many aspects of the subject; N. MARCUVITZ (ed.), *Waveguide Handbook* (1951, reprinted 1964), detailed information on wave-guide technology gathered during World War II; W.L. WEEKS, *Electromagnetic Theory for Engineering Applications* (1964), presents wave-guide theory in terms of equivalent transmission lines.

(W.L.W.)

Anthropology

Etymologically, anthropology is the science of man. In fact, however, it is only one of the sciences of man, bringing together those disciplines the common aims of which are to describe man and explain him on the basis of the biological and cultural characteristics of the populations among which he is distributed and to emphasize, through time, the differences and variations of these populations. The concept of race, on the one hand, and that of culture, on the other, have received special attention; and although their meaning is still subject to debate, these terms are doubtless the most common of

those in the anthropologist's vocabulary. (See RACES OF MANKIND; HUMAN CULTURE for extended discussions of the definition of those concepts.)

THE CONCERNS OF ANTHROPOLOGY

Anthropology, which is concerned with the study of human differences, was born after the Age of Discovery had opened up societies that had remained outside the technological civilization of the modern West. In fact, the field of research was at first restricted to those societies that had been given one unsatisfactory label after another: "savage," "primitive," "tribal," "traditional," or even "preliterate," "prehistorical," and so on. What such societies had in common, above all, was being the most "different" or the most foreign to the anthropologist; and in the early phases of anthropology, the anthropologists were always European or North American. The distance between the researcher and the object of his study has been a characteristic of anthropological research; it has been said of the anthropologist that he was the "astronomer of the sciences of man."

Early emphasis on "primitive" man

Anthropologists today study more than just primitive societies. Their research extends not only to village communities within modern societies but also to cities, even to industrial enterprises. Nevertheless, anthropology's first field of research, and the one that perhaps remains the most important, shaped its specific point of view with regard to the other sciences of man and defined its theme. If, in particular, it is concerned with generalizing about patterns of human behaviour seen in all their dimensions and with achieving a total description of social and cultural phenomena, this is because anthropology has observed small-scale societies, which are simpler or at least more homogeneous than modern societies and which change at a slower pace. Thus they are easier to see whole.

What has just been said refers especially to the branch of anthropology concerned with the cultural characteristics of man. Anthropology has, in fact, gradually divided itself into two major spheres: the study of man's biological characteristics and the study of his cultural characteristics. The reasons for this split are manifold, one being the rejection of the initial mistakes regarding correlations between race and culture. More generally speaking, the vast field of 19th-century anthropology has been subdivided into a series of increasingly specialized disciplines, using their own methods and techniques, that have been given different labels according to national traditions. The Table shows the terminology current in North America and in continental Europe.

Cultural and physical anthropology

Thus two large disciplines—physical anthropology and cultural anthropology—and such related disciplines as prehistory and linguistics now cover the program that originally was set up for a single study of anthropology. The two fields are largely autonomous, having their own relations with disciplines outside anthropology; and it is unlikely that any researchers today work simultaneously in the fields of physical and cultural anthropology. The generalist has become rare. On the other hand, the fields have not been cut off from one another. Specialists in the two fields still cooperate in specific genetic or demographic problems and other matters.

Although prehistoric archaeology and linguistics are each comprehensively dealt with in separate articles, their links with cultural anthropology are notable in the present context. In posing the problem of the evolution of mankind in an inductive way, archaeology contributed to the creation of the first concepts of anthropology, and archaeology is still indispensable in uncovering the past of societies under observation. In many areas, when it is a question of interpreting the use of rudimentary tools or of certain elementary religious phenomena, prehistory and cultural anthropology (in the strict sense) are mutually helpful. "Primitive" societies that have not yet reached the metal age are still in existence.

Relations between linguistics and cultural anthropology (in the strict sense) are numerous. On a purely practical level the anthropologist has to serve a linguistic apprenticeship. He cannot do without a knowledge of the language of the people he is studying, and often he has had to make the first survey of it. One of his essential tasks, moreover, has been to collect the various forms of oral expression, including myths, folk tales, proverbs, and so forth. On the theoretical level, anthropology has often used concepts developed in the field of linguistics: in studying society as a system of communication, in defining the notion of structure, and in analyzing the way in which man organizes and classifies his whole experience of the world.

The fourth anthropological discipline, for which as a matter of convenience, one can reserve the name cultural anthropology, maintains relations with a great number of other sciences. It has been said of sociology, for instance, that it was almost the twin sister of anthropology. The two are presumably differentiated by their field of study (modern societies versus traditional societies). But the contrast is forced. These two social sciences often meet. Thus, the study of colonial societies borrows as much from sociology as from cultural anthropology. And it has already been remarked how anthropology intervenes more and more frequently in urban and industrial fields classically the domain of sociology.

Exchanges with other disciplines

There have also been fruitful exchanges with other disciplines quite distinct from anthropology. In political science the discussion of the concept of the state and of its origin has been nourished by anthropology. Economists, too, have depended on anthropology to see concepts in a more comparative light and even to challenge the very notion of an "economic man" (suspiciously similar to the 19th-century capitalist revered by the classical economists). Anthropology has brought to psychology new bases on which to reflect on concepts of personality and the formation of personality. It has permitted psychology to develop a system of cross-cultural psychiatry, or so-called ethnopsychiatry. Conversely, the psychological sciences, particularly psychoanalysis, have offered anthropology new hypotheses for an interpretation of the concept of culture.

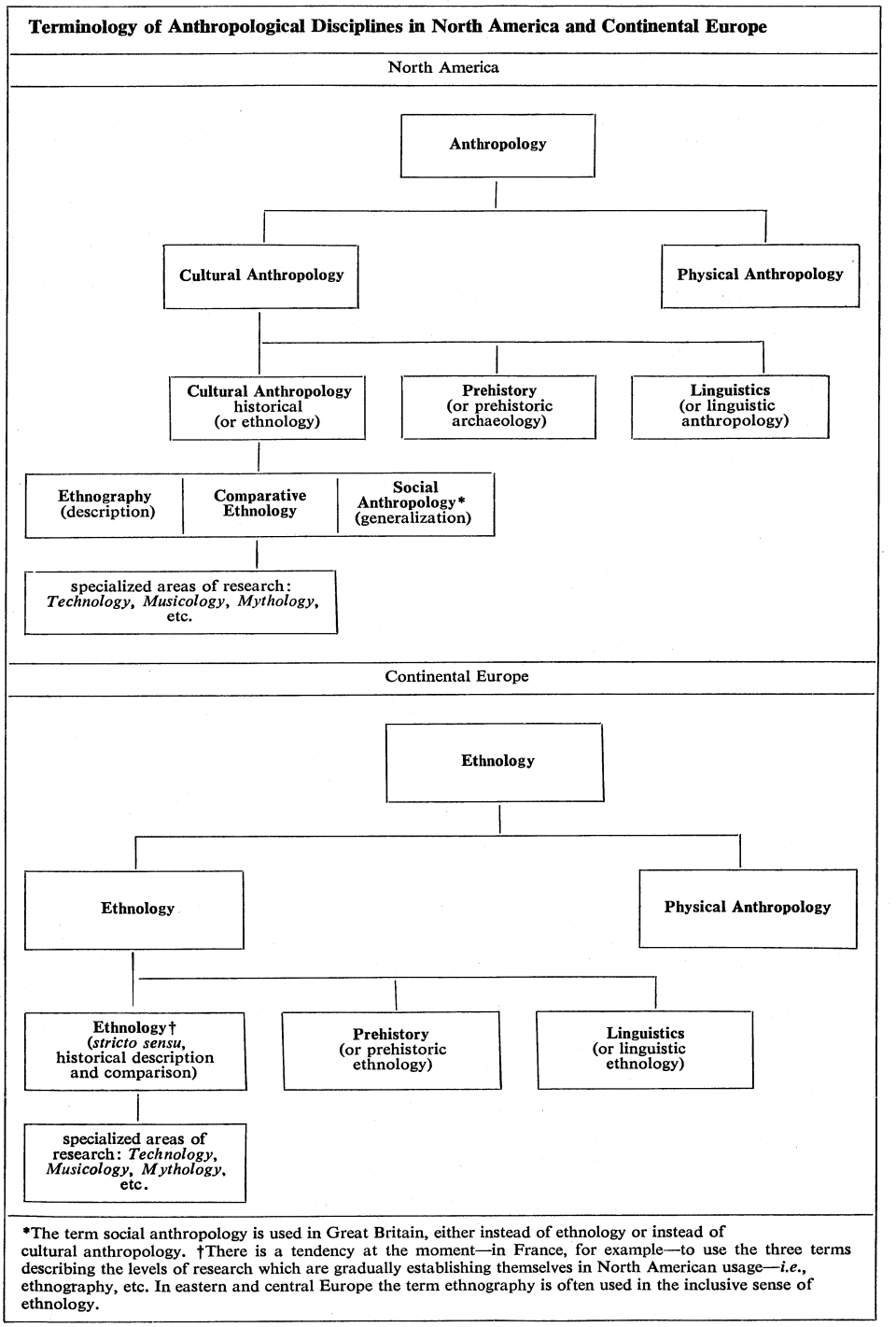
The link with history has long been a vital one because anthropology was originally based on an evolutionist point of view and because it has striven to reconstruct the cultural history of societies about which, for lack of written documents, no historical record could be determined. Anthropology has more recently suggested to historians new techniques of research based on the analysis and criticism of oral tradition. And so "ethnohistory" is beginning to emerge. Finally, cultural anthropology has close links with human geography. Both of them place great importance on man either as he uses space or acts to transform the natural environment. It is not without significance that some early anthropologists were originally geographers.

THE HISTORICAL BACKGROUND OF ANTHROPOLOGY

All human societies have been curious about how their customs originated and what the differences between their own culture and that of neighbouring societies might mean. Thus, in a sense they have all constructed their own anthropologies. But the interpretations put forward, even when they were founded partly on accurate observation, most often remained on the level of myth. Embryonic scientific thought began to appear in only a limited number of centres of civilization: in the classical Mediterranean world, in China, in the medieval Arab world, and in the modern Western world. Only in the West, however, did various ideas converge to bring about the birth of scientific anthropology in the 19th century.

A characteristic common to all these centres of civilization was the control that they exercised over vast areas and the opportunity that they enjoyed—through their soldiers, merchants, pilgrims, and missionaries—to gather observations on a wide variety of populations. Such a gathering of data was necessary in order even to begin to understand how men adapted to their environments, how they used their various economic, social, and political institutions, and how mankind evolved from simple to complex societies. Historians and philosophers among the ancient Greeks, Arabs, and Chinese all asked such

Terminology of Anthropological Disciplines in North America and Continental Europe



*The term social anthropology is used in Great Britain, either instead of ethnology or instead of cultural anthropology. †There is a tendency at the moment—in France, for example—to use the three terms describing the levels of research which are gradually establishing themselves in North American usage—*i.e.*, ethnography, etc. In eastern and central Europe the term ethnography is often used in the inclusive sense of ethnology.

questions. To take only the example of western Europe, many pertinent questions were posed by the French philosophers Jean Bodin and Michel de Montaigne as early as the 16th century, by the English philosophers Thomas Hobbes and John Locke in the 17th, and by the French philosophers Montesquieu, Rousseau, and Voltaire in the 18th, to mention only those who are often placed among the precursors of modern anthropology.

Historical changes affecting the birth of anthropology

19th-century beginnings. Modern anthropology began to take shape before the middle of the 19th century because of a series of innovations in the Western world. The last great phase of the discovery of the world had begun at the end of the 18th century. At the same time, political and intellectual revolutions had facilitated the questioning of certain religious dogmas, thus opening the way to the discussion of hitherto half-forbidden subjects. The 19th century, therefore, soon saw a revival of interest in and study of the origin of man, the unity or plurality of the human species, and the fixity or mutability of animal species.

Thus, the science of anthropology developed as an outgrowth of contemporary studies of the classification of human races; of the comparative characteristics of human anatomy; of the history of human settlements; of the classification of languages and the comparison of grammars; of the comparison between primitive and ancient societies; and of the historical development of man's economy and industry. Finally, about 1840, a principle for the study of human facts was proposed: the concept of evolution. This was even before Charles Darwin had published his celebrated *Origin of Species* (1859). This concept, arising in strong debates, provided the starting point for anthropology.

Evolutionism. Almost to the end of the 19th century, evolutionism determined the complexion of the new science. A major task of cultural anthropology was thought

to be that of classifying different societies and cultures and defining the phases and states through which all human groups pass—the linear interpretation of history. Some groups progress more slowly, some faster, as they advance from the simple to the complex, from the homogeneous to the heterogeneous, from the irrational to the rational. It suffices to quote an American anthropologist, Lewis Henry Morgan:

> As it is undeniable that portions of the human family have existed in a state of savagery, other portions in a state of barbarism, and still other portions in a state of civilization, it seems equally so that these three distinct conditions are connected with each other in a natural as well as necessary sequence of progress (*Ancient Society*, 1877).

Other quotations from a Scotsman, John F. MacLennan, or an Englishman, Edward B. Tylor, would take the same position.

Cultural anthropology, then, set out to analyze the totality of human culture in time and space. But by assuming a linear conception of history, it too often neglected the discontinuities and interferences of concrete history. Morgan, and particularly Tylor, however, sometimes felt the necessity of introducing the concept of the "diffusion," or spread, of cultural characteristics from one people to another—thus suggesting that characteristics could develop independently and converge and that a people could leap over "stages" of evolution by borrowing knowledge from others. Moreover, because it based itself on a theory that all mankind had a similar psychic outlook or that something called "human nature" was universal, anthropology also failed to take into account the fact that the same cultural trait can mean different things depending on the society in which it is found.

Impact of Marxist evolutionism

Marxism and the collectors. At the same time, in the second half of the 19th century another kind of evolutionism developed, that of Karl Marx and Friedrich Engels. Partly independent of anthropological evolutionism (Marx's *Critique of Political Economy* dates from 1859), partly linked to it (Engels' most important work appeared after Morgan's *Ancient Society* and made use of it), the Marxist theory laid stress on the causes of human evolution. A society was defined by its mode of production, on which its political, juridical, and ideological superstructures were allegedly based. These superstructures continued to exist after the mode of production had changed; and in the conflict that followed, this contradiction opened the way to a new type of society. Numerous anthropologists have taken the Marxist analysis into account, even if only to retain its historical view and to reject its economic determinism.

During this same period, especially toward the end of the 19th century, the tales of missionaries, traders, and travelling adventurers included an abundance of miscellaneous information that was collected in such works as Sir James Frazer's *Golden Bough* (1890) and Ernest Crawley's *Mystic Rose* (1902). These rather encyclopaedic collections of customs, religious and magical practices, and other curious data were read with relish by the intellectual community; the theories that accompanied the collections were equally appreciated by evolutionary-minded anthropologists, as the theories were meant to establish an evolutionary sequence of magical, religious, and scientific thought, using the data as evidence.

20th-century division. Since the beginning of the 20th century, cultural anthropology has stood out increasingly as a social science, whereas physical anthropology has moved forward of its own accord and asserted itself as a natural science, more certain of its cumulative character and of the continuity of its methodological developments, even when it launches into a new field of research.

DEVELOPMENTS IN CULTURAL ANTHROPOLOGY

Cultural anthropology has rediscovered some of the unappreciated contributions of its first masters. Thus in the eyes of some, Lewis Henry Morgan is a forerunner of the modern schools of thought. Cultural anthropology would continue to pose those kinds of problems that existed in the past: the role of evolutionism and of plural developments in cultural history; and relations between society and culture, between the individual and culture, and between the structure of society and its ever-recurring change.

By the beginning of the 20th century, many anthropologists had already begun to turn toward what might be called a more pluralistic viewpoint. To account for the variety of societies and cultures and the broadening of the differences that separated them, they suggested taking the total circumstances of each human group into account by considering the whole of its history, the contacts that it had had with other groups, and the favourable or unfavourable circumstances that had weighed on its development. Such a view was distinguished by a marked relativism: each culture represented an original development, conditioned as much by its social as by its geographical environment and by the manner in which it used and enriched the cultural materials that came to it from neighbours or others (through "diffusion") or from its own creativity (through "invention" and "adaptation").

20th-century schools of cultural anthropology. *Boas and the culture history school.* Cultural anthropology was also diversifying its concepts and its areas of research without losing its unity. Franz Boas, a German-born American, for example, was one of the first to scorn the evolutionist's search for selected facts to grace abstract evolutionary theories; and he inspired a number of students—Ruth Benedict, Alfred L. Kroeber, Margaret Mead, and Edward Sapir—to go out and seek evidence of man's behaviour among men in their natural environs, to go into the field to gather facts and artifacts and record observable cultural processes. He thus is known as the founder of the so-called culture history school of anthropology, which for much of the 20th century dominated American cultural anthropology.

Emphasis on field work and observation

Beyond this emphasis on field work and first-hand observation, it may also be said that Boas inclined toward what was called functionalism or the functional approach —an approach based on sociological theories of the late 19th and early 20th centuries that tended to liken societies to living organisms or machines, with interdependent parts. In the words of Melville J. Herskovits, one of Boas' students,

> the *functional* view, attempts to study the interrelation between the various elements, small and large, in a culture. Its object is essentially to achieve some expression of the unities in culture by indicating how trait and complex and pattern, however separable they may be, intermesh, as the gears of some machine, to constitute a smoothly running, effectively functioning whole (from *Man and His Works*, 1948).

Boas insisted upon this method of considering any single culture as a whole. Finally, by emphasizing the importance of collecting life histories, he drew attention to the problems posed by connections between culture and personality.

Mauss and the "sociological" school. In a similar way, Marcel Mauss, in France, influenced the characteristic tendencies of a whole generation of European sociologists and anthropologists, including Alfred Métraux and Claude Lévi-Strauss, and founded the Institute of Ethnology of the University of Paris; he also influenced such men as the noted British anthropologists Bronislaw Malinowski and Arnold R. Radcliffe-Brown. In general it may be said that Mauss, like Boas, was insistent upon studying social phenomena as a system—but in a slightly different fashion. Like many others of his time he conceived of systems as self-regulating or equilibrium-seeking, composed of elements that operate to maintain the integration or adaptation of the system. Mauss gave impetus, in fact, to what was called structuralism or the structural approach, which focussed more on society as an indivisible social organism than on society as an interrelation of individuals (the functionalist's emphasis). Like Boas, Mauss also tried to twin culture and personality—that is, cultural anthropology and psychology.

The "grand diffusionists." The large and influential American school of "culture history" anthropologists led by Boas should not be confused with a distinct and smaller group of Austro-German diffusionists, led by

The Kulturkreise school

Fritz Graebner and Wilhelm Schmidt, who constituted what has been called the "culture-historical" school in Europe. These latter, too, had rejected classical 19th-century evolutionism, but they were nevertheless inclined toward painting grand theories—principally the theory that out of a few ancient cultural centres or civilizations, born quite separately, there had developed all the array of cultures existing today. Diffusion, or the spreading of culture traits, in their view, was the prime force of human development, and all cultural development could be traced to a few inventive centres. Because they termed these original centres Kulturkreise, (or "cultural clusters"), they were also known as the Kulturkreise school of anthropology. This kind of pseudo-history was carried to even greater lengths by a British group of diffusionists, led by Grafton Elliot Smith and William J. Perry, who even named a single fountainhead of all cultural development—Egypt.

Functionalism and structuralism. Some schools of research that began to develop between the two world wars more or less vigorously rejected the historical approaches, sometimes denying any interest in them whatever. According to the cultural functionalists, including the followers of Malinowski, the only way to explain facts was to define the function that they performed currently in a given culture. The aim of all anthropological research, they held, should be to perceive the totality of a culture and the organic connection of all its parts. Consequently, comparison did not make sense: each culture was a unique reality. History, moreover, made no more sense; a culture was to be interpreted at one point in time, as if the age and the origin of the elements composing it were without importance. The only thing that counted was the function the elements performed *now*. Earlier anthropologists had talked of "survivals," customs or other cultural traits that survived from out of the past though no longer with any real function or meaning. But Malinowski would say, "There are no survivals"; everything current, according to the functionalists, has *some* function.

Whereas the name of Malinowski is supremely associated with the school of functionalism, the name of Radcliffe-Brown is known as one of the most important proponents of present-day structuralism. Relying on the concepts of formal mathematics and linguistics. Radcliffe-Brown and other structuralists tried to determine whether in cultural anthropology it was possible to reveal that which "suggests the character of a system" beyond empirical reality and which "alone is the true object of science" (Lévi-Strauss). A structure is not a sum of social relations, which are only the primary material from which the observer extracts "structural models." A structure is a system of which the members of the society being studied are not aware or only partly so. The model that the anthropologist constructs from the system is valid when the model's operation can account for all the observed facts. This exacting approach has proved particularly useful in studying kinship and marriage relations as well as myths. The difficulties of using this approach in other fields, as well as the fact that historical changes are difficult to include in this sort of static analysis, strengthen the objections that many workers in the field have raised against it.

Structure unperceived by society's members

Cultural psychology. One development of the interwar period led certain anthropologists to speak of a new subdiscipline, cultural psychology, or ethnopsychology, which is based on the idea that culture conditions the very psychological makeup of individuals (as opposed to the older notion of a universal psyche or human nature). In the 1930s, for instance, in her studies of the American Southwest, Ruth Benedict found that the ways in which the Pueblo Indians thought and reasoned were strikingly different from the ways in which their immediate neighbours thought and reasoned, even though their geographical environment was virtually identical. Her conclusion was that each culture over the ages had evolved and given to its members a unique "psychological set" or orientation toward reality and that this set actually determined how the members saw and processed information from the environment. Culture, in effect, affects the ways in which the mind works.

Studies in culture and personality have developed in many directions. Research into forms of child rearing, for instance, have called in question the universality of Freudian propositions concerning parent–child relationships. There have been many studies of value systems, which give a culture what has been called its "configuration," or of the personality types prized or rejected by each culture, or of the "national characteristics" of certain modern societies. The results of these studies have, however, been uneven in quality.

Neo-Marxism and neo-evolutionism Finally, certain theoretical tendencies of the 19th century have come back into favour. For political reasons Soviet anthropologists have conducted their research in the tradition both of Marxist analysis and of a fairly rigid evolutionism. Even their choice of subjects is sometimes linked to official ideology—as, for example, a program of religious anthropology aimed expressly at the "elimination of religious prejudice in the Russian population." Elsewhere, in France, for example, a brand of neo-Marxism has influenced a new generation of anthropologists to concentrate on analyses of primitive economies. Classical evolutionism, meanwhile, has been revived in the United States by some anthropologists who speak of "multilinear evolutionism" or many paths to modernization.

Current status of cultural anthropology. It is true that cultural anthropology has not reached a state of complete coherence. This is clear from the persistence of divergent national traditions and from the way in which research can be impregnated with explicit or implicit ideologies. It is also true that different schools of thought coexist in the same country and that cultural anthropology is not therefore based on a unified body of concepts, whereas a science is defined above all as a homogeneous language for interpreting a specific level of reality. A "science" of culture would seem possible only if anthropologists could free themselves of ethnocentrism and produce concepts and other elements that were universal, objective, and theoretically significant. The functionalists think they have fulfilled these conditions. The structuralists challenge this and, in their turn, try to fulfill the conditions. Thus cultural anthropology—as opposed, for example, to linguistics—has developed only very partially a terminology independent of a national or private language. These limitations are still encountered by most of the social sciences. But anthropology's primary aim—to permit cross-cultural comparability—makes the problem even more serious.

The problem of making cultural anthropology "scientific"

Anthropology is undergoing a crucial test of another kind. Its traditional objects of study—"primitive" or "traditional" cultures—seem to be disappearing. Either they are dying out because they find it impossible to adapt themselves to a modern world or they are transforming under the direct or indirect influence of modern industrial societies. Moreover, those that do remain at a folk level often take exception to being placed among societies that are the subject of anthropological study, seeing this as a manifestation of condescension and a vestige of domination.

Much anthropological research and study has entered the library or laboratory. One of the criticisms of Boas and others engaged in pure fieldwork was that they were collectors rather than systematizers. There is thus a considerable wealth of ethnographic data to be analyzed, collated, classified, and interpreted in order to be made useful. Files of information are being arranged in what are called Human Relations Area Files. More and more typologies are being constructed, typologies based on political systems or technology, or systems of kinship. In addition, new readings of the material are being attempted in the hope that mathematical formulations or models might be obtained. Also emerging is the study of insufficiently known societies by techniques of simulation.

Many anthropologists refuse to turn to the laboratory and continue to do fieldwork, either among Western populations or among modernizing, formerly colonial popu-

lations. They are joined in this task by researchers native to those populations. For some anthropologists these field studies provide an opportunity for a true anthropological experiment, determining how people respond to modernizing influences and how elements of the old culture evolve into those of the new. Such anthropologists tend to reject the concept that social systems seek integration and "equilibrium." Instead they propose a more "dynamic" interpretation of traditional societies and emphasize the role played therein by tensions and conflicts.

Economic development and anthropology

In any case, at a time when the problems of development are among the primary cares of the world, a growing number of anthropologists are devoting themselves to research the results of which can be used in political policy and decision making—whether they are employed directly by interested governments, or lent by foreign governments or international organizations, or recruited by foundations for study and development.

A significant development in the latter half of the 20th century has been the emergence of more and more non-Western anthropologists. Originally, cultural anthropology was a Western interest and endeavour, and it has continued to be dominated by Westerners. Even in non-Western countries where anthropology institutes and university departments have begun to multiply somewhat —as in Japan, India, and some Latin-American nations— anthropologists have remained rather constricted. Japan is a good example. Cultural anthropology as an independent science there is still young, having arisen largely only since World War II; and most Japanese anthropologists in the schools have had to be hybrid teachers, attaching themselves to sociology or social science departments and teaching sociology or some other related discipline in addition to anthropology. Not only have anthropology courses been few but also funds for field studies have been limited, so that there have been few lengthy and intensive studies; what research there has been has focussed largely on Japanese or other East or Southeast Asian communities. Furthermore, Japanese anthropologists have shared a problem faced by many non-Western researchers, in that the native language in which they write has not been as readily accessible to foreigners as have been western European languages. "International communication," the Japanese anthropologist Takao Sofue has noted, "has [thus] been seriously restricted with the result that Japanese scientists have been isolated from effective criticism from abroad" ("Social Anthropology in Japan," *American Behavioral Scientist*, 12:15–17, Jan.-Feb. 1969). It has also meant, of course, that they have not been sufficiently widely read abroad to make their influence felt. This problem, though, is not so serious in non-Western countries like India, where a European language constitutes a major language of scholarly communication.

The problems of applied cultural anthropology. From the anthropologist's point of view, applied studies—that is, research meant to give practical aid and guidance to governments and other organizations—have in many ways been an undoubted gain. Concerned as they so often were with the effects of social change, they offered the nearest approach to the controlled experiment in the social sciences. The specialized inquiries greatly deepened the knowledge of particular aspects of primitive society and culture, especially of economic and political organization, land tenure, and law. Scientific value apart, work in the applied field also offered to many anthropologists the purely human satisfaction of aiding backward peoples in their struggle to meet and master the forces of Western civilization.

The concrete gains derived by colonial governments were more difficult to assess, partly because the officials were not bound to act upon the anthropological findings and partly because the value of the findings was not always wholeheartedly accepted. Sometimes, it is true, the anthropologist found himself embarrassed by the excessive confidence of his employers that he had the key to all problems. More often, the employers were inclined to question whether anthropology was in fact as helpful and the information it provided as indispensable as enthusiasts would make it out to be. Some impatience was felt with the "academic" anthropologist who would insist on comprehensive studies when only some specific information was asked for, or who seemed to deal in a complicated fashion, using complicated language, with issues that to the practical man appeared straightforward. To all this anthropologists could reply that, though the knowledge they sought was not indispensable to government, it facilitated informed and smooth government.

But anthropologists also had to face another, more disturbing criticism—that they overemphasized the importance of tradition and were hostile to modern development. Nor was this view limited to colonial administrators; educated Africans and Indonesians openly expressed their distrust of a science the primary interest of which was in "primitive" peoples and which might play into the hands of reactionaries and upholders of "colonialism."

The theoretical researcher versus the technician

If these objections did not promise too well for the future of applied anthropology, anthropologists themselves had grown more cautious. They came to fear that the applied work might entice too many of the younger anthropologists away from general and theoretical research, so that the very progress of the discipline might be endangered. Conversely, the man fully committed to applied work, like the permanent government anthropologist, would be in danger of losing touch with universities and academic centres, and hence with the advances achieved in his discipline. He would turn into a mere technician, perhaps still useful to his employers but no longer truly representing anthropological knowledge.

There were graver problems of an ethical nature. A change of roles is forced upon the anthropologist when he is consulted on the best way to implement government policies. To be sure, he might see no cause for disagreeing with the policy, and the best way of imposing it might well be understood to be the one best serving the interests of the native peoples. Even so, the anthropologist, in abandoning the point of view of the scientist, must pronounce upon the merits and demerits of particular courses of action and thus introduce value judgments. Nor will the issues always be clear-cut and uncontroversial; in that case the anthropologist might have to take sides and argue from his own political and moral convictions. And if his recommendations had little chance against administrative considerations or the dictates of "higher policy," personal frustrations would be added to the dubiousness of his position.

On the other hand, if the anthropologist presented his facts without adding recommendations or warnings, he would be furnishing information that might be put to uses with which he could not in good conscience agree. Or again, he might be tempted to restrict his advice to the most efficient means for achieving certain ends, dismissing the ends themselves, the policy to be implemented as not of his concern—which would hardly diminish his ethical commitment.

All these issues were widely and on occasion heatedly debated among anthropologists. In an attempt to clear the air the Society for Applied Anthropology published in 1951 a carefully worded code of ethics. It appealed to the social conscience of the individual research worker and to his responsibility at all times to uphold the moral tenets of civilization—respect for the individual and for human rights and the promotion of human and social well-being. Not all anthropologists were prepared to endorse this assumption of a moral mission on the part of the "disinterested" scientist. The dilemma, then, though vital for the future of applied anthropology, remained unresolved. (Pa.M./Ed.)

DEVELOPMENTS IN PHYSICAL ANTHROPOLOGY

Early investigations. The history of physical anthropology has been, in large part, a history of man's attempt to determine his place in nature, to compare himself with other primates, and to interpret the physical differences among the races. A basic concept that constituted the organizing assumption of primate taxonomy, from its inception to well after the formal appearance of a theory of evolution in 1859, was the idea of a "great chain of be-

The "missing link"

ing," or a hierarchical arrangement of nature. This great explanatory theory made necessary a correlative concept, that of the famous "missing link." The missing link proved to be of less value to scientists than it was to, *e.g.*, the American showman P.T. Barnum, who entertained the public with specimens of every conceivable link between groups of animals, including for good measure a mermaid.

One useful result of the belief in unilinear gradation was the search for previously unknown forms that would complete man's knowledge of the chain of being. With this inclusive charter, the scientists of western Europe were ordained to take an inventory of nature and to determine the appropriate place of each newly described form. In practice, it of course supplied the theory for understanding not only man's place in nature but that of all other organisms and was extended to cultures as well. In the absence of genetics and a concept of culture, personality traits as well as skin colour were used for classification. The 19th century was filled with works devoted to the positioning of the anthropoid apes, monkeys, and newly discovered "races." While some persons investigated the anthropoids, others were impressed with the possibility that various groups of unknown aborigines might be the missing link or links. Thus, the "savage Hottentot" or the "stupid native of Nova Zembla" (Novaya Zemlya) was pressed into service to fill the gap between anthropoid apes and man.

Orthogenesis or straight-line evolutionism

After the publication of *Origin of Species* by Charles Darwin in 1859, the idea of evolution became increasingly important, and the chief accomplishment of the period between 1859 and about 1900 lay in the recognition of a considerable time depth for man. The system of arranging the primates remained unilinear, but it was extended far back into the geological past. Both fossil and contemporary races of man were arranged according to their assumed degree of morphological primitiveness. The findings of extinct forms of animals by paleontologists, the excavation of older cultures by archaeologists, and the discovery of Neanderthal skulls all established the antiquity of man as well as of other forms of organic life, and thus the evolutionary sequence became of foremost interest.

By the end of the 19th century several useful classifications of all the races of the world had been completed. Many of the differences among the races were quite well-known. Similarly, differences between man and the anthropoid apes had been inventoried and also between modern man and fossil man. A great deal of attention was paid to the number of differences, and it was assumed that the degree of relatedness was adequately indicated by the degree of morphological similarity.

Genetics and blood typing. The year 1900 may be taken as a turning point in the development of a new conceptual approach, although no single date can fully comprehend a shift in such an exceedingly complex continuum. Two major events that were to have far-reaching consequences occurred: (1) the rediscovery of Gregor Johann Mendel's genetic principles and (2) the discovery of the ABO blood groups. Mendel had formulated the basic principles of heredity in 1865 but these had passed unnoticed. The inheritance of the blood groups was not at first appreciated, but within ten years it became a focus of research and thus a building stone in modern racial studies. Many new concepts came into use with the increasing influence of genetic theory, and attention was immediately directed to the processes of change, those ways in which gene frequencies are modified. At the same time, increasing attention was paid to the significance of traits and to the ways in which they were interrelated. Experimental studies were devised to demonstrate the functional significance of differences in morphology. The concept of straight-line evolution (orthogenesis) was qualified, and examples of reversals or major shifts in the direction of evolution were objectively examined. The concept of the missing link went into discard.

The realization that a change in way of life that introduced a species to a new "adaptive plateau" and brought about many radical changes in its structure obviated the need for a belief in so-called missing links. The techniques of anthropometry, or the physiological measurement of man, were reinforced by many other techniques, such as blood typing—all supplying the physical anthropologists with a much more versatile set of research tools.

Current status of physical anthropology. Much light has been thrown upon man's relation to other primates and upon the nature of the transformation of his skeleton in the course of evolution from early man to modern man, a span of at least 500,000 years. Discovery of the South African man-apes, the first in 1924, and of similar forms in East Africa beginning in 1959 has revealed unanticipated data—that diverse combinations of traits can coexist and that erect posture preceded the great expansion of the brain in human evolution.

Studies of varieties of man and race

Many of the processes responsible for the differentiation of man into different races, although he still remained a single species, *Homo sapiens*, are known: selection, genetic drift, migration, and mutation. Objective methods of isolating various kinds of traits and dealing mathematically with their frequencies, as well as their functional or evolutionary significance, make it possible to understand the composition of human populations and to formulate hypotheses concerning their future. Thus, it is possible for a person to learn about his own genetic constitution with reference to traits ranging from blood types to the fissural patterns of his molar teeth and to know the frequency and distribution of these traits in the various races about the world. He is also able to obtain some estimate of the probability with which his children will inherit these features.

Fields of research. Insight into the field of physical anthropology may be obtained by reviewing the general fields of investigation in which particular researchers are actively engaged. Human ecology, or relationship of the human organism to its environment, bridges the biological and social sciences. It provides insight into the important problems of population size and stability. One very immediate challenge is to determine the varying rate of change that may occur in populations of different sizes.

Human evolution is another area that serves as a focus for research. The essential problems are not only the complete description of fossil forms but the evaluation of the significance of particular traits. New concepts of radiant and parallel evolution have come to the fore. It is clear that the course of human evolution may swing about so that what were formerly considered relatively modern forms may actually have preceded morphologically earlier ones. For example, the more modern kind of Neanderthal man apparently preceded the more primitive or early form. Fossil man of considerable antiquity has been found in Europe, Asia, and Africa. No area in the world is without skeletal remains, so that researchers may everywhere engage in skeletal studies.

Primatology is an area in which man's place in nature is no longer the major point of focus. Researchers are concerned with experimental, medical, and ecological aspects of primates and it is from these studies that much of the functional significance of various bone and muscle complexes has been derived. The primates provide a natural laboratory of many kinds of experiments in physical adaptation to fundamentally different ways of life or adaptive zones.

Studies of primates

Genetics is a fourth major area of research. The study of inherited traits in individuals and the behaviour of the genes responsible for these traits in populations is essential to understanding human variability. Although the blood groups have provided the bulk of the data, many other traits are being analyzed and theories of their inheritance tested. As a result of research in this area it is possible to describe races in terms of gene frequencies and to calculate amounts of race mixture.

Growth studies, both of human beings and other primates, engage the attention of many physical anthropologists in medical and dental schools as well as in independent clinics and universities. Methods of assessing rates

of growth, skeletal age (compared with chronological age), and genetic, endocrinological, and nutritional factors are some of the aspects involved in these studies. The relation between growth and socio-economic status and other cultural features receives considerable attention.

Anthropometry and other measurement. Anthropometry or measurement has been a mainstay of anthropological research for well over a century. More attention is being given to the judicious selection of measurements than to the simple techniques with which calipers are handled. Statistical considerations are especially important in genetic and anthropometric research, and part of the history of statistics is identical with the history of the development of these two research interests. Quite certainly the end results of growth can only be studied by measurement. Newer growth studies have followed children through their morphological and biochemical changes with the aim of discerning why children grow.

Many of the applications of physical anthropology lie in the field of measurement. Thus, the problems involved in providing military and civilian clothing for large numbers of people depend on measurement and statistical treatment. Substantial savings have been made possible by measuring the people of a particular area and adjusting the clothing tariffs to these known distributions of body sizes. There is sufficient variation within most countries so that the geographical variations in size of the body and intermembral proportions are of practical importance. Human constitution is another area of research interest. Several descriptive systems exist for classifying persons who vary from a lateral or relatively squat body build through a relatively muscular to a linear body build. The components of body build, the different tissues and dimensions, were also being studied. The various areas of interest referred to above are not mutually exclusive; ideally they should all form part of the training of researchers in the origin and evolution of man.

BIBLIOGRAPHY. Histories of anthropological science include: T.K. PENNIMAN, *A Hundred Years of Anthropology*, 3rd ed. rev. (1965), which covers all of anthropology; and P. MERCIER, *Histoire de l'anthropologie* (1966), which covers only cultural anthropology. The principal textbooks are: M.J. HERSKOVITS, *Man and His Works* (1948); F.M. KEESING, *Cultural Anthropology: The Science of Custom* (1958); and J. POIRIER (ed.), *Ethnologie générale* (1968).

On physical anthropology only a few works of synthesis can be cited: P. TOPINARD, *Eléments d'anthropologie générale* (1885), the oldest of the treatises; A.E. MOURANT, *The Distribution of the Human Blood Groups* (1954); W.C. BOYD, *Genetics and the Races of Man: An Introduction to Modern Physical Anthropology* (1950); R. KHERUMIAN, *Génétique et anthropologie des groupes sanguins* (1951); G. HEBERER, G. KURTH, and I. SCHWIDETZKY-ROESING, *Anthropologie* (1959; Eng. trans., *Anthropology A to Z*, 1963); ASHLEY MONTAGU, *An Introduction to Physical Anthropology*, 3rd ed. (1960); J. COMAS, *Manual of Physical Anthropology*, rev. ed. (1960); C.S. COON, *The Living Races of Man* (1965); R. MARTIN and K. SALLER, *Lehrbuch der Anthropologie*, 4 vol. (1957–66); and H.V. VALLOIS, *Les Races humaines*, 7th ed. (1967).

In cultural anthropology—aside from two works by the "fathers" of the discipline, L.H. MORGAN, *Ancient Society* (1877); and E.B. TYLOR, *Anthropology: An Introduction to the Study of Man and Civilisation* (1881)—some of the classic general works might be mentioned: FRANZ BOAS, *The Mind of Primitive Man* (1911), and *Race, Language and Culture* (1940); BRONISLAW MALINOWSKI, *A Scientific Theory of Culture, and Other Essays* (1944); A.R. RADCLIFFE-BROWN, *Structure and Function in Primitive Society* (1952); A.L. KROEBER (ed.), *Anthropology Today* (1953); CLAUDE LEVI-STRAUSS, *Anthropologie structurale* (1958; Eng. trans., 1963); G. BALANDIER, *Anthropologie politique* (1967; Eng. trans., 1971); and M. MAUSS, *Oeuvres* (1968).

Studies of individual peoples that have become classics include: W.H.R. RIVERS, *The Todas* (1906); M. GRANET, *Fêtes et chansons anciennes de la Chine* (1919; Eng. trans., *Festivals and Songs of Ancient China*, 1932); BRONISLAW MALINOWSKI, *The Argonauts of the Western Pacific* (1922), and *Coral Gardens and Their Magic* (1935); A.R. RADCLIFFE-BROWN, *The Andaman Islanders* (1922); MARGARET MEAD, *Coming of Age in Samoa* (1928); P. SCHEBESTA, *Bambuti, die Zwerge vom Kongo* (1932); FRANZ BOAS, *The Religion of the Kwakiutl Indians* (1930); R.F. FORTUNE, *Sorcerers of Dobu: The Social Anthropology of the Dobu Islanders of the Western Pacific* (1932); R.W. FIRTH, *We, the Tikopia: A Sociological Study of Kinship in Primitive Polynesia* (1936); M. GRIAULE, *Masques Dogons* (1938); M.J. HERSKOVITS, *Dahomey: An Ancient West African Kingdom* (1938); M. LEENHARDT, *Gens de la grande terre* (1937); E.E. EVANS-PRITCHARD, *The Nuer* (1940); and E.R. LEACH, *Political Systems of Highland Burma: A Study of Kachin Social Structure* (1954). (Ed.)

Anthropology, Philosophical

The term anthropology was first used in the philosophical faculties of the German universities at the end of the 16th century to refer to the systematic study of man as a physical and moral being. In the 18th century there was a tendency to split the study into physical anthropology, which investigated man as a natural species, and cultural anthropology, which considered man as a social being individualized by historical and geographical factors. This split arose in response to a real need for a division of intellectual work, a need that continued to increase in recent times. As a result, the realm of anthropology has broken down into disciplines that are more and more specialized and concerned with increasingly restricted aspects of man.

An illustration of this proliferation of specialized approaches to the study of man is seen in a work entitled *Anthropology Today*, edited by an American anthropologist, Alfred L. Kroeber, in 1953. The work was an attempt to state the position of the discipline in the middle of the 20th century. It gathered the contributions of 100 scholars, each of whom tried to explain what could be discovered about man in the context of his distinct, narrow perspective. Contemporary anthropology thus appears to be a jumble of parts on the basis of which it would be vain to attempt to define the whole.

Philosophical anthropology arose as a remedy for this dissolution of the study of man. It attempts to regroup the specialized disciplines within a common purpose; it seeks to give a sense of order to the investigation of man, in the absence of which the human sciences neglect the essence, or humanity, of man. The order sought by philosophical anthropology is not a purely theoretical order, limited to the realm of knowledge; it is an order with a practical validity of perhaps greater impact than the conclusions of the specialized human sciences.

Man in the 20th century seems to have lost his sense and respect for what is basically and uniquely human. This phenomenon is attested to not only by the paroxysms of Hitlerism and Stalinism but also by many aspects of daily life in the modern world. The most dangerous threats are those phenomena that do not necessarily appear as such; for example, the results of the uncontrolled experimentation with economic and technological factors. The German philosopher Immanuel Kant declared that the human person should be considered an end in itself; increasingly, however, the human person is treated as a means to be used in the service of obscure and contradictory interests that disjoint the image of man. Philosophical anthropology, as a reflective process that has its goal and point of departure in the human condition, has given itself the task of safeguarding humanity from such abuse.

HISTORICAL BACKGROUND OF PHILOSOPHICAL ANTHROPOLOGY AS A DISCIPLINE

The modern discipline of philosophical anthropology had its origin in the 1920s. The late appearance of this discipline is somewhat paradoxical, because Western philosophy has constantly been concerned with man. During the middle of the 5th century BC, the Sophist Protagoras articulated the principal formula of philosophical anthropology: "Man is the measure of all things"; not long afterward, Socrates adopted for himself and incorporated into his teaching the saying borrowed from the oracle of Delphi: "Know thyself." If anthropology is taken to mean the empirical investigation of man, then this discipline has several magnificent monuments in the ancient civilizations of the West: for example, the medical writings of Hippocrates, the biological and political writings of Aristotle, and the works of the Greek historians He-

rodotus and Thucydides. Thus, philosophical anthropology seems to have existed long before it was recognized as an independent discipline.

The themes of man, world, and God

Man, the world, and God have constituted the important themes of Western thought from the beginning of its recorded history; the relative significance of these three themes, however, has varied from one epoch to another. Western thought has laid greater stress on the existence of the individual human being than have the great speculative systems of the East; for example, in the Hindu Brahmanism personal destiny dissolves in the All. Nevertheless, recognition of the uniqueness of the human form did not guarantee that man would receive a privileged status in Western thought. The appearance of philosophical anthropology, as a modern discipline, presupposed a certain autonomy for the human condition in relation to the world and to God; it demanded that all deterministic notions of man be eliminated.

Ancient periods. Ancient thought was characterized by its emphasis on the notion of cosmos, or the order of the universe; an understanding of the cosmos provided the key to universal intelligibility. Cosmology had a priority over theology and anthropology. The relative movements, connections, and oppositions of the planetary gods depended upon the structure and laws of the extraterrestrial world, which in turn played a great role in the evolution of events in this world; the celestial macrocosm imposed its law on the human microcosm. The disciplines of astrology and alchemy were practical consequences of this subordination of man to transcendent influences. According to the teachings of Pythagoras, Plato, and the Neoplatonists, the human realm was inferior to the transcendent in dignity; its truth was degraded and obscured by bodily existence. As a result, man was not self-sufficient; he did not form an autonomous region of being because, in the final analysis, his destiny was governed by extrinsic influences.

Medieval period. Medieval culture preserved the ancient structure of the cosmos. There was, however, a change. For the Greeks, the cosmic order wielded authority over the gods themselves, but Christian revelation brought forth the idea of an almighty God who was Creator of the world and of men. Theology superseded cosmology insofar as it endowed cosmology with religious meaning and import. The terrestrial realm was situated between heaven and hell, the positive and negative poles of each man's destiny. The human being did not exist in and for himself; his status was that of a creature of God whose thought and actions were to be ordered in such a way as to reflect the supremacy of religious values. The Middle Ages, while they had no discipline of anthropology in the strict sense, did have a section of theology that concerned the fallen nature of man and the possibility of salvation, which was offered to him by divine grace in accord with the redemption effected by Christ.

Renaissance. The rise of philosophical anthropology can be located at the moment man was removed from the ascendancy exercised over him by either cosmic discipline or divine authority. Man did not thereby cease to view himself within the framework of the world; he did not reject the presence of God; he did, however, disengage himself sufficiently from these bonds to become a centre of interest in his own eyes. It was in the cultural context of the Renaissance that the centre of gravity of reflective thought descended from heaven to earth. In ancient literature the educated people of the West rediscovered a clear conscience instead of the guilty conscience of Christianity; at the same time, the great inventions and discoveries manifested the right of a new initiative for a humanity that seemed to be escaping both cosmic and Christian predestination. Man could take pride in his accomplishments and regard himself with admiration. The themes of the dignity and excellence of man were particularly dear to the Italian Humanists and to their imitators. Renaissance Humanism was not yet philosophical anthropology; it remained fascinated by the theme of man as the image of God.

The 16th and 17th centuries. The thought of Michel de Montaigne, the 16th-century skeptical author of the *Essais*, represents the first attempt at anthropological reflection—that is, reflection that centred on man, exploring his different aspects in a spirit of empirical investigation, freed from all ties to dogma. At the end of the 16th century and the beginning of the 17th century, Francis Bacon, the precursor of modern Empiricism, set before philosophers and other scholars the example of the hardy navigators, the discoverers of the New World. In the generation before Bacon, however, Montaigne was also the discoverer of a new world, one in which man would reflect upon himself and strive to define the authentic nature of *homo humanus* ("human man").

The thought of Montaigne

The anthropology of Montaigne began with a turning in upon himself; it gave priority to that reality which was within. Montaigne, however, was also the witness of a renewal of knowledge brought about by numerous discoveries that made the horizons of the traditional universe expand greatly. For him, self-awareness already reflected an awareness of the surrounding world; it wondered about the "savages" of America and about the cannibals that were so different from him and yet so near; it compared the intelligence of man with that of beasts and accepted the idea of a relationship between animal existence and human existence. Objective facts were gathered as a result of an uneasiness about judging values. Montaigne's anthropology was an attempt to locate humanity in the midst of the totality of things; the being of man became a circle, whose circumference extended everywhere and whose centre was nowhere.

The classical philosophy of the 17th century abandoned the way opened up by Montaigne; "wavering and varying" man seemed to be an insufficient support for the eternal truths whose primacy philosophy affirmed. Blaise Pascal, a French scientist and religious writer, described man as neither angel nor beast and said that it was therefore fruitless to try to understand him as a fallen god or as an animal raised to a higher state. Despite this attempt to place man in his proper perspective, however, Pascal denounced the ego as loathsome. At the same time René Descartes depended upon the individual consciousness only as the quickest means to divine certitudes. For the Dutch Jewish Rationalist Benedict de Spinoza and the Cartesian philosopher Nicolas Malebranche, the personal self was devoid of both reality and value. Philosophical anthropology could not assert itself until after the demise of the great Rationalistic schemes of reality.

In his *Essay Concerning Human Understanding* in the late 17th century, the British philosopher John Locke imposed a new style of reflection on European thought, one that made use of the empirical approach and demanded an awareness of the limitations imposed upon human knowledge as such. In anticipation of the formula of Immanuel Kant, the great 18th-century German philosopher, Locke stated that man should recognize himself as an "intellectual native of this world"; that man is not an absolute subject, but an incarnate consciousness with all the inadequacies and limitations of human reality. Locke's philosophy—as well as that of the Earl of Shaftesbury and David Hume in Britain and of Étienne Bonnot de Condillac in France—can be considered philosophical anthropology; the *Encyclopédie* of Jean d'Alembert and Denis Diderot also had an anthropological orientation.

The 18th-century Enlightenment. In the 18th-century Enlightenment, the distinctive character of philosophical anthropology became evident; it was a branch of philosophy, closely associated with the human sciences, attempting to relate the accumulation of new objective data about man with the values that constitute the human consciousness. Traditional philosophy had been interested only in rational essences and refused to include within its scope the human domain in all its diverse, concrete aspects. The French naturalist Georges Buffon devoted two volumes of his 44-volume *Histoire naturelle* (1749–1804) to man, treating him as a zoological species. Since that time the science of anthropology has continued to multiply and diversify its approaches to man; it has also maintained the necessity of converting the data of the other sciences into its approach to man.

Association of philosophical anthropology with studies of human sciences

The Enlightenment thinkers rediscovered that man is not only a natural being but also a cultural being; the traditional philosophy of nature was complemented by a "philosophy of history" (an expression coined by Voltaire in 1765). In 1725, Giambattista Vico, a Neapolitan thinker known for making cultural history into a "new science," affirmed in his *Scienza nuova* that "the social world is certainly the work of men" and endeavoured to elucidate the norms that govern man in his social environment. Following him, such philosophers as Montesquieu, Voltaire, Johann Herder, Kant, and G.W.F. Hegel developed a historical, juridical, and political anthropology.

The 19th century. The immense progress of the social sciences was one of the significant aspects of the history of thought in the 19th century. Nevertheless, the specialists in these sciences—and in all the human sciences—became more restricted and specialized in their outlook. They made excessive pretensions for their own sciences but were prisoners of their specialization, unable to view man in his totality or in his essence. The specialist in the religious sciences saw everywhere only ritual behaviour; the specialist in economics reduced all individual or social life to preoccupations with material self-interest. Conceptual totalitarianism arose—Freud analyzed man only in terms of complexes, censorship, and repression; Marx made economic production the sole and universal key to intelligibility.

Philosophical anthropology developed as a remedy for this epistemological anarchy by maintaining, in contrast to the empirical sciences, the primacy of values in an understanding of man. Scientific endeavour, it held, is indispensable, but philosophical reflection has its place prior to and subsequent to the work of science; it must order the incoherent mass of results obtained from the sciences to conform to the true essence of man.

Prior to the advent of the human sciences, ancient philosophy was given the name metaphysics—that is, the study of the principles of being beyond nature or the natural sciences. This new branch of philosophy can be recognized as a "metahumanity"—that is, a reclaiming of human equilibrium in the rising tide of sciences and technologies.

MAN'S STATUS AS EXISTING IN THE UNIVERSE

Resistance to the concept of man as a natural being. The idea that man is a natural being and that it is necessary to attribute to him the status of a living thing among other living things seems self-evident today. Such an affirmation was present in the biological writings of Aristotle, but it was long eclipsed from Western thought. The evolutionary theories of Darwin in the late 19th century were strongly resisted by a public that considered them to be a humiliation and degradation.

Judeo-Christian concept of man. A major obstacle to man's appreciation of his natural being was the Judeo-Christian concept of transcendent meanings of man: the creation story in the book of Genesis made man a creature among other creatures, but not a creature like other creatures; man was born as the final act of divine initiative and given responsibility for the Garden of Eden; he was the object of creation and the beneficiary of a direct relationship with the Creator. The Fall and redemption, the categories of sin and grace, thus concern only the descendants of Adam, who have been given a nature radically different from that of the animals and plants, which are at the disposition of man in view of his unique destiny. Because of his unique vocation, man is considered to be a transnatural, or supernatural, being. He alone participates after this world in an eternal life that is far more important than the natural and temporal life, and he alone will survive once the carnal shell, in which the soul is imprisoned for a time, has been destroyed.

Desacralization and dissection of human bodies

Throughout the Middle Ages and even afterward, the dissection of corpses was opposed by religious prohibitions and was punished as a crime. Anatomy and medicine could not become empirical disciplines until this taboo was suppressed. The symbolic date in this matter was 1543, when a Belgian physician, Andreas Vesalius, a professor at the University of Padua, published a precise atlas of the human body, entitled *De humani corporis fabrica* ("On the Fabric of the Human Body"). In the intellectual order, the revelation of the makeup of the human body was analogous to the great land discoveries that were expanding man's horizons; the desacralized body was opened to profane investigation. The naturalization of the human organism was the first step toward a science of the natural man, toward a natural history of the human species.

Intellectualist concept of man. To establish philosophical anthropology, it was not enough to recognize that man's body is a body among other bodies. It was still possible to maintain that man's body is not man, that the essential human reality has nothing in common with the material organism with which it finds itself linked for a time. Descartes proclaimed, "I am not this collection of members that is called the human body"; and Pascal affirmed,

> I can easily conceive of a man without hands, feet, head (for it is only experience that teaches us that the head is more necessary than the feet). But I cannot conceive of man without thought.

Philosophers in the 17th century, who were predominantly Rationalists and spiritualists, denied the anthropological perspective; they sought refuge in a dualism that made the organism something different from the spirit, considering it the inferior partner with little value. For Descartes, the body is only a more or less crude machine, which contains nothing that can provide an explanation for its association with an immaterial and invisible spirit.

Intellectualistic philosophy, which developed out of Descartes's dualism, maintains the isolation of a rational faculty in man that refuses to let itself be caught in the trap of bodily existence. Philosophical anthropology, however, stresses incarnation; in opposition to Cartesian dualism, which separates the body and the spirit, it affirms the unity and identity of these two aspects of man. Thought is intimately related to the body; experience shows no such thing as disincarnate thought. For this reason, anthropology adopts a monistic conception of the human being. This is not an ontological monism (the doctrine that there is only one kind of substance, or ultimate reality), as in Materialism, which denies the spirit, or as in spiritualism, which denies the body. The anthropological monism is epistemological (concerned with knowledge) rather than ontological (concerned with Being); it forms the conceptual basis for a study that endeavours to describe the reality of man in his concrete presence. It follows the approach of "experimental," or inductive, philosophy, which, developing slowly out of the work of the Italian scientist Galileo and Francis Bacon, had little by little freed itself of the Cartesian mechanistic conception of science. After having viewed himself with the eye of the theologian, who scrutinizes the work of grace, and with the eye of the artist, who is struck by formal or structural beauty, man finally began to consider himself as a being of nature, who is as much a valid subject of scientific investigation as other living things.

Man as a natural being. In the biblical account, Noah, the builder of the ark, is also the leader of the menagerie entrusted to him; it would have been paradoxical to include him in the list of beasts carried on board the ark. The question of man's relationship to the beasts, however, eventually had to be determined. A decisive step in the origin of physical anthropology was the encounter of Western man with the great anthropoid apes of Africa (Angola) and of Asia (the Sunda Islands) at the beginning of the 16th century. Until then Europe had known only the smaller sized monkeys, which were too far removed from the human species to present any confusion. The discovery of the chimpanzee and the orangutan (the name means "man from the woods" in the Malay language), however, raised the question of the confines of humanity because these animals—by their size, their posture, their walk—could be confused with man. Questions arose whether the anthropoid, who resembles man, is an animal or a man and why it should be considered an

Discovery of the anthropoid apes

ape and not a man. Following the development of experimental philosophy, these questions could be posed in empirical terms and investigated scientifically, without reference to the teachings of revelation.

The study of the natural history of man began, in modern times, when the first specimens of anthropoid apes were introduced into Europe; they aroused popular curiosity and were submitted to the study of scholars. The first precise description of an anthropoid was given in 1641 by the Dutch physician Nicolaes Tulp, the master in Rembrandt's painting "Anatomy Lesson of Dr. Nicolaes Tulp." The ape that Tulp studied and represented in a drawing was numbered among the animals, although the threshold that separated it from man was not specified.

Work of Edward Tyson. The separation between man and ape was delineated at the end of the 17th century by the English physician Edward Tyson, a leading member of the Royal Society of London, one of the oldest and most prestigious scientific societies in Europe. Tyson had the opportunity to study the remains of a young chimpanzee from Angola that had died in London several months after its arrival. His research resulted in a monumental work that was officially published by the Royal Society under the title *Orang-Outang, sive Homo Sylvestris: Or, the Anatomy of a Pygmie Compared with That of a Monkey, an Ape, and a Man* (1699). This treatise, one of the landmarks in both anthropology and comparative anatomy, is remarkable for the empirical approach used in the investigation. Tyson examined both the exterior and interior of the chimpanzee with the same critical and analytical eye that Vesalius had used in drawing up his illustrations of the human body. His precise measurements, his complete exploration of the external and internal structures of the animal, and his minutely detailed sketches permitted him to pose what is perhaps the central problem of physical anthropology: whether it is possible to find among the anatomical or physiological characteristics of the ape the justification for asserting a radical difference between the ape and man, all the similarities notwithstanding.

Problem of the difference between the ape and man

The problem was clearly set out. Tyson did not refer to a plan of God, according to which man received an immortal soul that was denied to the animals. He analyzed in the greatest detail the similarities and dissimilarities between a chimpanzee and a man. A table that summarized his work indicated 48 physiological characteristics that made the ape similar to man and 34 characteristics that, on the other hand, connected it more closely to animals. This table, which was quite accurate, was used half a century later by Buffon in his *Histoire naturelle*. Most of the distinctive marks of the human species were already brought to light in Tyson's work. He emphasized the fact that the ape is a quadrumane (having four hands) rather than a quadruped (having four feet); unlike the human foot, its foot has an opposable big toe (*i.e.*, like the thumb). The arrangement of the internal organs allows the straight posture that makes the ape similar to man in this respect. Tyson included the form and mass of the brain among the similarities, as well as the speech apparatus, and declared that he was unable to determine from a strictly anatomical point of view why the ape is incapable of thinking and speaking. His conclusion was that anatomy is not everything and that other factors are involved:

> Since therefore in all respects the brain of our Pygmie [chimpanzee] does so exactly resemble Man's . . . there is no reason to think that agents do perform such and such actions because they are found with organs proper thereunto: for then our Pygmie might be really a man. . . .

Tyson's exploration of the boundaries between animality and humanity left the central question unanswered; the solution had to be sought in a dialogue between the physical order and the metaphysical. The key point of Tyson's work, however, is that the science of man as a natural being henceforth had an empirical terminology and a methodology that was independent of mythico-theological presuppositions. The idea of a continuity between animality and humanity definitively outweighed that of a radical break between them; man re-entered the ranks of natural living beings.

Theory of the "great chain of being." This idea of the continuity between animal and man developed with the popularization of the theme, Hellenic in origin, of the "great chain of being." This explanatory theory set up a hierarchical order among natural beings, from the most simple to the most complex: from mineral to vegetable, then to animal and to man, and finally to divine beings, superior to man. This arrangement of an ascending gradation of the forms of life was the basis for the comparative analysis of structures; it had been found as early as in the biological writings of Aristotle, who established similarities of structure common to the various degrees of natural history. Christianity had incorporated the concept of the chain of being, which was coordinated without great difficulty with the idea of the days of creation. In the book of Genesis God performs his creative act in a gradation from lower to superior beings, with the appearance of man crowning the work of the six days. Theologians completed the ascending series by placing above man the celestial hierarchies of angels and transcendent powers, the intermediate degrees between man and God.

The concept of the chain of being, which was generally accepted until the 18th century, encouraged scholars to look for intermediary forms—"missing links"—between the various degrees on the scale; for example, zoophytes (invertebrate animals resembling plants, such as sponges) formed the link between the vegetable order and the animal order. For Tyson, the Pygmy was the missing link between animality and humanity. Richard Blackmore, a British physician who was influenced by Tyson's views, writing in 1714, noted that man, being half body and half spirit, ought to be considered as the equator "that divides in the middle the whole creation." It was now recognized that the human species was located at a critical point of variation along the chain that made anthropology a special type of study.

The "missing links"

Work of Linnaeus. Linnaeus (Carl von Linné), an 18th-century Swedish naturalist, included man in his classification of living species. He described him more precisely than ever before—in the first edition of his fundamental work, the *Systema Naturae*, in 1735. The tenth edition of the *Systema* placed man among the mammals in the order of primates, alongside the apes and the bat. Linnaeus wrote,

> I cannot discover the difference between man and the orangutan, although all of my attention was brought to bear on this point, except for laying hold of some uncertain characteristics.

The distinctive characteristic of *Homo sapiens* (man) is his use of reason, which does not depend on any physiological characteristics; "it is diurnal, or awake by day; it varies according to education and according to the effects of the climate"; that is, the variations within the human genus are of a cultural nature.

Work of Buffon and Blumenbach. The *Systema Naturae* stopped at man; it did not encompass the angels. For Linnaeus, natural history does not have access to the supernatural, and he refused to venture an opinion on higher beings. Buffon's *Histoire naturelle*, which appeared beginning in 1749, adopted the same attitude: it devoted two volumes to an objective study of the human species through the most diverse of its aspects. Although man was the object of a more detailed study than the dog or the horse, his place was in the natural kingdom alongside the dog and the horse. Thus, the naturalization of the chain of being was completed: the mythical–theological scheme was replaced by the logical ordering of a classification.

Buffon's work, the first truly systematic exposition of the question of man's place, was continued by the German scholar Johann Friedrich Blumenbach, recognized as the father of physical anthropology, in his *De Generis Humani Varietate Nativa* ("On the Native Variations of the Human Race"). In this work, which was written in 1775 when he was only 23 years old, Blumenbach observed that "the innumerable varieties of mankind run into each other by insensible degrees." Following the

early Greek philosophers Anaxagoras and Aristotle, he insisted that man alone truly has hands that make him capable of fashioning tools, as Benjamin Franklin, the American statesman and scientist, had said. He observed that only man speaks, and only he is capable of laughing and crying.

With the work of Blumenbach, physical anthropology made a notable contribution to philosophical reflection; more theoretical considerations about man were now possible. Man's erect posture, which freed the anterior limbs, enabled the development of the hand, the source of technological initiatives. At the same time, this upright posture gave man an enlarged field of vision, which in turn resulted in an increase of action—a quadruped cannot subdue an area it cannot view. Although man is weak, naked, and deprived of natural defenses, his physical structure is the basis of his autonomy; and his intelligence is his most efficacious weapon in assuring his survival.

Man as a product of evolution. The development of a natural history of man was thus made possible by the naturalization of the human being. This natural history was originally an investigation of existing animals, but it had to be broadened to include the past so that living things could be more accurately classified. The notion that species were stable, which had been accepted as a dogma in accord with religious teaching, contradicted the evidence of experience. A static classification could not explain the dynamic relations between isolated species. Thus, a primitive time line of natural history was developed. The relationship of families led to the idea of a filiation between them according to an order of succession. The interpretation of fossils, which gave scholars views of the history of life in depth, aroused impassioned debates. From them have arisen the concepts of mutation (the process by which the hereditary constitution of a cell is altered), transformism (the theory that one species is changed into another), and evolution. These concepts were already being formulated in the 18th century, were later reviewed and corrected by the French naturalist Lamarck and by Darwin, and became the guidelines for the scientific and philosophical investigation of the 19th century.

Addition of the element of time

When the dimension of time was added to natural history, the science of anthropogenesis (the study of the origin and development of man) required a place within the context of cosmogenesis (the study of the origin and development of the universe) and in the broad perspective of paleontology (the study of past geological periods) and of prehistory. There was resistance, however. In the first half of the 19th century, Jacques Boucher de Perthes, a French antiquarian, discovered some handworked flints and other stone tools in northern France and identified them as fossils of an early culture. It took more than 20 years for his conclusions to win acceptance among his contemporaries. In fact, his views were not officially recognized and accepted until 1859, the year that Darwin published his *Origin of Species* and Paul Broca, a French surgeon and anthropologist, founded at Paris the Société d'Anthropologie; this was just three years after the discovery of the human fossils of Neanderthal man.

Philosophical anthropology has to take into account new facts about the past development of man whenever they are discovered. The *homo perennis* ("perennial man") of traditional philosophy, who was always and will be forever the same, is an illusion. The study of prehistory has enabled man to follow the trials and errors of nature during the slow genesis of the human race. The maturation of man's physical shape was correlative with the development of his mental activity and with the emergence of intelligence. If the human race has a history of development, however, then it must be recognized that man might continue to develop in the future. Philosophical anthropology must also seek the future form of the human species. If it is not necessary to accept the venturesome speculations of a Teilhard de Chardin, a French Jesuit paleontologist who viewed all things as evolving toward a greater unity and perfection, it is probably not possible to reject entirely the idea of a further development of the human species.

MAN'S STATUS AS A SOCIOCULTURAL BEING

The thought of Renan

Ernest Renan, a French historian and philologist, claimed in the middle of the 19th century to have elaborated an "embryogeny of the human spirit" in which the progressive development of culture in the universe was recapitulated. Renan formulated his thought in his work *L'Avenir de la science* (*The Future of Science*):

> One who has learned nothing of the history of our globe and the beings who have successively peopled it from geology, nothing of the laws of life from physiology, nothing of the laws of the forms of being and the general plan of animate nature from zoology and botany, nothing of the structure of the universe from astronomy, nothing of the science of humanity in its becoming from ethnography and history—can such a one boast of knowing the law of things, as I said, of knowing man whom he studies only abstractly and in his individual manifestations?

This text, written before Darwin's *Origin of Species*, expressed a growing attitude that the development of the human sciences refuted the illusion of an eternal subject, as man had been conceived by the Rationalist and Idealist philosophers. For Descartes, the "I" of his "I think, therefore, I am" maintained a direct relationship with itself, as the subject and object of the reflective process, and with God, independent of any relationship with the natural or historical world; the consciousness of the thinking being was an absolute, the product of a clear apprehension that secured in each individual the same consciousness, irrespective of the particular circumstances of space and time. Classical philosophy, the abstract science of essences, was dead; it gave way to a philosophical anthropology—that is, to a reflection on the concrete conditions of human existence.

Man as distinguished from animals. The philosophical anthropology of Renan encompassed both the natural sciences—such as geology, physiology, zoology—and sciences of a different order—such as ethnography (*i.e.*, descriptive anthropology) and history. In the 18th century, when Linnaeus and Buffon initiated the natural history of the human species, they took into account distinctive characteristics that were not of a physiological order. It was not possible to describe *homo sapiens* in the objective vocabulary that was sufficient to describe the dog, the donkey, or even the chimpanzee.

Work of Linnaeus. In the later editions of *Systema Naturae*, Linnaeus presented a summary of the diverse varieties of the human species. The Asian, for example, is "yellowish, melancholy, endowed with black hair and brown eyes," and has a character that is "severe, conceited, and stingy. He puts on loose clothing. He is governed by opinion." The African is recognizable by the colour of his skin, by his kinky hair, and by the structure of his face. "He is sly, lazy, and neglectful. He rubs his body with oil or grease. He is governed by the arbitrary will of his masters." As for the white European, "he is changeable, clever, and inventive. He puts on tight clothing. He is governed by laws."

This anthropological sketch with its naïve assertions may reveal that the scientific study of man cannot be like that of other animals. Linnaeus attempted a rudimentary mixture of physiological characteristics and psychosociological traits. The mentality, the clothes, the political order were taken into account; physical anthropology was not distinguished from cultural anthropology; the variations among men were recognized as variations of culture as well as of nature.

Work of Buffon. When the naturalist characterizes the diverse variations of a certain animal species—whether it be fish, mammals, or birds—he does not take any cultural traits into account because there are none. The cultural factor is a distinctive characteristic of the human species, one in which the distinction between the animal kingdom, in the restricted sense of the term, and the human kingdom is found. Buffon, whose anthropology was more comprehensive than that of Linnaeus, stated clearly the separation between man and the higher apes,

which for some were intermediaries between animality and humanity.

> The ape . . . is, in fact, only a mere animal wearing on its exterior a mask of a human figure, but in the interior it is deprived of thought and of all that makes man; an animal . . . still essentially different from man by nature, by temperament, and also by the length of time necessary for its rearing, for gestation, for the growth of the body, for the length of its life, that is, by all the real habits that constitute what one calls the nature of a particular being.

Buffon's analysis emphasized the close mutual connection of physical, mental, and cultural factors on the human level. For him, the most significant phenomenon in man's development is that of articulated language; the rise of the human intelligence is tied to the advent of language. Buffon stressed, however, that the presence or absence of speech does not seem to be tied to the existence of an organic speech apparatus.

> It is true that it is not for lack of organs that animals do not speak; several species that can be taught to pronounce words and even to repeat rather long phrases are known . . . but no one has ever succeeded in making them give birth to the idea that these words express.

This is why "the most stupid of men can command the most intelligent of animals . . . because he has a reasoned plan, an order of actions, and a series of means by which he can force the animal to obey him." Therefore, according to Buffon, there exists "an infinite distance" between animal and man, and the essential cause for this is that "man is a being with reason, and the animal is one without reason."

The theme of the continuity or of the discontinuity between animality and humanity, together with the opposition between nature and culture, was one of the greatly disputed issues during the Enlightenment; it has remained a fundamental preoccupation of anthropology. For Buffon, the variations among men had to be explained either by the influence of climate or by the mode of living; the opposition between these two notions reflected the opposition between deterministic understandings of man based upon natural or cultural explanations. Buffon devoted more than 200 pages of his natural history of man to the study of the "Variétés de l'espèce humaine" ("Varieties of the Human Species"); this treatise, updated from edition to edition as a result of new information, formed the outline for one of the first attempts at a comparative ethnology (originally the science treating the distribution of mankind into races). He made precise studies of the dwarfs of Madagascar, of the Patagonians in the southern part of South America, the Hottentot in South West Africa, the American Indians, and the natives of several of the Pacific islands recently explored by the English navigator James Cook and the French navigator Louis-Antoine de Bougainville.

Man as distinct from other men. The nascent cultural anthropology had been given momentum by the enlarging of the human horizon begun in the 15th century when Spanish and Portuguese adventurers had undertaken the systematic investigation of the Earth. Their reasoned transgression of the limits of the Old World, which had been centred in the Mediterranean Basin, was itself an experiment in thought. The voyage around the Earth was a voyage around humanity that destroyed and renewed previously accepted truths. Cultural anthropology is the product of the confrontation between the classical values of the West and the opposing values and customs of the newly discovered civilizations.

Discovery of different civilizations

If physical anthropology was born out of Western man's encounter with the anthropoid apes, cultural anthropology was made necessary by his encounter with people in the rest of the world. The "savage" appeared to manifest a style of humanity that was a contradiction of the certainties, the beliefs, that had sustained the Europeans for centuries. The shock was such that the naked Indian and the cannibal were at first assumed not to belong to the human race; this approach enabled Europeans to avoid the problem. This facile solution was, however, rejected by Pope Paul III in 1537 in his bull, or document, *Sublimis Deus* ("The Transcendent God"). According to Paul, the Indian savages were men; they had souls and, as such, could be initiated into the Christian religion. This was a key decision in the history of anthropology, for it recognized that there is an intrinsic diversity among men, that values must be accepted as relative, and that there must be a search for new generalizations about man. Little by little, it became clear that the traditional thought of the West—including metaphysics, moral philosophy, psychology, and even religion—could no longer claim to have an absolute validity. The white, adult, "civilized," and Christian European could no longer be taken as the universal man. Cultural anthropology thus began with an unprecedented humiliation for Western man and his self-understanding. The confrontation of cultures brought with it the necessity of establishing a broader culture in which confrontation could take place. Anthropology appeared as the process necessary to construct such a culture.

Man's discovery of the anthropoid apes gave rise to the natural history of man, and, while locating his place among the animals, it also made necessary a reflection on the distance between animal and man. The theory of evolution provided the conception of a continuous development of the animal condition and the human condition, the stages of which were studied in paleontology. The encounter of Western man with the "savages" gave evidence of the distance between man and man. The idea quickly developed that the geographic distance between the inhabitant of a European city and the man of the forest or of the swamp corresponded in reality to a chronological arrangement. The "savage" was identified as a "primitive," as a "man of nature," who remained close to an initial state from which a privileged part of humanity had been able to remove itself by a continued effort at community and individual advancement. A study of the history of man endeavoured to bring to light the successive stages that the human species had passed through along the way to the present civilized societies. The themes of "civilization" and "progress" were among the principal preoccupations of the Enlightenment.

Man as a product of culture. In its early days, it was hoped that anthropology would order nature and culture within the same perspective, with nature designating the point of departure, the original stage from which civilization progressively moved. The "savage" was fit into this perspective as an intermediate form, a transition figure that bridged the gap between ape and man; the zero degree of humanity was located at the limit or point at which the hominization (forming of man) of the anthropoid was accomplished. This hope was a delusion, for it soon appeared that the "absolutely savage savage" conceived by Buffon was only a fanciful image. The primitives—whatever they were—were never truly primitive; they were always deeply affected by the particular cultural system to which they belonged. For this reason Maurice Merleau-Ponty, a 20th-century French Phenomenologist, concluded that "it is the nature of man not to have a nature."

The view of Merleau-Ponty

By his statement Merleau-Ponty meant that cultural phenomena are inherent in humanity; they are everywhere and always linked to the physical presence of the human being. Deprived of all consciousness, the body of man is not man; it is only a cadaver. Philosophical anthropology has had to accept the fact that the bodily incarnation of the living human is necessarily joined to a social incarnation: this double incarnation characterizes the human species and accounts for its uniqueness.

In many of the locations in which prehistoric fossils have been discovered, the remains of humans have been found with the bones of wild or domestic animals. It has been determined that the prehistoric dog was similar to contemporary dogs; they lived in society with their masters just as their descendants do. It has also been determined, however, that man has been considerably modified and that this modification is reflected less in his physical characteristics than by cultural signs. Scholars recognize successive ages of man (such as Paleolithic, Neolithic, and protohistoric) as a result of the study of materials discovered in the vicinity of skeletons: weap-

ons, tools, utensils of every kind, and decorative items are evidence that man's way of living has not stopped changing, even though there has been no noteworthy change in the way of life of the animals.

Clearly, the first men lived a difficult existence. They took shelter in grottoes and caves and lived a life similar to that of the animals; but they learned to use fire, to make tools for work, and to provide rudimentary defenses. Eventually man began to cultivate the soil and to raise animals. The invention of writing about the 3rd millennium BC, when the basic social structure was shifting from the clan through regional sovereignties to the empire, marked the threshold of the historical era, in which the human world took on a rational form.

Dependence of the newly born human on society

The human being that arrives in the world at the end of the ninth month of gestation is certainly not powerful enough to face the difficulties of existence. According to Friedrich Nietzsche, a 19th-century German philosopher, "Man is an unfinished animal." Young animals, even though not self-sufficient at the time of their birth, are much closer to self-sufficiency than the young born of men. Man is the species with the longest infancy; according to the Dutch anthropologist Louis Bolk, the birth of the child seems premature. Although long, the length of gestation for man is insufficient; it is prolonged by an extra-uterine gestation during which the infant builds up its bodily organs and at the same time undergoes the fashioning effected by the family and social environment.

Prior to any conscious and organized education, even before the beginnings of articulated language, the infant undergoes an initial education that establishes patterns of sensorimotor coordination and sets into action a training process in accord with the values of the surrounding environment. By the time of self-consciousness, an individual has already been structured by years of continual fashioning, which have left him with only restricted latitude for self-initiative. The child learns to speak with the words of others; he learns to feel and to act in terms of pedagogical schemata imposed by the environment. Even prior to birth, the intra-uterine existence of the human embryo is influenced by the mother's way of life. Thus, the respective confines of biological and cultural existence cannot be exactly determined, but the idea of some form of biological inscribing of culture onto the human species cannot be rejected. According to Adolf Portmann, a Swiss biologist, "human heredity, in its own proper way, is not essentially genetic, but social." Man's historical and cultural generation by means of pedagogy is indissolubly related to his physical generation by natural means.

Man as the creator of culture. The representation of a naked human by an artist, or even by a photograph, can be dated with precision by an expert. Even without clothing, which would reflect a particular fashion, it is possible to date the portrayal of the body of a man or a woman, because it is remodelled by the demands of culture. More generally, human intelligence cannot claim to grasp nature such as it is in itself; every statement about nature is in the same category as a representation of culture, which is a phenomenon of consciousness. As Wilhelm Dilthey, a German philosopher and historian of ideas, wrote in 1883,

> Nature is alien to us. It remains something exterior to us and not interior. Society is our world. We live with the interplay of reciprocal reactions that animate it, . . . for inside and outside ourselves we perceive the states and forces that comprise the system of society in constant change—the supreme sign of life.

Man's use of symbols

Natural, or physical, anthropology thus becomes an integral part of cultural anthropology. Man's relationship with the universe of his environment is inscribed in a "mental space" (*i.e.*, the whole of the psycho-socio-cultural contents of the thought of men in a given time and a given society) that arranges the meanings expressed by words and ideas and permits their mutual communication. For Ernst Cassirer, a 20th-century German cultural philosopher, man is a symbolical animal (*homo animal symbolicon*): the concrete world in which each man lives out his daily life (*Lebenswelt*) is represented by generally accepted symbols that are appropriate to a particular society at a certain moment of its history.

Thus, for the cultural anthropologist, humanity is the product of its own total past history and activity, and each individual is both the product and the support of a collective consciousness that defines a particular moment in the history of the human spirit. Each epoch interprets the sum of its traditions, norms, and values in such a way as to impose a model for behaviour on daily life as well as on the more specialized domains of morals and religion and of art and science. In the early 18th century, Vico affirmed that, because man is the author of the moral world, he is able to construct a scientific analysis of the forms or symbols of culture in terms of any of the great forms of intelligibility that inspire the law, politics, language, and religion of different peoples. At the end of the 18th century, the German Johann Gottfried von Herder, a precursor of Romanticism, asserted, "we live in a world that we ourselves have created."

The philosophy of history attempts to discover a rational law for these successive social forms. Cultural anthropology merely describes empirically the various observable mental systems in their diverse places and times. Man thus appears as the prisoner of the established culture, enmeshed in the spirit (*Zeitgeist*) and world view (*Weltanschauung*) of his times; and there is the risk that anthropological reflection will degenerate into historical relativism—that is, a description of markedly different cultures, each of which is closed in upon itself as if it were an absolute of a new type. The sociology of knowledge indicates that each individual existent is the prisoner of the cultural space in which it originated; its ideas, its customs, its behavioral patterns, in virtue of its social predestination, must be regulated according to the ideal model (the *Idealtypus* of the German sociologist Max Weber) prevailing in its environment and time period.

Cultures known only in terms of specific men and works

The assertion that the thought of the individual is only a particular affirmation of the general state of consciousness overlooks the fact that culture is a creation of individuals. It cannot explain, moreover, the continual transformation of "mental space" throughout history. A cultural moment is always only a pause in the process of becoming; the total situation consists of a gathering of a great many personal affirmations that make up something like a statistical table. The great cultures, throughout their history, pass through successive phases of growth, peak, and decadence; immobility would be tantamount to the stagnation of creative forces. Neither the formation nor the transformation of cultures can be understood unless the interventions of individual personalities are acknowledged. Whether it is an ancient African culture or the European Enlightenment, the formation and transformation of a culture are understandable only through specific works, whose authors cannot be subordinated to some reductive abstraction. It is true that the human being is a product of culture, and it is equally true that culture is produced by men; the two truths balance each other and include each other.

Philosophical anthropology must, therefore, be distinguished from cultural anthropology. The latter is an inventory of the great systems of culture; the former endeavours to find the process of personality that attempts to express itself through the symbols or terminologies of the community. Philosophical reflection considers the cultural vocation to be an integral part of the human being. Man's social incarnation, like his bodily incarnation, is a prerequisite for his existence in the world; far from being a form of oppression or of abdication, it is a means of liberty.

MAN'S STATUS AS AN INDIVIDUAL PERSON

Max Scheler, a German philosopher and Phenomenologist, wrote in 1928:

> I can state with some satisfaction that the problems of anthropology have become today in Germany the true centre of all philosophical studies, and that, outside the circles of philosophers, the biologists, physicians, psychologists, and sociologists also are endeavouring to constitute a new view of the essential structure of man.

This text makes it possible to date the advent of philosophical anthropology as the branch of thought that recognizes in the personal unity of the human being both the point of departure and the goal of the reflective process. Certain characteristic tendencies of Western philosophy in the first half of the 20th century—among others, Phenomenology (concerned with the description of phenomena as consciously experienced), Existentialism (stressing the concrete in human personality, as opposed to abstract essences), and Personalism (regarding personality as the key to the interpretation of reality)—centre on the human condition as it is lived by the individual. In the history of thought, philosophical anthropology appeared when the pursuit of truth was no longer considered to be the attempt to attain an abstract, absolute truth but to define an incarnate truth, a concrete truth conditioned by personal existence.

The "individual" of the 19th century

Rise of the concept of person. In the course of the 19th century, the material and intellectual framework of the traditional civilization was gradually broken up under the impact of the Industrial Revolution and political revolutions. The change in life style coincided with the appearance of new philosophical categories and the rise of two opposed tendencies: Socialism, which stressed the role of the state or the social group; and individualism, which gave precedence to the interests of the individual over those of the group. The "individual" of the 19th century withstood the new types of conformism of the masses; he rejected the law of numbers and made a new absolute out of his own will. The German Anarchist philosopher Max Stirner, the Danish Christian thinker Søren Kierkegaard, and Friedrich Nietzsche each professed a radical individualism. But their individual seems to have a basically negative attitude: he affirms himself against all others; he is the exception to the rule, the unique against the state (Stirner), the Christian against the church (Kierkegaard), or the superman against humanism in all its forms (Nietzsche). The concept of the person in 20th-century Personalism, however, attempts to be a positive notion that integrates the components of man rather than excluding them: rather than displaying an exceptional prophetic character, this person is a totality of the values that compose every human being. Instead of opposing man to communities, Personalism attempts to conciliate the individual with the group.

The awareness of self

Far from being an immediate datum of consciousness and of action, the individual's awareness of his own existence is a late acquisition. In primitive cultures, man remains ignorant of his own identity; his name designates his role in the society. The thresholds of his own thought are as unknown to him as the thresholds of his own body. Man in some of the primitive societies, for example in Australia or Africa, had no precise notion of the meaning of his own life and death. For these societies, collective existence absorbed individual existences; after contact with the West, however, the self-consciousness that arouses individualization became manifest, at the expense of the ancient culture, which had been founded on the interplay of collective representations. Hellenistic thought initiated something of a Copernican—or Socratic—revolution by making the thinking subject the centre, or source, of truth. The effects of this rational individualization were, nevertheless, disguised in the early Greek thought by the emphasis given to the order of the cosmos and in the later Christian thought by the emphasis given to God. Philosophical anthropology seeks to place a greater emphasis on man than did the earlier models of thought.

The scholars of the East have not been proponents of the anthropological revolution. In India the affirmation of individual existence is often considered a form of error and evil. Salvation is found in the renunciation of individuality, in the total effacement of self in the heart of Nonbeing. This point constitutes one of the principal differences between the spirituality of the East and that of the West, which would more readily accept Goethe's proposition that the quality of personality alone can assure the happiness of men.

In Western societies, the growth of the child seems to reflect the development of the history of thought. Like man in the ancient societies, the newly born and the nursling are not centred upon themselves; they relate directly to their environment and are incapable of distinguishing their bodies from the nearest objects. Their individuality seems to dissolve and disperse in their ambience. Beginning about age three, the child finally attains a rather precise awareness of his own existence. Several more years pass before the young person discovers the autonomy of his thought; and it is only after the years of adolescence that the balance of subjective individuality and moral responsibility are attained—if they are ever attained.

The status of the individual person varies with the historical context of the prevailing civilization. In each age the rights and duties of the individual, his norms and values, and his principles of knowledge are questioned. Philosophical anthropology acknowledges the relativity of principles in the changing of the times and yet endeavours to help man prevail against forces that menace the recognition of his humanity. It may be seen as a reaction to the totalitarian systems of the 20th century: Italian Fascism, Soviet Communism under Stalin, and German Nazism were powerful ideologies calling for the annihilation of the individual character of the person. The philosophical protests of the German Phenomenologist Max Scheler, of the Russian Existentialist Nikolay Berdyayev, of the Jewish philosophical theologian Martin Buber, and of the French Personalist Emmanuel Mounier offered answers to this challenge to humanity; the philosophies of the person and the philosophies of existence present to each man the means to centre himself upon himself.

Person as a concrete reality. Philosophical anthropology is not the expression of a contemplative or speculative attitude: its emphasis on the reality of the concrete man means that man cannot be defined in terms of the reflective consciousness alone. The identity of man must be found in a more comprehensive conception.

The "I" of intellectualism. The *cogito* ("I think") of Descartes was spoken in the first person, but the "I" was only the focus of a provisional application of the universal truth that a thinking subject is necessarily an existing subject; and Descartes was preoccupied with its universality, not its personal character. In the intellectualism of Western thought from Spinoza to Kant, the "I" was merely such an application of a transcendent truth that was one and indivisible. Even after Hume's skeptical criticism, a trace of this rationalistic understanding of the individual remained in Kant's transcendental "I"—the subject of knowledge, whose existence cannot be known from concrete evidence but is postulated as necessary. Kant felt obliged to complete his critical philosophy with *Anthropologie in pragmatischer Hinsicht abgefasst* (1798; "Anthropology Composed from a Pragmatic Viewpoint"), in which he was concerned with the diverse forms of empirical human nature and the practical aspects of existence. It was as if each man could live a double life divided between the speculative exercises of the pure spirit and the inglorious activities of daily life.

Philosophy of incarnation

The "self" of anthropology. Philosophical anthropology is a philosophy of incarnation, in which each existing individual is seen as doubly involved in concrete reality: he is identified with a body, and he is in solidarity with other individuals in the social groups of which he is necessarily a part. The "I" of intellectualism is isolated from both itself and others; the "self" of anthropology is intimately involved with its body and with the communities (such as family, state, nation, church, and political party) to which it is tied by relationships that it could not deny without denying itself. The unity of this self is not immediately evident as though it were a presupposition of reason; rather, the self searches for itself in a dialogue with the diverse elements that constitute its concrete singularity. Its search is a task that is identified with the total course of its life, through the diversity of situations that must be confronted one after the other. According to the philosophical anthropologist, there is no

universal formula of truth, nor is there a definitive criterion for it. Truth takes on the sense of this search for self, about which nothing can be affirmed before the conclusion; the truth or falsity of a life is expressed in the development of that life, and it remains in question until the end of that life. The search for the truth is tantamount to the truth of the search.

Johann Friedrich Richter, a German Romantic poet, wrote about a day in his childhood, about the year 1770, when, at the doorway to his father's house, as he contemplated the surrounding countryside, he was struck by the thought: "Ich bin ein Ich" ("I am a self"). This illumination can be considered the inaugural moment for a philosophical anthropologist; it is the counterpart of the Cartesian *cogito ergo sum*, which was the point of departure for Rationalistic intellectualism. "I am a self," as perceived by the young Richter, meant the coming to consciousness of one's own personal existence in its irreplaceable uniqueness with its responsibilities: I exist by myself and in myself; I have been born and I will die; I have control of myself and of the body that is mine; it is in my name that I must confront the world and men throughout the uncertain span of time that stretches out before me. Martin Heidegger, a contemporary German Existentialist thinker, has insisted that the existing human subject has been "thrown into the world"—that is, the beginning of his history does not depend on himself—and yet he must give meaning to this ordeal of life that he did not ask for and that perhaps can have no meaning.

The person as the subject of values. In philosophical anthropology, the individual human being is considered the subject of fundamental values. He does not create these values; rather he receives them from the various pressures of the world in which he lives. Nevertheless, they only have authority over him to the extent that he commits himself to them. This commitment to values serves as a framework for the exercise of a freedom that differs significantly from the theoretical and speculative freedom of traditional philosophy. The traditional question of the relationship between human freedom and the idea of an all-powerful God is also raised for the philosophical anthropologist. To the extent that the individual has not given himself existence and finds himself thrown into the world, the individual cannot benefit from a complete freedom without limit. Complete freedom is reached by way of a conditional freedom—that is, a search for a certain amount of order in which the individual attempts to imprint his personal mark on the circumstances that seem to determine the course of his life. In this view, there is, for man, no divine right to freedom, as though it were his prerogative; man's freedom must constantly be liberated, and it is always provisional, insofar as the subject imposes upon himself an obligation to the essential responsibilities in his life.

Man's concrete freedom

Threats to the concept of person. Philosophical anthropology can, perhaps, be described in terms of what it rejects. It is opposed to an ontology (philosophy of Being) of a Rationalistic type because it emphasizes the singular over the universal; it intends to be a philosophy of the living experience of concrete man, without any reference to ontological essences. Man must be understood as a man in the human world; the human condition is not the divine condition, and man is not a god in miniature or a fallen god. It also rejects the threats to the self posed by other theories. Materialism, for example, in all its forms tends to reduce personal existence to an extrinsic condition of a physical or biological nature. Other threats of this kind may be seen to come from social theories. The French sociologist Émile Durkheim, for example, saw the thought of the individual as a support for collective symbols; the individual's ideas and values do not belong to him, but rather he receives them already formed from the groups to which he belongs, and he submits himself passively to their discipline.

Each of these themes of modern thought presents a threat to philosophical anthropology. Biological Materialism inspired the thought of Freud to a great extent. Psychoanalysis denounces the illusions of consciousness; it analyzes individuality as a combination of elementary forces below the level of consciousness, which largely remains a prisoner of pressures it cannot control. Freudian therapy attempts to increase conscious control over the unconscious; but its critics regard its therapeutic methods as erring through an excess of rationalism, while its basic conception errs from an excess of irrationalism. Similarly, the Historical Materialism of Karl Marx subordinates the consciousness of individuals to the economic regimen and to the relationships between the social classes. Although the Historical Materialists seek to liberate the human person, this liberation is relegated to the remote future; meanwhile, the human consciousness in this view seems to have been stripped of all initiative.

Psychoanalysis, Marxism, and Structuralism

A final, and perhaps still more radical, threat to personal existence is found by many in contemporary Structuralism, which analyzes social institutions in terms of highly abstract structural relationships. The structural anthropology of the French social theorist Claude Lévi-Strauss seems, in effect, to some of its critics, to actually eliminate man. Man's real behaviour is analyzed according to a logical model of what the Structuralists perceive to be his collective unconscious, a kind of cybernetics, or logical control system, presiding over the thought of all individuals and the functioning of all groups. The result of this approach, as some scholars have said, is that the individual man has ceased to exist in it.

Structuralism, Marxism, and psychoanalysis are currents of thought that have been greatly influenced by the development of the human sciences. The fact of their existence indicates the role that philosophical anthropology must play in relation to physical and cultural anthropology. In these sciences man is no longer presented as the subject but becomes the object of doctrinal syntheses. The proper role of philosophical reflection is to recover the individual consciousness from all the axiomatic systems that attempt to overwhelm or reduce it. Philosophical anthropology affirms that in the last resort each human being—despite the multiple temptations of the world and of other men, and despite all internal and external influences—must assume the responsibility of his own existence. The dangers of alienation (separation in some way from the social group) are numerous; they can alter somewhat the course of a person's history, but they cannot determine it absolutely.

The person as a mystery. Philosophical anthropology is concerned with the irreducible mystery of each personal life, about which the models and schemata of psychology can provide no more than a limited understanding. Following Wilhelm Dilthey and his student Bernhard Groethuysen, the historical schemata of personality in successive cultures—such as ancient man, medieval man, and the Renaissance Humanist—can be defined; it is also possible to follow the ancient tradition of the theory of temperaments and of physiognomy (the art of discovering temperament and character from outward appearances) and attempt to classify personalities in terms of their character. Philosophical anthropology can borrow an approach from such endeavours, but it cannot reduce itself to an individual psychology of historical models or to a psychophysiological study of character if it is to maintain the primacy of the metaphysical, free choice of the self to seek to find itself. The factual data remain subordinated to the individual's affirmation of himself as a first person—that is, an "I" or a "self"—whose fundamental choice transfigures whatever elements the choice sets into operation. It is only within this perspective that the forces pressing upon the individual can be understood and distinguished: the organic and neurobiological order of forces related to bodily existence; the intellectual order of forces related to rational reflection; and the forces in the order of values related to personal choice. By reflection upon his own existence, the individual realizes a sublimation of the bodily orientations and establishes the meaning of his personal destiny. Such a hierarchical schema of forces has meaning only when incarnated in a concrete exis-

tence. A similar statement can be made about any schema that is proposed.

Philosophical anthropology as an orientation

Philosophical anthropology does not constitute—and never has constituted—a school of thought. Rather, it represents an orientation common to various tendencies in contemporary thought. In this sense, it embraces all those philosophers who are concerned with aspects and degrees of personal being in the world: both the bodily and the social incarnation of the self, in the state of both dependence and autonomy with respect to the world and to others.

These concerns are present in contemporary Phenomenology, which was inspired by the German Edmund Husserl and his student Max Scheler, in whose view personal consciousness is the link to the origin and the goal of concrete truth. The Existentialism of Karl Jaspers, Nikolay Berdyayev, Martin Heidegger, Gabriel Marcel, Maurice Merleau-Ponty, and Jean-Paul Sartre accords a priority to the individual consciousness as the source of all concrete commitments. The Personalism of Emmanuel Mounier is a philosophy of responsibility and fidelity to oneself, pursued by moral and political action. In Anglo-Saxon thought, the anthropological desire is found in William James's Pragmatism—a philosophical movement according to which meaning and truth are to be found in practical consequences; in John Dewey's instrumentalism—a philosophical movement according to which ideas are instruments for transforming problems into solutions; and among the numerous disciples of these masters.

In conclusion, philosophical anthropology can be considered as the modern form of humanist thought directed against both the excesses of a triumphant Rationalism that slights the incarnation of the human consciousness and the threats of a dissolution of the personality by the deterministic notions of the natural sciences and the human sciences. Merleau-Ponty wrote:

> I am not the result or the intersection of multiple causalities that determine my body or my "psychism"; I cannot conceive of myself as nothing but a part of the world, as the simple object of biology, psychology, and sociology, nor close over myself the universe of science. Everything that I know of the world, even through science, I know from a viewpoint that is my own . . .

This human presupposition is the basis of philosophical anthropology.

PHILOSOPHICAL ANTHROPOLOGY IN RELATION TO GOD

Theistic philosophical anthropology. Philosophical anthropology is devoted to the task of determining truth on a human scale by conciliating man's relationships with the natural world and the cultural world. For most philosophical anthropologists, this locating of man in the totality of things must also take into account man's relationship with God, the privileged source of all truth for the traditional philosophies of the West and for most of the spiritualities of the East.

Philosophical anthropology has attempted to effect the shift, initiated by Hellenistic thought, that replaces God with man as the guideline for the affirmation of truth. The world's great religions have each developed a certain conception of the reality and the destiny of man that has depended upon their conception of the divinity.

Concept of "religious man"

The concept of *homo religiosus* ("religious man") has varied at different times and in diverse cultures, but it has included certain fundamental categories—such as purity and impurity, obedience and disobedience, fidelity, salvation, and damnation—that are common to the principal forms of religion. The study of man seems to have been subordinated to an eschatology (doctrine of the last things) in which individual actions find meaning in a symbolic framework that superimposes upon them a supernatural value.

Christian anthropology. Each theology gives rise to its own anthropology. The anthropology of Christianity has been of considerable importance for Western philosophy in that, since the Middle Ages, metaphysics has developed as a by-product of man's theological reflection, which has remained dependent on the Judeo-Christian revelation as recorded in the Bible. Western man has become aware of himself in terms of the *homo Christianus* ("Christian man"), who is normatively described in the Old and New Testaments. Man's existence is that of a creature included in the totality of creation and responsible before his Creator. Marked by the sin of Adam and redeemed by the incarnation of Christ, each creature is to follow his own way within a history of salvation that began long before his birth and will be completed after death in the glory promised to the elect or in the eternal damnation promised to the reprobate.

There is, therefore, a Christian anthropology; it is a religious anthropology and not a philosophical anthropology. Until the end of the 17th century, the theological conception of truth, which was commonly accepted as absolutely authentic, described human nature in terms of its supernatural destiny. Philosophical anthropology is thus seen as the consequence of a reversal of values that began in the 18th century with the decline of theological absolutes.

Human approach to God. Religious truth had been understood as eternal and immutable; the classic formula, first expressed in the 5th century by St. Vincent of Lérins, was that religious truth is "what is believed everywhere, at all times, and by all." It began to appear, however, that such a definition had no meaning in the human domain. Theological discourse is a human discourse: it is not God but, rather, man speaking about God. Men can approach the truth of God only in human terms; theology thus becomes a search, the results of which must continually remain in question. Every affirmation of the divine truth is a human affirmation marked by the particular circumstances of the place and time in which it is formulated. Religion has been involved in the transformation of the forms of Western civilization since the Renaissance. The religious forms (symbols, institutions, structures) have borrowed from the general culture intellectual and scientific language for the requirements of doctrine and apologetics as well as architectural and musical language for the celebration of the liturgy.

The decay of theology that began in the 18th century is the most evident sign of this change in the meaning of truth; the traditional plan for a philosophical approach to God was abandoned in favour of a philosophical approach to man. Nevertheless, God is not thereby eliminated; rather, man and God are perceived in terms of the relationship between them, which is limited by the concrete circumstances of the individual man.

The philosophers of the age of the Enlightenment—such as Locke and Hume, Condillac and Kant—had believed that the transcendence of God doomed to failure any attempt to encompass him within the framework of human discourse. For the philosophical anthropologist, however, revelation is not the manifestation of God as he is in himself but is a message from God to man that is expressed in the concrete terms of a particular language and culture. More than a response or a completed formula, revelation is a question and a challenge; it is the task of each man to find its meaning for his personal benefit. In this approach, the freedom of the individual subject is affirmed; the presence of God, however, remains the indispensable foundation for the placing of man in his particular context.

Revelation as a concrete message

The correlation between anthropology and theology is reflected in the early developments in the comparative study of religions. At the beginning of the 18th century, the French thinker Bernard Fontenelle observed: "To the extent that men have become more perfect, the gods have become more so." Friedrich Schiller, a German poet and dramatist, emphasized in 1789 that "man depicts himself in his gods." In the same period, Johann Gottfried von Herder, a philosopher and theologian, affirmed: "It is necessary to read the Bible in a human manner, for it is a book written by men for men." Religion was thus humanized. Nevertheless, the Bible was not desacralized; on the contrary, it acquired a new sacredness. Reading the Bible in a human way was understood to require a new interpretation of revelation

that would include the notion of the search for God in the history, the language, and the customs of men. Herder also wrote in 1778: "The fact that religion is integrally human is a profound sign in recognition of its truth."

A connection between the image of God and the image of man is found in the Judeo-Christian revelation. The change in the religious consciousness in modern times is recognized as merely restoring and developing this theme: the history of humanity is also a history of divinity. The present dialogue between man and God is seen as a provisional moment that prolongs the past forms of dialogue and will yield to new forms in the future.

This human approach to the divine, freed from theological presuppositions and orthodoxies, is described particularly in the work of Pascal and Kierkegaard. For them, the "wretchedness of man without God," his aloneness, and his mortality find their meaning and fulfillment only in the experience of grace, which inexplicably raises man up from his worthlessness and gives meaning to an existence that, outside of a relationship with God, would have had no meaning. This overcoming of anguish, this liberation from the sin of existence, enables man to know that he will enjoy the plentitude of eternal life promised to the believer by revelation. This interpretation safeguards the possibility of a philosophical anthropology because God is now reached through his relationship with the individual, through his discovery by the individual. The presence of God has not been obliterated; rather, the evidence of man takes precedence over the evidence of God, and the search for God becomes an aspect of the search for man.

Atheistic philosophical anthropology. The relationship just described is, however, an uncertain formula and is disputed. From the moment that the primacy of anthropology is affirmed, the possibility of a complete elimination of theology arises. The dialogue between the human and the divine tends more and more to restrict the part of God. The Renaissance Humanists celebrated man as a second God, in the image of God (*homo alter Deus*); certain modern thinkers would consider God as another man, the inverse image of the human being (*Deus alter homo*).

In the 19th century, the German Materialist Ludwig Feuerbach, in his work *Das Wesen des Christentums* (1841; Eng. trans., *The Essence of Christianity*, 1957), declared that man is not a shadow of God; it is God who is the shadow of man, an illusory phantasm that man nourishes out of his own substance. He states:

> The Christian religion is the relation of man to himself, or more exactly, to his own essence, but to his own essence as to another being. The divine being is nothing other than the human essence, or more exactly the essence of man, separated from the limits of the individual man, that is, real, bodily, objectivized; that is, the essence of man contemplated and honoured as another being distinct from himself. All the characteristics of the divine being, then, are the characteristics of the human essence.

Anthropology, thus, took over the entire realm of theology. Max Stirner, in his work *Der Einzige und sein Eigentum* (1845; Eng. trans., *The Ego and His Own*, 1907), considered God to be the product of an alienation of the human being, contrived for the purpose of having an illusory being to keep the human from being himself. God is the major obstacle to the blossoming of human freedom in its radical fullness, which requires the suppression even of the idea of divinity. The theme of the death of God was taken up again by Nietzsche and, in the mid-20th century, by such thinkers as Jean-Paul Sartre, William Hamilton, and Thomas J.J. Altizer.

Criticism of atheism

Modern atheism can possibly be considered the triumph of philosophical anthropology: by eliminating all transcendence, man recovers the totality of the human realm; he becomes both the author and the actor of all truth. But in fact, to many thinkers, contemporary consciousness seems not to justify such a desacralization of truth and of civilization. Delivered from God, man is threatened by the determinism of the natural sciences, which consider man to be a complex of determined reactions; or, under the influence of the social sciences, the individual is absorbed into some collective representation of the environment. Some thinkers proclaim that "man is dead"; that is, personal identity and individual existence are only the illusions of Idealists whose influence has waned. For them man exists only as either biological or social matter that has its own meaning and validity without regard for particular human beings. Within this perspective, all justification of being and of coherent thought disappears; the reality of the world in its becoming assumes the form of a disorganized frenzy.

The need for a relationship between anthropology and theology. There is a striking correlation between the themes of the death of God and the death of man; this correlation seems to verify the mutual interrelation of theology and anthropology. The stress given by atheistic Existentialism to the anguish of the human being thrown into an absurd world emphasizes the essential function of the divine as the foundation of values and the guarantor of the meaning of existence. The desacralization, or profanation, of man is illustrated by such phenomena as the modern attempts at genocide or the haphazard development of a technological society that threatens man with a new period of barbarianism.

The philosophical anthropologist has not been able to avoid taking sides in the crises of contemporary society. The task of his science is to preserve the human meaning of human existence. For most philosophical anthropologists, the theological dimension, or some reference to God, is indispensable in order to hold in check the contradictory threats that would destroy the image of man and to maintain the opening to transcendence in which each individual consciousness can discover its meaning, its values, and its freedom. If the principal preoccupation of ancient thought was to prove the existence of God, the primary task of contemporary thought is to prove the existence of man.

BIBLIOGRAPHY. By restricting the vast literature discussing man to only those sources that accept and consider the data of the natural sciences and of the cultural sciences the bibliography of philosophical anthropology becomes very short.

General works include MICHAEL LANDMANN, *De Homine* (1962), a German selection, philosophical rather than anthropological, with an important bibliography; J.S. SLOTKIN, *Readings in Early Anthropology* (1965), a good selection of important historical texts, mainly of Anglo-Saxon thinkers; G. GUSDORF, *Les Sciences humaines et la pensée occidentale*, 6 vol. (1966-72), a general history of the sciences of man on the basis of an anthropological philosophy; ERNST CASSIRER, *An Essay on Man* (1944), a useful and accurate sketch; and ERNST KROEBER (ed.), *Anthropology Today* (1953), which illustrates the present state of inconsistency of the sciences of man.

The topic of man as a natural being has been studied by German biologists—*e.g.*, JACOB VON UEXKULL, *Theoretische Biologie*, 2nd ed. (1928); A. PORTMANN, *Biologie und Geist* (1956); and A. GEHLEN, *Der Mensch* (1940)—who intend to give a philosophical meaning to the physical structure of man. The anthropology of TEILHARD DE CHARDIN, *Le Phénomène humain* (1955; Eng. trans., *The Phenomenon of Man*, 1959), extends from cosmology to theology. Man as a cultural being is the main theme of W. DILTHEY, *Einleitung in die Geisteswissenschaften* (1883), the starting point of a new approach to man; ERNST CASSIRER, *Philosophie der symbolischen Formen*, 3 vol. (1923–29; Eng. trans., *The Philosophy of Symbolic Forms*, 3 vol., 1953–57), a classical analysis of cultural institutions; and B. GROETHUYSEN, *Anthropologie philosophique* (1952), a series of historical sketches of human personality from antiquity to the Renaissance.

Philosophical anthropology for some contemporary philosophers is a positive approach to the concrete reality of man, considered not only as a rational being, but as a whole living existence. Such is the position of the German Phenomenologist MAX SCHELER, *Wesen und Formen der Sympathie*, 2nd ed. (1923; Eng. trans., *The Nature of Sympathy*, 1954); *Die Stellung des Menschen im Kosmos* (1928; Eng. trans., *Man's Place in Nature*, 1961); the Jewish thinker MARTIN BUBER, *Ich und Du* (1923; Eng. trans., *I and Thou*, 1957); the Existentialist philosopher of Russian origin NIKOLAI BERDYAEV, *Cinq Méditations sur l'existence* (1936); and MAURICE MERLEAU-PONTY, *Phénoménologie de la perception* (1945; Eng. trans., *The Phenomenology of Perception*, 1962). CLAUDE LEVI-STRAUSS, *Anthropologie structurale* (1958; Eng. trans., 1963), is a summary of the Structuralistic view of man.

(G.P.G.)

Antibiotic

Antibiotic is derived from the word antibiosis, which was first used in the 1880s to describe the destruction of one living thing by another. Antibiosis exists throughout the world of living things; bacteria, for example, attack other organisms in order to obtain nourishment and to reproduce, as do molds, which also exist as parasites on other species. Organisms, however, have developed various protective devices (*e.g.*, secretion of toxic substances), and it has been known for many years that antagonism can exist among micro-organisms. Early application of bacterial antagonism involved the use of a harmless micro-organism to eliminate a disease-causing one. Present-day use of antibiotics, however, is based on utilization of an active inhibitory substance obtainable from antibiotic-producing organisms. Antibiotics now are defined, therefore, as chemical substances or metabolic products of organisms, generally micro-organisms, that are detrimental to other micro-organisms. Most of the useful antibiotics are produced by species of the mold *Penicillium* and various *Streptomyces* bacteria. Many species that synthesize antibiotics have been isolated from soil.

The most useful antibiotics destroy or inhibit many kinds of disease-causing micro-organisms and have relatively nontoxic properties when injected into animals, including man. Antibiotics can be conveniently grouped according to the types of functions they inhibit. Certain antibiotics interfere with bacterial cell-wall synthesis; these include the penicillins, cephalosporin derivatives, vancomycin, and cycloserine. Antibiotics that interfere with protein synthesis in bacterial cells include the tetracyclines, chloramphenicol, erythromycin, lincomycin, novobiocin, streptomycin, gentamicin, kanamycin, and neomycin. Some antibiotics affect the cell membranes of bacteria, yeasts, or fungi; this group includes polymixin, colistin, nystatin, and amphotericin B. An antibiotic that affects nucleic acid metabolism in certain fungi is called griseofulvin.

Types of antibiotic action

HISTORICAL BACKGROUND

Attempts to use substances derived from one organism to inhibit or kill others began at least 2,500 years ago, when the Chinese became aware of the curative properties of moldy curd of soybeans and used this substance to treat boils, carbuncles, and similar infections. The first suggestion that bacterial antagonism might be important in treating disease was made by two Frenchmen, Louis Pasteur and Jules-François Joubert in 1877 after they discovered that bacteria, which are responsible for an infection called anthrax, grew rapidly if inoculated into urine free from micro-organisms but died if so-called common-air bacteria were present. The failure of an anthrax infection to develop in animals under similar conditions led to the conclusion that an injection of anthrax-causing bacteria did not cause illness if ordinary bacteria were given at the same time; this was the first clear demonstration of an antibiotic effect.

Discovery of penicillin

The antibiotic era did not begin, however, until about 1928, with the discovery by Sir Alexander Fleming, at St. Mary's Hospital in England, that a growth culture of the pus-producing bacterium, *Staphylococcus aureus*, had disappeared in an area in which a green mold was growing. Since the organism that produced the substance that killed the bacteria was a species of *Penicillium*, Fleming named the substance penicillin. Attempts to treat human infections with this material were not encouraging, however, because the substance was unstable and lacked potency; not until several years later did several workers at Oxford University examine the possibility that stable penicillin might be produced in large enough quantities to treat human disease. In 1941 the drug was used to treat serious infections; the results were dramatic, as patients who received penicillin made rapid and complete recoveries. World War II interfered with the large-scale manufacture of penicillin in Great Britain. As a result, methods for its mass production, purification, and stabilization were, instead, developed in the United States. The development of penicillin was the beginning of an era that has been called the golden age of chemotherapy. Since 1948, a large number of substances, which inhibit or kill bacteria and some fungi, have been discovered.

GENERAL FEATURES

Range of antibiotic activity. Antibiotics, often called antimicrobial drugs, are chemical substances synthesized by micro-organisms. Bacteriostatic antibiotics inhibit bacterial reproduction but do not kill bacterial cells. If a bacteriostatic antibiotic is removed from exposed micro-organisms, they usually reproduce normally. Bactericidal antibiotics kill micro-organisms by direct action. Antibiotics also may be distinguished on the basis of the breadth or spectrum of their activity. Broad-spectrum antibiotics (*e.g.*, tetracyclines, chloramphenicol), for example, may affect not only a large number of species in one bacterial genus but also unrelated micro-organisms; *e.g.*, bacteria, rickettsiae, Bedsonia, Mycoplasma. Narrow-spectrum antibiotics (*e.g.*, penicillin, streptomycin) usually have limited effects and may act against only a few bacterial species.

Sensitivity and resistance of micro-organisms

No antibiotic is effective against all micro-organisms. Micro-organisms susceptible to a specific agent are sensitive to that agent; unaffected micro-organisms are considered resistant. The range of activity of antimicrobial substances probably is related to several factors, including the availability of appropriate sites on the micro-organism to which the antibiotic can become attached and the production of substances by micro-organisms that inactivate the antibiotic. Examples of the latter are two enzymes (organic catalysts) that are synthesized by bacteria; the enzyme penicillinase destroys some types of penicillin, while the enzyme cephalosporinase inhibits the activity of the antibiotic cephalothin.

Importance. With few exceptions, the antibiotics are the only drugs that can cure diseases by eliminating their causes; *i.e.*, a variety of micro-organisms. The importance of antibiotics in saving lives and in reducing the duration of certain illnesses is evident from comparisons of fatalities caused by certain common infections before and after antibiotics became available. In the 1930s, before antibiotics were used, the fatality rate from pneumonia caused by pneumococcus (*Diplococcus pneumoniae*) ranged from 20% to 85% in the U.S.; by the 1960s, however, the fatality rate had decreased to about 5%. The use of antibiotics has reduced the number of patients who die from chronic infection of a heart valve from almost 100% to about 5%. Epidemic spinal meningitis caused by meningococcus bacteria (*Neisseria meningitidis*) once killed 20% to 90% of untreated victims; antibiotic treatment has reduced the fatality rate to about 2%. The effective use of chloramphenicol has reduced the death rate of typhoid fever from approximately 10% to 2% or less. Antibiotics also have played a significant role in the treatment of venereal diseases such as syphilis and gonorrhea.

In addition to curing certain established infections, the antibiotics are valuable in preventing some infectious diseases. By preventing throat infections caused by *Streptococcus* bacteria, penicillin, for example, is effective in preventing recurrences of rheumatic fever in susceptible individuals. The development of a venereal disease may be prevented by early administration of the proper antibiotic. Prolonged survival of children with cystic fibrosis of the pancreas is due in part to daily administration of an antibiotic. The use of antibiotics also has resulted in better control of so-called communicable diseases. Treating a patient with "strep" sore throat with penicillin or another effective antibiotic, for example, destroys the causative bacteria, thereby decreasing the chances for spreading the infection to other individuals. Appropriate antibiotics can eliminate diphtheria-causing bacteria from people who carry them but do not have disease symptoms.

Absorption and excretion. Antibiotics vary in the manner in which they are absorbed and excreted. Some antibiotics taken by mouth cannot get into the bloodstream in sufficient concentrations to be effective; in addition, anti-

biotics taken by mouth may be inactivated either by acid in the stomach or by intestinal enzymes. These antibiotics, therefore, are not absorbed from the intestinal tract into the bloodstream and must be given parenterally; that is, injected directly into a muscle or vein. Some antibiotics, however, can be administered either orally or parenterally. Antibiotics are excreted by the kidneys in urine or by the intestinal tract in feces; many antibiotics are transformed or metabolized in the liver before excretion. Such antibiotics combine with various substances in the liver to form conjugated compounds that are no longer active. Antibiotics taken by mouth but not absorbed in the intestinal tract usually are excreted in the feces. Antibiotics may be excreted in the urine in unchanged, inactivated, or in both forms.

Reactions. No antibiotic developed thus far is free of the potential to produce undesirable effects. Both incidence and type of reactions vary widely among different antibiotics and for individual ones. The number of reactions associated with the use of antibiotics increases both with the number of antibiotics available and with the number of people treated. In addition, the repeated administration of one or more drugs increases chances for reactions from repeated exposure.

Undesirable effects

Three general types of reactions are caused by antibiotics. Toxic effects may be produced by most of them. Hypersensitivity reactions, known for all antibiotics developed thus far, vary in type and severity and range from mild skin rashes to death. Various types of biological alterations in the host (*e.g.*, changes in types of intestinal micro-organisms) also may occur after the administration of any antibiotic.

Most individuals experience alterations in normal microbial populations (microflora) of the intestinal, upper respiratory, and genitourinary tracts after treatment with antibiotics. In some cases suprainfection, evidence of the appearance of a second infection during treatment of the first one, also may occur. Several factors increase the possibility for suprainfection. Children under three years of age are more likely than older ones to develop suprainfection, as are individuals with lung diseases other than tuberculosis. In addition, the range of antimicrobial activity of single or combined drugs is a factor.

Bacterial resistance. Resistance to antibiotics does not occur in all organisms or with all drugs. After 25 years of widespread use of penicillin G, for example, no important degree of resistance had developed in species of *Streptococcus* that usually cause throat infections or in pneumococcus, the most common bacterial cause of pneumonia. In contrast, penicillin-resistant *Staphylococcus*, which developed within a short time after penicillin was first widely used, are commonly found in hospitalized patients.

Some bacterial species isolated before the development of an antibiotic may be resistant to it; other species, however, apparently develop resistance only after a drug has been used for some time. Bacteria may develop resistance to antibiotics in one of several ways; *i.e.*, by mutation, transduction, transformation, or conjugation. Some species of micro-organisms, which are sensitive to a particular antibiotic, may undergo a spontaneous change or mutation in their genetic material without being exposed to the drug, thereby becoming resistant to it. Transduction involves a type of virus, called a bacteriophage, that reproduces in bacteria. A bacterial gene containing the information necessary to develop resistance to a specific antibiotic may become attached to a virus that attacks bacteria (bacteriophage) as it reproduces in a resistant bacterium. This gene then may be transferred by the virus to a sensitive bacterial strain, which becomes resistant to the drug. During transformation, a sensitive bacterial cell may become resistant to a specific drug by utilizing a gene that originally was part of the genetic material of a resistant bacterium. During sexual mating in bacteria, called conjugation, one bacterial cell may transfer genetic material responsible for resistance to a specific antibiotic to another bacterium; cellular contact is necessary, but resistance to as many as five or six drugs may appear rapidly and simultaneously in a previously sensitive bacterium. Among bacteria known to transfer drug resistance by conjugation are those that cause typhoid fever and some types of food poisoning (*Salmonella* species), those commonly found in the intestines of animals including man (*Escherischia coli*), and those that cause one form of dysentery (*Shigella*). Among antibiotics to which bacteria develop resistance during conjugation are streptomycin, the tetracyclines, chloramphenicol, ampicillin, and kanamycin. The transfer of resistance factors from one bacterium to another usually takes place in the intestines of animals, including man.

Transfer of resistance in conjugation

The general problem of the development of bacterial resistance to antimicrobial agents is important. The development of new antibiotics often offers only temporary solutions because micro-organisms also gradually develop resistance to these agents.

TYPES OF ANTIBIOTICS

An effective antibiotic must affect invading micro-organisms without damaging the cells of the patient, and, remarkably, many substances with this selective activity have been developed. It is probable that antibiotics act directly on the micro-organisms that cause infections rather than enhancing the defense mechanisms of the patient; however, no antibiotic cures an infection only by destroying the causative micro-organism. It is likely that even the bactericidal drugs require the effective intervention of defense mechanisms of the patient in order to destroy disease-causing micro-organisms.

Antibiotics that act on the bacterial cell wall. The antibiotics that inhibit or kill bacteria by interfering with their cell-wall synthesis include the penicillins, the cephalosporin derivatives, and vancomycin. Although these antibiotics differ in chemical structure, all inhibit the synthesis of a molecule called a mucopeptide, which is a necessary component of the bacterial cell wall. These drugs affect only bacteria that are reproducing. Since exposure to this type of antibiotic prevents the formation of a normal cell wall, the internal contents of the micro-organisms are lost, and cell death follows. Such bacteria are more susceptible to the environment than are bacteria with intact cell walls.

Inhibition of mucopeptide synthesis

The penicillins. Penicillin is synthesized by various species of *Penicillium*. During a metabolic process called fermentation, *Penicillium* produces several kinds of penicillin. If fermentation is altered by the addition of certain chemicals, 6-aminopenicillanic acid only is produced; penicillin G (benzyl penicillin) is formed under different conditions. Penicillins may be divided into two classes, biosynthetic and semisynthetic. Biosynthetic penicillins (penicillin G; benzathine penicillin, or Bicillin; and phenoxymethyl penicillin, or penicillin V), which are formed during the process of fermentation, sometimes are altered slightly in structure by chemical methods. Semisynthetic penicillins are made by adding certain compounds to 6-aminopenicillanic acid. With the exception of ampicillin, the semisynthetic penicillins resist breakdown and inactivation by the enzyme penicillinase, which is synthesized by certain bacteria. The penicillins, bacteriostatic in low concentrations and bactericidal in higher ones, may cause many undesirable effects (*e.g.*, many hypersensitivity reactions ranging from minor to severe, occasionally fatal).

Biosynthetic penicillins

Among the bacteria sensitive to penicillin G are those that cause throat infections (*Streptococcus* species), pneumonia (pneumococcus), spinal meningitis (meningococcus), gas gangrene (*Clostridium*), diphtheria (*Corynebacterium*), and syphilis (*Treponema pallidum*). Although 90 percent of the *Staphylococcus aureus* bacteria isolated outside hospitals are sensitive to penicillin G, the percentage decreases to 10 percent inside hospitals.

Penicillin G (which can be administered by mouth or injected into a muscle, beneath the skin, or into a vein) also has been applied to particular areas of the body; such topical application of the drug, however, often is accompanied by hypersensitivity reactions. A derivative of penicillin G called procaine penicillin is absorbed slowly after intramuscular injection. The action of penicillin G, which is rapidly excreted by the kidneys, can be pro-

longed by administration of substances such as probenecid (Benemid) that compete with penicillin G for available excretory structures. Probenecid commonly is used to prolong the action of antibiotics.

Benzathine penicillin (Bicillin), which has the same antimicrobial spectrum as penicillin G, also can be given by mouth or parenterally. Benzathine penicillin, however, is present in the blood longer than is penicillin G. Phenoxymethyl penicillin (penicillin V) is stable in stomach acid and readily absorbed from the small intestine.

Semisynthetic penicillins

Semisynthetic penicillins include methicillin, oxacillin, cloxacillin, dicloxacillin, nafcillin, and ampicillin. Methicillin (Staphcillin), effective against many strains of penicillinase-producing (*i.e.*, penicillin G-resistant) strains of *Staphylococcus aureus*, is less active against several other micro-organisms (*e.g.*, *Streptococcus* and pneumococcus) than is penicillin G. Oxacillin (Prostaphlin), a penicillinase-resistant semisynthetic penicillin, generally is more effective against penicillinase-producing *Staphylococcus* than is methicillin. Cloxacillin (Tegopen) has a range of antimicrobial activity similar to that of oxacillin. Dicloxacillin (Dynapen) appears to be more effective against *Staphylococcus* than the other semisynthetic penicillins.

Nafcillin (Unipen) is stable in stomach acid, penicillinase resistant, and similar to oxacillin in its effects against penicillinase-producing *Staphylococcus*. The antimicrobial spectrum of nafcillin generally is similar to that of penicillin G. Nafcillin is well absorbed from the intestinal tract after oral administration and may also be given parenterally.

Ampicillin (Polycillin, Penbritin, Omnipen) has broad-spectrum effects, inhibiting the reproduction of a number of bacteria. Destroyed by contact with penicillinase enzyme, it is, therefore, ineffective for penicillin G-resistant *Staphylococcus* and other bacteria that synthesize the enzyme. Although it is stable in stomach acid, ampicillin absorption in the intestinal tract is retarded by the presence of food; higher and more rapidly attained blood levels follow intramuscular injection. Most of the ampicillin injected into a muscle is excreted in unchanged form in urine. The drug appears in active form in bile in concentrations much higher (approximately 30 times) than those in blood.

Cephalosporin derivatives

Other antibiotics that affect the bacterial cell wall. The cephalosporin derivatives are semisynthetic antibiotics derived from a substance synthesized by the mold *Cephalosporium acremonium*, which first was isolated from marine waters. Although the fungus synthesizes three substances with antibiotic activity, only the active part (7-aminocephalosporanic acid) of one, cephalosporin C, is used to derive the three drugs presently used. These drugs, cephalothin (Keflin), cephaloridine (Loridine), and cephaloglycine (Kafocin), are synthesized by chemical methods in which various components are added to 7-aminocephalosporanic acid. Undesirable reactions to cephalosporin derivatives range from various allergies to kidney damage.

Cephalothin (Keflin) is bactericidal, and low concentrations are highly effective against a large number of bacteria; *e.g.*, many *Streptococcus* species, pneumococcus, *Staphylococcus aureus*, *Corynebacterium*, meningococcus, *Salmonella* (including the typhoid bacillus), and *Shigella*. Cephalothin is not well absorbed from the gastrointestinal tract and is administered by injection. About 25 percent of a cephalothin dose is metabolized to a weak antimicrobial compound, which is excreted in the urine.

Cephaloridine (Loridine) is bactericidal, with a range of antimicrobial activity similar to that of cephalothin. Cephaloridine is poorly absorbed from the gastrointestinal tract and, like cephalothin, must be administered parenterally to produce a therapeutic effect. Approximately 75 percent of a dose of cephaloridine is rapidly excreted in the urine. Cephaloglycine (Kafocin) appears to have less antimicrobial activity than cephalothin. Cephaloglycine is absorbed from the gastrointestinal tract and is excreted in active form in the urine, in which it reaches concentrations that inhibit various strains of bacteria. Because it appears in relatively low concentrations in blood, cephaloglycine is not used to treat systemic infections. A high concentration in urine, however, makes it useful in treating certain infections of the urinary tract.

Vancomycin and cycloserine

Vancomycin is a bactericidal antibiotic synthesized by *Streptomyces orientalis*, a bacterial species isolated from soil obtained in Indonesia and India. Organisms highly sensitive to vancomycin include several kinds of *Streptococcus*, *Staphylococcus aureus*, and pneumococcus. Micro-organisms that cause tetanus (*Clostridium tetani*), gas gangrene (*Clostridium welchii*), and diphtheria (*Corynebacterium diphtheriae*) also are highly sensitive. Among the organisms resistant to vancomycin are *Mycobacterium tuberculosis*, yeasts, and fungi.

Vancomycin, which is not absorbed when taken by mouth, is injected into a vein because intramuscular injection is very painful. Most injected vancomycin is excreted by the kidneys.

Among the reactions produced by vancomycin are skin rashes and shock; the most important undesirable effects, however, involve the ears and kidneys.

Cycloserine (Oxamycin, Seromycin), a broad-spectrum antibiotic produced by *Streptomyces orchidaceus*, is active against the tubercle bacillus *Mycobacterium tuberculosis*, including strains resistant to other antituberculotic drugs, and *Staphylococcus aureus*.

When given orally, cycloserine is rapidly absorbed from the stomach and small intestine. Reactions, which usually involve the central nervous system, tend to appear within two weeks after treatment is begun and to disappear when the drug is withdrawn.

Antibiotics that interfere with protein synthesis. A large number of antibiotics inhibit micro-organisms by interfering with their ability to synthesize proteins. Some antibiotics (*e.g.*, chloramphenicol, the tetracyclines, erythromycin, lincomycin) interfere with protein synthesis by binding to ribosomes the cellular structures responsible for the production of proteins. Others (*e.g.*, streptomycin, gentamicin, kanamycin, and neomycin) cause bacteria to produce abnormal proteins as well as inhibit protein synthesis.

The tetracyclines. Six tetracycline compounds—chlortetracycline (Aureomycin), oxytetracycline (Terramycin), tetracycline (Achromycin, Panmycin, Polycycline, Steclin, Tetracyn), demethylchlortetracycline (Declomycin), methacycline (Rondomycin), and doxycycline (Vibramycin)—are available for the treatment of human infections. The tetracyclines are synthesized by several species of *Streptomyces* bacteria found in soil. Tetracycline is a semisynthetic drug derived from chlortetracycline.

Effectiveness of tetracyclines

These drugs have a broad spectrum of antibiotic activity, which overlaps that of some of the penicillins, streptomycin, and chloramphenicol. In addition, the tetracyclines also are effective against micro-organisms that resist other drugs. These micro-organisms include the causative agents of typhus fever and Rocky Mountain spotted fever (rickettsiae), amoebiasis (*Entamoeba histolytica*), and one form of pneumonia (*Mycoplasma pneumoniae*), as well as agents that cause various eye infections. There are differences, however, in the effectiveness of the various tetracyclines against bacteria. In general, doxycycline is more active than methacycline, while the latter is more effective than chlortetracycline and demethylchlortetracycline. Tetracycline and oxytetracycline are the least effective.

Although the tetracyclines are absorbed from the stomach and small intestine, absorption is impaired by the presence of milk, milk products, aluminum hydroxide gels, and salts of calcium and magnesium. If a tetracycline produces stomach irritation, therefore, substances other than milk or milk products usually are used to alleviate the problem. Chlortetracycline, oxytetracycline, and tetracycline also are available in forms that can be injected into a vein. The tetracyclines are removed from the blood by the liver, concentrated there, transferred with bile to the intestines, and then absorbed. The tetracyclines are excreted in various proportions in feces and urine; *i.e.*, about 35 percent of a dose of tetracycline is excreted in urine, while about 12 percent of a dose of

chlortetracycline follows this route. The tetracyclines also produce a variety of undesirable effects ranging from hypersensitivity reactions to nausea and liver damage. Oral or parenteral administration of the tetracyclines may lead to the development of suprainfections by bacteria and fungi resistant to them.

Other antibiotics that interfere with protein synthesis. Chloramphenicol (Chloromycetin), a product of *Streptomyces venezuelae* metabolism, also has been synthesized chemically. There is no difference in antimicrobial activity of synthetic and natural products. The drug is effective against a wide range of micro-organisms, including bacteria, rickettsiae, and *Mycoplasma.* Chloramphenicol is rapidly absorbed from the gastrointestinal tract, partially inactivated (about 15 percent of a dose) in the liver, and rapidly excreted in urine. The most serious effect caused by chloramphenicol is depression of blood cell formation in bone marrow. This effect, which is not dose-related and may be an allergic response to the drug, is observed most frequently after prolonged administration. Anemia, caused by depression of red blood cell formation, is a common side effect of chloramphenicol treatment. Since it is a dose-related reaction, anemia usually disappears within a short time after therapy is stopped. Other side effects caused by chloramphenicol include skin rashes, diarrhea, blurred vision, and suprainfection.

Erythromycin

Erythromycin, a product of a soil organism called *Streptomyces erythraeus*, may be either bacteriostatic or bactericidal, depending on the organism and the drug concentration. Among the micro-organisms against which it is effective are *Staphylococcus aureus*, several species of *Streptococcus*, pneumococcus, and *Mycoplasma pneumoniae*. Preparations are available for injection into veins or oral administration. The antibiotic is concentrated in the liver and transferred in active form with bile to the intestines and partially absorbed. When large doses are given by mouth, a considerable quantity of the drug is excreted in the feces. Erythromycin administration is accompanied by a low risk of undesirable effects (*e.g.*, skin rashes, fever).

Lincomycin (Lincocin), a product of *Streptomyces lincolnensis*, differs chemically from other antibiotics; it is a member of a series of antimicrobial substances whose structure contains an amino acid and an amino sugar (called an octose) that are joined by an amide bond. Its spectrum of activity resembles that of erythromycin. Urinary excretion of lincomycin is limited and varies considerably, depending on whether the drug is taken orally or injected; significant quantities of lincomycin, however, are eliminated in the feces. Diarrhea is common in patients treated orally with lincomycin. Among other reactions are nausea, vomiting, abdominal pain, skin rashes, generalized itching, and reduction in the number of white blood cells.

Novobiocin, an antibiotic derived from *Streptomyces niveus* and *S. spheroides*, is both bacteriostatic and bactericidal; its antimicrobial spectrum is similar to that of penicillin G and erythromycin. *Staphylococcus aureus* and pneumococcus are the most sensitive bacteria, but several species of meningococcus, *Streptococcus*, and *Haemophilus*, including the organism that causes whooping cough, are inhibited by the drug. Novobiocin is well absorbed from the gastrointestinal tract and is given either by mouth or by injection. Most of the antibiotic is present in the feces after oral administration. The incidence of reactions to novobiocin is relatively high; the most common reactions are extensive skin rashes and fever. Yellow discoloration of skin, blood serum, and eyes, usually caused by a product of the metabolism of the antibiotic, also is a common side effect. Other side effects include nausea, vomiting, diarrhea, liver damage, loss of hair, and intestinal hemorrhage.

Streptomycin

Streptomycin, the first antimicrobial agent developed after penicillin, is synthesized by a soil organism called *Streptomyces griseus*. The active drug consists of three compounds, streptidine, streptose, and N-methyl-L-glucosamine. Among the organisms sensitive to the antibiotic are pneumococcus, *Salmonella* (including the typhoid bacillus), meningococcus, and some strains of *Staphylococcus*. Streptomycin was the first antibiotic effective in treating tuberculosis. Many initially sensitive micro-organisms have become resistant to streptomycin probably because it has mutagenic ability; *i.e.*, it increases the chances for changes or mutations in the hereditary material of cells. Since it is not absorbed from the gastrointestinal tract, streptomycin must be injected. Approximately 50 percent of an injected dose is excreted in unchanged form in urine; smaller amounts are transferred from the liver to the intestines with bile and eliminated in feces. Streptomycin causes many side effects; *e.g.*, various types of skin rashes, fever, anemia, nausea, vomiting, deafness, kidney irritation.

Gentamicin (Garamycin), a broad-spectrum antibiotic derived from *Micromonospora purpurea*, consists of three closely related components called gentamicin C_1, C_2, and C_{1a}. Among the micro-organisms against which this bacteriostatic and bactericidal antibiotic are effective are *Staphylococcus aureus*, pneumococcus, *Pseudomonas*, several *Streptococcus* species, the tubercle bacillus, and *Mycoplasma pneumoniae*. Since insignificant amounts of gentamicin are absorbed from the gastrointestinal tract, intramuscular injection is used to produce therapeutic concentrations in the blood. Although it is excreted by the kidneys, only about 40 percent of a dose of gentamicin is recovered in the urine after several days, suggesting that the drug accumulates in the body. After a few days, however, the amount of the drug excreted in urine approximates the dose administered. Among the reactions to gentamicin are nausea, vomiting, headache, and skin rashes. The drug also occasionally causes ear damage, which tends to involve the portion of the auditory nerve that controls balance.

Kanamycin (Kantrex), produced by the soil organism *Streptomyces kanamyceticus*, has a broad spectrum of antimicrobial activity. Sensitive organisms include *Salmonella*, *Shigella*, the organism that causes brucellosis (*Brucella*), *Staphylococcus aureus*, and the tubercle bacillus. Since it is poorly absorbed from the gastrointestinal tract, most of an ingested dose of kanamycin is eliminated in feces; to achieve a therapeutic effect, therefore, the drug is injected. Kanamycin is excreted primarily by the kidneys. The two most important reactions to kanamycin are damage to the auditory nerve and the kidney. Other undersirable effects include various skin rashes, itching, and diarrhea.

Neomycin

Neomycin is synthesized by the soil organism *Streptomyces fradiae*. The drug has a broad spectrum of activity and is highly effective against the causative agents of one form of dysentery (*Shigella*), food poisoning (*Salmonella* species), diphtheria (*Corynebacterium diphtheriae*), tuberculosis (*Mycobacterium tuberculosis*), anthrax (*Bacillus anthracis*), Weil's disease (*Leptospira*), and *Staphylococcus* infections. Neomycin usually is administered by injection. When administered by mouth, which may occur in patients undergoing intestinal surgery, only about 3 percent is absorbed into the bloodstream. Kidney damage and deafness occur in a significant number of people treated with neomycin. Individuals who receive neomycin orally for prolonged periods may have large quantities of fat in the feces.

Antibiotics that affect cell membranes. The antibiotics that produce their antimicrobial effects by acting on cell membranes can be divided into two groups. Polymyxin and colistin are surface-active agents (detergents) and thus can become oriented between lipid (fat) and protein films; such action may disrupt lipoprotein-containing bacterial membranes so that they no longer function as effective barriers and allow cell contents to escape. The antifungal activity of nystatin and amphotericin B is dependent on their binding to a compound that is present only in the cell membranes of sensitive yeasts and fungi. Interaction between the compound, a sterol, and the drugs causes changes in the cell membrane that result in leakage of cell contents and death.

Polymyxin actually refers to a group of closely related antibiotic substances, polymyxins A,B,C,D,E, which are

Polymyxin and colistin

synthesized by a soil micro-organism, *Bacillus polymyxa*. The compound used to treat human infections, polymyxin B, is active against several bacteria (*e.g.*, *Salmonella*, *Shigella*) and is especially useful in treating infections caused by *Pseudomonas aeruginosa*. Polymyxin is injected into muscles or veins. Elimination from the kidneys continues for a few days after administration of the antibiotic is stopped. Among the many reactions that may be caused by polymyxin are skin rashes, itching, dizziness, speech abnormalities, and fever.

Colistin (sodium colistimethate) is produced by *Aerobacillus colistinus*. Although its exact chemical structure is not yet known, it may be identical with, or closely related to, polymyxin E. The compound used to treat infections is sodium colistimethate; it has a range of antimicrobial activity similar to that of polymyxin B and is used primarily to treat infections caused by *Pseudomonas aeruginosa*. The reactions produced by colistin are similar to those associated with the use of polymyxin.

Nystatin (Mycostatin), a product of the soil organism *Streptomyces noursei*, is active only against certain yeasts and fungi. The drug, used primarily for topical application, is not absorbed into the body from the skin and mucous membranes to which it is applied locally. Reactions to topical applications of nystatin are rare.

Streptomyces nodosus, a soil organism, is the source of amphotericin B (Fungizone). Ineffective against bacteria, rickettsiae, and viruses, amphotericin B is used to treat various yeast and fungal diseases; *e.g.*, histoplasmosis, North American blastomycosis, sporotrichosis, and coccidiodomycosis. Injection of amphotericin B into a vein often is accompanied by chills, fever, vomiting, and headache. Other reactions may be more serious; *e.g.*, convulsions, anemia, liver damage, kidney damage.

Antibiotics that affect nucleic acid metabolism. One known antibiotic, griseofulvin, exerts its antimicrobial effects by altering nucleic acid metabolism. It increases the rate of synthesis of deoxyribonucleic acid (DNA) and increases the incorporation of carbon into ribonucleic acid (RNA) in fungi. Griseofulvin (Fulvicin, Grifulvin, Grisactin), produced by *Penicillium griseofulvum* and *Penicillium janczewski*, is used to treat ringworm (*Microsporum*), athlete's foot (*Epidermophyton*), and diseases of the hair and nails (*Trichophyton*). Griseofulvin is ineffective against diseases caused by any other fungi, yeasts, or bacteria.

Griseofulvin is always administered by mouth, and most of the drug is excreted in unchanged form in the feces. Interestingly, the drug has more affinity for diseased skin than for normal skin. Few serious reactions occur after griseofulvin treatment.

BIBLIOGRAPHY. W. CARTER and K.S. MCCARTY, "Molecular Mechanisms of Antibiotic Action," *Ann. Intern. Med.*, 64: 1087–1113 (1966), a review of current knowledge of the mechanism of action of some commonly used antibiotics, with details of biochemical alterations they produce in bacteria; A. FLEMING, "History and Development of Penicillin," *Penicillin: Its Practical Application* (1946), an interesting account by the discoverer of penicillin; H.W. FLOREY, "Historical Introduction," *Antibiotics*, vol. 1, ch. 1 (1949), a review of the development of the concept of antibiosis and its practical application; J.P. UTZ, "Antimicrobial Therapy in Systemic Fungal Infections," *Am. J. Med.*, 39:826–830 (1965), a short summary of the treatment of infections produced by yeasts and fungi; L. WEINSTEIN and N.J. EHRENKRANZ, *Streptomycin and Dihydrostreptomycin* (1958), a monograph describing the chemistry, antibacterial effects, pharmacology, reactions, and clinical uses of streptomycin; L. WEINSTEIN and K. KAPLAN, "The Cephalosporins and Their Semi-synthetic Derivatives," *Ann. Intern. Med.*, 72:729–739 (1970), a review of the historical development, microbiological behaviour, pharmacology, effects, and clinical applications; L. WEINSTEIN, "Antibiotics," *The Pharmacological Basis of Therapeutics*, 4th ed., ch. 55–61 (1970), a detailed review of antimicrobial agents, including their history, chemistry, microbial activity, pharmacology, toxicology, and clinical uses; B. WEISBLUM and J. DAVIES, "Antibiotic Inhibitors of Bacterial Ribosomes," *Bact. Rev.*, 32:493–528 (1968), a detailed description of the mechanism of action of the antibiotics that suppress, inhibit, or alter bacterial protein synthesis.

(L.W.)

Antigonus I Monophthalmus

Antigonus I Cyclops, or Monophthalmus (the "One-Eyed"), 382–301 BC, was a Macedonian general under Alexander the Great and, after the latter's death in 323 BC, the most famous among his successors (*diadochoi*). He was the founder of the Antigonid dynasty, which ruled in Macedonia until 168 BC. From 323–301 he reigned (king from 306) over large parts of Asia Minor, laying claim to the whole of Alexander's empire.

Military campaigns

In 333 Alexander had appointed him satrap of Phrygia, and upon Alexander's death he also received the governorship of Pamphylia and Lycia from the regent Perdiccas. He then formed an alliance against Perdiccas with Antipater, the governor of Macedonia, Ptolemy of Egypt, and Lysimachus of Thrace and Craterus (all of whom had served under Alexander). Perdiccas was murdered, and Antipater became regent. In 321 Antipater appointed Antigonus commander in chief of his army in Asia and sent him against Eumenes, the satrap of Cappadocia and an adherent of Perdiccas. Antigonus defeated Eumenes and then besieged him unsuccessfully in the mountain fortress Nora. Polyperchon succeeded Antipater as regent, and Antigonus joined forces against him with Cassander (Antipater's son), Ptolemy, Lysimachus, and Eumenes in 319. When Eumenes, his rival in Asia Minor (Anatolia), went over to Polyperchon, he defeated him with the aid of Seleucus and Peithon, the satraps of Babylonia and Media, at Gabiene. Then, wishing to eliminate all possible rivals, he had both Eumenes and Peithon executed; Seleucus managed to escape to Egypt.

Antigonus was now in complete control of Asia Minor, but Ptolemy, Lysimachus, Cassander, and Seleucus allied themselves against him in the first coalition war (315–311) in an attempt to thwart his plan of reuniting Alexander's empire. Antigonus occupied Syria and proclaimed himself regent. In order to win the support of the Greek city-states, whose resistance to subjugation presented the chief stumbling block to the formation of a Hellenistic monarchy, he announced to his assembled army that all the Greeks should be free, autonomous, and ungarrisoned. This political slogan was to be sounded again and again—almost immediately by Ptolemy and for a final time by the Romans in 196. With the aid of his officers in Greece, Antigonus drove out Cassander's Macedonian forces of occupation there and formed the island cities in the Aegean into the League of the Islanders, preparatory to his invasion of Greece. His ally, the city of Rhodes, furnished him with the necessary fleet.

While he was engaged in conquering Caria, his son, Demetrius Poliorcetes, was defeated at Gaza by Ptolemy and Seleucus (312). Seleucus returned to his former province, Babylonia. In view of this new threat from the East, Antigonus decided to make peace with all of his adversaries except Seleucus. All of the *diadochoi* confirmed the existing boundaries and the freedom of the Greek cities. Antigonus, no longer regent but merely *stratēgos* (officer in charge) of the whole of Asia, was to rule in Syria and from the Hellespont to the Euphrates.

Activities in Greece

Then Ptolemy attacked Cilicia, and the second coalition war (310–301) against Antigonus broke out. In Greece in 307, Antigonus' son Demetrius ousted Demetrius of Phaleron, Cassander's governor of Athens, and re-established the old Athenian constitution. The grateful Athenians honoured Antigonus and Demetrius as divine saviours (*theoi sōtēres*). Cassander's influence in Greece was now broken, and in 306 Demetrius defeated Ptolemy's fleet near Salamis on the island of Cyprus and conquered the island. This victory gave Antigonus control of the Aegean, of the eastern Mediterranean, and of all of the Near East except Babylonia. The assembled army proclaimed him king, and his friends adorned him with the diadem. For his part, he appointed Demetrius king and coregent and sent him the diadem. This was to become a traditional ceremony in the Hellenistic monarchy.

In 305, after Antigonus had vainly attacked Egypt, Ptolemy also assumed the title of king, and Cassander, Lysimachus, and Seleucus followed suit. The partition of

Alexander's empire into five states had now been formally established. In 305 Antigonus sent Demetrius to conquer Rhodes, which had refused him armed support against Ptolemy. After a year's unsuccessful siege he concluded a peace treaty and an alliance with the island state, guaranteeing it autonomy and neutrality in his conflicts with Ptolemy. This concession was necessary because in the meantime Cassander had invaded Attica and was besieging Athens. Demetrius drove him out of central Greece, and the Athenians bestowed on him a new religious honour, *synnaos* ("having the same temple") of the temple of the goddess Athena.

In 303 he occupied Corinth, Sicyon, and Argos in the Peloponnese, and Achaea, Elis, and almost all of Arcadia joined his side. In 302 Antigonus and Demetrius crowned their success by renewing the pan-Hellenic league, which Philip II of Macedonia had formed in 338. Ambassadors from all the Hellenic states—with the exception of Sparta, Messenia, and Thessaly—elected Antigonus and Demetrius protectors of the new league at Corinth. It was to be an "eternal" treaty, extending to the descendants of the kings. Each member state furnished a contingent of troops for a league army that was commanded by the kings or their deputies. The league was to ensure a general peace in Hellas, but first and foremost it was to aid Antigonus against Cassander.

Final campaigns

Now at the zenith of his power, Antigonus demanded Cassander's unconditional submission. He wanted possession of Macedonia, the native land of his dynasty, and to establish his dominion over Alexander's former empire. The other *diadochoi*, however, warned by Cassander's fate, now joined forces to attack the omnivorous old man. From Babylonia Seleucus invaded Asia Minor, Ptolemy attacked Syria, and Lysimachus moved into the western part of Asia Minor. Docimus, the regent of Phrygia, and Phoenix, the *stratēgos* of Lycia, deserted Antigonus. He, in turn, recalled Demetrius, left his capital city Antigoneia (which he had founded on the Orontes in 306), and crossed the Taurus Mountains. Lysimachus, who was waiting for Seleucus, avoided an engagement. In vain Antigonus sent a corps of raiders into Babylonia in order to divide his enemies' forces. In 301 the united armies of Lysimachus and Seleucus engaged the forces of Antigonus and Demetrius at Ipsus in Phrygia. Demetrius made the error of pursuing the enemy's cavalry too far, and as a result Antigonus, age 80, lost the battle and his life.

Antigonus had been an excellent strategist who, until then, had never lost a battle. He had a genuine admiration for Greek civilization. He founded several cities, especially in Asia Minor, and united several small communities into unitary, large centres; Lebedus (Lebedos) and Teos, for example. Several Greek artists graced his court; Apelles painted his portrait in profile because of the missing eye (the cause of which is unknown).

BIBLIOGRAPHY. C. WEHRLI, *Antigone et Demetrios* (1968), a comprehensive biography with a discussion of sources and literature; H.H. SCHMITT, *Die Verträge der griechisch-römischen Welt von 338 bis 200 v. Chr.* (1969), treats the foundation of the Hellenic union; R.H. SIMPSON, "Antigonus the One-Eyed and the Greeks," *Historia*, 8:385–409 (1959), discusses the policy of Antigonus toward the Greek states; C. HABICHT, *Gottmenschentum und griechische Städte* (1956), itemizes the cult honours of Antigonus.

(H.V.)

Antigonus II Gonatas

Antigonus II Gonatas (the meaning of his surname is unknown), the son of Demetrius I Poliorcetes and grandson of Antigonus I, was king of Macedonia from 276 BC. He rebuilt the power of Macedonia and safeguarded Macedonian interests by keeping Greece divided without attempting to conquer the whole country. While his father Demetrius was busy fighting in Macedonia and Asia Minor against Lysimachus of Thrace, Ptolemy I of Egypt, Pyrrhus of Epirus, and Seleucus I, Antigonus, as his regent, was engaged in maintaining their hegemony in Greece, which had been achieved in 287 BC. When his father was taken prisoner by Seleucus in 285, Antigonus assumed the title of king on the death of Demetrius in 283 but did not count the beginning of his reign until 276. Although he had only a few bases in Greece, Antigonus laid claim to Macedonia when Seleucus was murdered in 281. His claim was disputed by Seleucus' successor, Antiochus I. Antigonus took part in the defense of Greece against the invading Celts (279). In the following year he concluded a peace with Antiochus. This surrendered his claim to Macedonia. Thereafter Antigonus' foreign policy was marked by friendship with the Seleucids.

In 277 he crossed the Hellespont and defeated the Celts near Lysimacheia. After this success he was acknowledged king by the Macedonians in 276. Pyrrhus, returning in 274 after the failure of his campaign in Italy, drove Antigonus out of Upper Macedonia and Thessaly. Although he retained only a few Macedonian cities, Antigonus followed Pyrrhus when the latter marched into the Peloponnese; and when Pyrrhus died in Argos in 272, Antigonus' control over Macedonia was assured. He was now also the chief of the Thessalian League and on good terms with neighbouring Illyria and Thrace. He secured his position in Greece by keeping Macedonian occupation forces in the cities of Corinth, Chalcis on Euboea, and Demetrias in Thessaly, the three "shackles" of Hellas.

Beyond that he supported the pro-Macedonian faction in various cities in the Peloponnese and the rise to power of tyrants in Sicyon, Argos, Elis, and Megalopolis. In order to keep Greece in a state of complete dependency by controlling the straits and the supply of grain from the southern Russian region, Macedonia—its vigour restored—needed only to gain mastery over the Aegean Sea. To avert this danger, King Areus of Sparta and the city of Athens—urged on by Ptolemy II of Egypt—declared a war for the liberation of Greece (the Chremonidean War, 267–261). Although the Egyptian fleet had blockaded the Saronic Gulf, Antigonus defeated the king of Sparta near Corinth in 265 and then besieged Athens. In 263–262 the city capitulated. Athenian officials were replaced by Antigonus' appointees, and Athens became no more than a Macedonian provincial city.

The Chremonidean War

Immediately after the Chremonidean War, Antigonus joined forces with the Seleucid Antiochus II against their common enemy, Ptolemy II. Whether his naval victory of Cos, which secured Antigonus the Aegean Sea and the League of the Islanders, belongs to this (255) or to the Chremonidean War (261) is uncertain. In 255 a peace was concluded with Ptolemy, and by marrying his stepbrother Demetrius the Fair to Berenice of Cyrene, Antigonus established Macedonian influence in this neighbour country of Egypt.

But his position in Greece was now shaken by a number of reversals. In 253 Alexander, Antigonus' nephew and regent, revolted in Corinth with Ptolemy's help and declared himself an independent monarch. Antigonus lost Corinth and Chalcis, the two bases from which he dominated southern Greece. As the Aetolians had occupied Thermopylae, he was cut off from Athens and the Peloponnese. After Alexander's death, however, Antigonus gave Nicaea, Alexander's widow, to his son Demetrius in marriage and by means of a strategem regained Corinth in 244. In the meantime the Achaean League was becoming a dangerous opponent. Since 251 it had been under the leadership of Aratus of Sicyon and was receiving financial aid from Ptolemy II. In vain, Antigonus sent gifts to win over Aratus. In 243, without a declaration of hostilities, Aratus made a surprise attack on Corinth and forced the withdrawal of the Macedonian occupation troops. Megara, Troezen, and Epidaurus also deserted Antigonus. He made no attempt to regain these territories but instead formed an alliance with the Aetolian League, which made unsuccessful raids of pillage into the Peloponnese. Nevertheless, by defeating the Egyptian fleet at Andros *c.* 244 Antigonus was able to maintain his hegemony in the Aegean. After a life of endless warfare he died in 239 at 80 years of age.

Personally, Antigonus was unassuming, short of stature, and snub-nosed. In Macedonia the cult of the ruler, so

Antigonus and his court

usual in the other Hellenistic states, was unknown. He chose his friends not because of their noble ancestry but for their personal abilities. He conceived his monarchic rule in philosophical terms—*i.e.*, by the strict observance of his duties as a ruler. Once, when his son treated some subjects arbitrarily, he said to him: "Do you not understand that our kingship is a noble servitude [*endoxos douleia*]?" This paradoxical concept of monarchy envisaged the ruler as bearing the burden of his office, serving the people and the law. In his youth Antigonus had been a student of Zeno, the founder of Stoicism. He had been taught by him in Athens and in 276 invited him to his court in Pella in Macedonia. The philosopher, however, did not come and instead sent two of his students, Persaeus and the Theban Philonides. Persaeus wrote a treatise on kingship, was the mentor of Halcyoneus, the son of Antigonus, and became commandant of Corinth in 244. When Zeno died in 263 the King lamented that he had lost the only man whose judgment of his public actions he valued, and he prevailed upon the Athenians to bury him in state. Among the literati at his court was the historian Hieronymus of Cardia, who recorded the war with Pyrrhus, and the poet Aratus, a native of Cilicia, author of the much read didactic poem on astronomy, *Phaenomena*.

BIBLIOGRAPHY. W.W. TARN, *Antigonos Gonatas* (1913, reprinted 1969), a comprehensive biography; C.F. EDSON, "The Personal Appearance of Antigonus Gonatas," *Classical Philology*, 29:254–255 (1934), describes his physical deformities; M. CHAMBERS, "The First Regnal Year of Antigonus," *Am. J. Philology*, 75:385–394 (1954), provides the dating for his reign; H. VOLKMANN, *Historia*, 16:155–161 (1967), comments upon the notion of noble servitude (in German).

(H.V.)

Antioch

Antioch (Turkish Antakya), a populous city of ancient Syria, and since 1939 the chief town of the *il* (province) of Hatay in southeastern Turkey, is near the mouth of the Orontes River (Turkish Asi Nehri).

History. Antioch was founded (300 BC) by Seleucus I Nicator, a former general of Alexander the Great, as a centre for military control and for the diffusion of Hellenic culture in Syria (see SYRIA AND PALESTINE, HISTORY OF). The city's strategic location, commanding both the east–west and north–south military and commercial communications in northwestern Syria, contributed to its growth and prosperity in Hellenistic, Roman, and early Byzantine times. The beautiful suburb of Daphne, five miles to the south, was a favourite pleasure resort and suburban residence for Antioch's upper classes; and the seaport Seleucia Pieria, at the mouth of the Orontes, was the city's harbour. Antioch was the centre of the Seleucid Kingdom until 64 BC, when it was annexed by Rome and made the capital of the province of Syria.

The Roman city

In addition to serving as the western terminus of the caravan routes over which goods were brought from Persia and the Far East, Antioch was the headquarters of the Roman garrison of Syria, one of whose principal duties was the defense of the eastern border from Persian attacks. During the military anarchy of the 3rd century AD, the city was captured by the Sasanian King Shāpūr I (*c.* AD 260) and for a time (AD 261–272) came under the control of the independent kingdom of Palmyra.

Antioch was one of the earliest centres of Christianity; it was there (*c.* AD 40) that the followers of Christ were first called Christians. It served as the headquarters for the missionary journeys of St. Paul (*c.* 47–55) and furnished a number of early martyrs.

The 4th century was a period of prosperity and political importance. In addition to being the headquarters of the governor of Syria, Antioch was the seat of the Count of the East, a new official administering the provinces from Mesopotamia and Syria on the north to Egypt on the south. Because the church of Antioch had the distinction of having been founded by the Apostles Peter and Paul, its bishop ranked with the bishops of the other apostolic foundations, Jerusalem, Rome, and Alexandria (Constantinople was accepted in this category later). Thus the bishops of Antioch were influential in theology and in ecclesiastical politics.

Antioch served as the headquarters of the emperor Julian the Philosopher (or "the Apostate") during his brief reign (361–363) when he attempted, unsuccessfully, to eradicate Christianity and restore pagan religion. The life of the city during the last half of the 4th century is well known from the writings of the pagan teacher and man of letters Libanius and from the sermons and treatises of the gifted preacher St. John Chrysostom, who was born there.

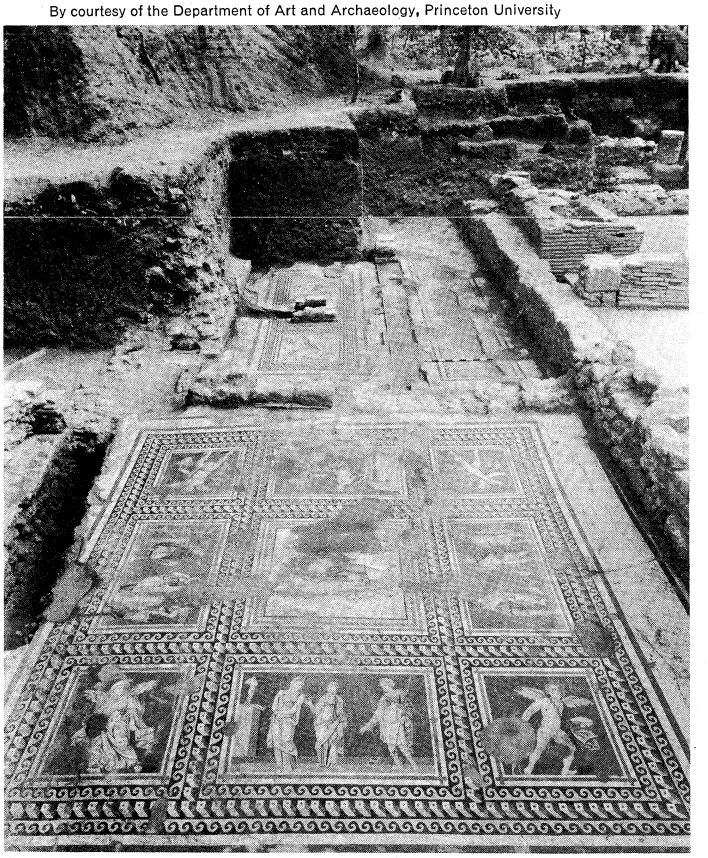
Commercial importance

The 4th and 5th centuries brought prosperity from the olive oil produced on the plantations in the region east of Antioch; many of these were owned by wealthy residents of the city who exported the oil. Later, when the silkworm was brought from the East in the reign of Justinian I (527–565), mulberry trees for the feeding of the worms were grown in the same area, and Antioch developed a flourishing silk industry.

In the 5th and 6th centuries Antioch, with the remainder of Syria, was heavily involved in the doctrinal and political disputes of the Monophysite controversy. Syria took the Monophysite part, which was opposed to the orthodox theology of the government in Constantinople, and there were riots and bloodshed in the city when imperial officials attempted to enforce orthodox doctrine on the unwilling population.

The 6th century brought an unparalleled series of disasters from which the city never fully recovered. A fire in 525 was followed by earthquakes in 526 and 528. Captured temporarily by the Persians in 540 and 611, Antioch was absorbed into the Arab Caliphate in 637. Under the Arabs, it shrank to the status of a small town. Because of its location, however, it served as a strongpoint in the border wars that followed. The Byzantines recaptured the city in 969, and it served as a frontier fortification until taken by the Seljuq Turks in 1084. In 1098 it was captured by the crusaders and became the capital of

By courtesy of the Department of Art and Archaeology, Princeton University



The House of the Red Pavement, Antioch, showing rooms 1 to 3, *c.* 5th century. Mosaic floor depicts the Four Seasons (corner panels) and mythological scenes.

the Frankish principality of Antioch. Taken by the Mamlūks in 1268 and by the Ottoman Turks in 1517, Antioch remained under Ottoman control until World War I.

The present town is a small Turkish provincial capital (population of 57,855 in 1965). Its activities are based mainly on the agricultural produce of the adjacent area, including the intensively cultivated Amik plain. The chief crops are wheat, cotton, grapes, rice, olives, vegetables, and fruit; the town possesses soap and olive-oil factories and cotton ginning and other processing industries. Silk, shoes, and knives are also manufactured.

Archaeology. Excavations were conducted at Antioch from 1932 to 1939 under the auspices of Princeton University, the Musées Nationaux de France, and a group of U.S. museums. Valuable information was obtained concerning the topography and history of the city. A major contribution was the discovery of a large number of fine mosaic floors, excavated in private houses and public buildings, which furnished unexpected material that made necessary new assessments of ancient art. Dating largely from the Roman imperial period, many of the floors represent copies of paintings which otherwise would have been unknown. In addition, the mosaics provide new illustrations of contemporary life. Unique in the history of ancient art is the topographical mosaic found in a villa at Daphne, dated in the later part of the 5th century AD, which illustrates a tour of Antioch such as might have been taken by an ancient visitor. The mosaic depicts the famous colonnaded streets, private houses, and public buildings, with people shown in a variety of activities. Discoveries of silverware illustrate the city's reputation in antiquity as a centre of silversmith's and goldsmith's work.

The "Chalice of Antioch"

The celebrated "Chalice of Antioch," found near Antioch in 1910 and once thought to be the chalice used at the Last Supper, is dated by art historians in the 4th or 5th century AD.

BIBLIOGRAPHY. G. DOWNEY, *A History of Antioch in Syria: From Seleucus to the Arab Conquest* (1961), a comprehensive history, based on the original sources and utilizing the results of the excavations, illustrated with maps, plans, and photographs, and *Ancient Antioch* (1963), a shorter version of the foregoing, intended for the general reader, with additional material on the intellectual life of the city and fuller illustration; G. HADDAD, *Aspects of Social Life in Antioch in the Hellenistic-Roman Period* (1949), a scholarly study based on the original texts; R.A. PACK, *Studies in Libanius and Antiochene Society under Theodosius* (1935), a study based on contemporary texts of social and political problems during a period of the city's prosperity; G. DOWNEY, *Antioch in the Age of Theodosius the Great* (1962), an essay that attempts to re-create all aspects of the life of the city during a period for which ample evidence is preserved; G.W. ELDERKIN *et al.* (eds.), *Antioch-on-the-Orontes*, vol. 1–4 (1934–52), the reports on the excavations, with some studies of the objects found; D. LEVI, *Antioch Mosaic Pavements*, 2 vol. (1947), a complete collection of the mosaics, with scholarly commentary; C.R. MOREY, *The Mosaics of Antioch* (1938), a study of the mosaics and their significance for the history of ancient art; J.J. RORIMER, "The Authenticity of the Chalice of Antioch," in *Studies in Art and Literature for Belle Da Costa Greene*, ed. by D.E. MINER, pp. 161–168 (1954), an account of the history of the chalice and of the scholarly studies of its date. For the history of Antioch during the period of the Crusades, see *A History of the Crusades*, ed. by K.M. SETTON, vol. 1–2 (1957–62), references to Antioch given in the indexes.

(G.D.)

Antiochus III the Great

Antiochus III the Great was, after Seleucus I, the most successful of the Seleucid kings of the Hellenistic Syrian empire. Though he failed in his wars against Rome, he rebuilt the Syrian empire in the east. He reformed the empire administratively by reducing the provinces in size, established a ruler cult (with himself and his consort Laodice as divine), and improved relations with neighbouring countries by giving his daughters in marriage to their princes.

Born in 242 BC, the son of Seleucus II, Antiochus succeeded his brother Seleucus III as king in 223 BC. He retained from the previous administration Hermias as chief minister, Achaeus as governor of Asia Minor, and Molon and his brother Alexander as governors of the eastern provinces, Media and Persis. In the following year, when Molon rebelled and assumed the title of king, Antiochus abandoned a campaign against Egypt for the conquest of southern Syria, on the advice of Hermias, and marched against Molon, defeating him in 220 BC on the far bank of the Tigris and also conquering Atropatene, the northwestern part of Media. Shortly thereafter he had Hermias killed and was thus rid of most of the influences from the previous administration. In the same year, Achaeus set himself up as king in Asia Minor, but a mutiny in his army kept him from attacking Antiochus.



Antiochus III, coin, late 3rd–early 2nd century BC. In the British Museum.
By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.

Wars of expansion

Antiochus was now free to conduct what has been called the Fourth Syrian War (219–216), during which he gained control of the important eastern Mediterranean sea ports of Seleucia-in-Pieria, Tyre, and Ptolemais. In 218 he held Coele Syria (Lebanon), Palestine, and Phoenicia. In 217 he engaged an army (numbering 75,000) of Ptolemy IV Philopator, a pharaoh of the Hellenistic dynasty ruling Egypt, at Raphia, the southernmost city in Syria. His own troops numbered 68,000. Though he succeeded in routing the left wing of the Egyptian army, his phalanx (heavily armed infantry in close ranks) in the centre was defeated by a newly formed Egyptian phalanx. In the subsequent peace settlement, Antiochus gave up all his conquests except the city of Seleucia-in-Pieria.

After the Syrian war, he proceeded against the rebel Achaeus. In alliance with Attalus I of Pergamum, Antiochus captured Achaeus in 213 in his capital Sardis and had him executed in a barbaric manner. After the pacification of Asia Minor he entered upon his later to be famous eastward campaign (212–205), pressing forward as far as India. In 212 he gave his sister Antiochis in marriage to King Xerxes of Armenia, who acknowledged his suzerainty and paid him tribute. He occupied Hecatompylos (southeast of the Caspian Sea), the capital of the Parthian king Arsaces III, and forced him to enter into an alliance in 209 and the following year defeated Euthydemus of Bactria, though he allowed him to continue to rule and retain his royal title. In 206 he marched across the Hindu Kush into the Kābul valley and renewed a friendship with the Indian king Sophagasenos.

Returning westward via the Iranian provinces of Arachosia, Drangiana, and Carmania, he arrived in Persis in 205 and received tribute of 500 talents of silver from the citizens of Gerrha, a mercantile state on the east coast of the Persian Gulf. Having established a magnificent system of vassal states in the east, Antiochus now adopted the ancient Achaemenid title of "great king," and the Greeks, comparing him to Alexander the Great, surnamed him also "the Great."

After the death of Ptolemy IV, Antiochus concluded a secret treaty with Philip V, ruler of the Hellenistic kingdom of Macedonia, in which the two plotted the division of the Ptolemaic empire outside Egypt. Antiochus' share was to be southern Syria, Lycia, Cilicia, and Cyprus; Philip was to have western Asia Minor and the Cyclades. Antiochus invaded Coele Syria, defeated the Ptolemaic general Scopas at Panion near the source of the Jordan River in the year 200, gained control of Palestine, and

granted special rights to the Jewish temple state. But Philip, marching along the Dardanelles, became involved in a war with Rhodes and Pergamum, both of whom appealed to Rome for help against Macedonia, informing Rome of the alliance between the two Hellenistic kings. Rome intervened decisively in the system of Hellenistic states. Philip was defeated by the Romans in the Second Macedonian War (200–196), and Antiochus refused to help him. Instead, taking advantage of the Romans' involvement with Philip, Antiochus marched against Egypt. Though the Romans had sent ambassadors to Ptolemy V, they could not lend him any serious assistance. When peace was concluded in 195, Antiochus came permanently into possession of southern Syria—which had been fought over for 100 years by the Ptolemies and Seleucids—and of the Egyptian territories in Asia Minor. He also gave his daughter Cleopatra in marriage to Ptolemy V. Egypt practically became a Seleucid protectorate.

War with Rome

In his insatiable expansionist drive, Antiochus occupied parts of the kingdom of Pergamum in 198 and in 197 Greek cities in Asia Minor. In 196 BC he crossed the Hellespont into Thrace, where he claimed sovereignty over territory that had been won by Seleucus I in the year 281 BC. A war of harassment and diplomacy with Rome ensued. A number of times the Romans sent ambassadors demanding that Antiochus stay out of Europe and set free all the autonomous communities in Asia Minor. To meet these demands would have meant the actual dissolution of the western part of the Seleucid empire, and Antiochus thus refused. Tensions with Rome increased further when the great Carthaginian general Hannibal, who had fled from Carthage in the aftermath of defeat by the Romans in the Second Punic War, found refuge with Antiochus in 195 BC and became his adviser. Antiochus offered an alliance to Philip of Macedonia, whom he had previously forsaken, but was rebuffed. Philip, Rhodes, Pergamum, and the Achaean League joined Rome. Only the Aetolians, discontent with Rome's growing influence in Greece, called upon Antiochus to be their liberator and appointed him commander in chief of their league. Relying on them Antiochus landed in Demetrias in the autumn of 192 with only 10,500 men and occupied Euboea. But he found little support in central Greece. In 191 the Romans, numbering more than 20,000, cut him off from his reinforcements in Thrace and outflanked his position at the pass of Thermopylae (in Greece). With the remainder of his troops Antiochus fled to Chalcis on Euboea and from there by sea to Ephesus; his fleet was wiped out by the combined naval forces of Rome, Rhodes, and Pergamum. Meeting no resistance, the Roman army crossed the Hellespont in 190. Antiochus was now eager to negotiate on the basis of Rome's previous demands, but the Romans insisted that he first evacuate the region west of the Taurus mountains. When Antiochus refused, he was decisively defeated in the Battle of Magnesia near Mt. Sipylus, where he fought with a heterogeneous army of 70,000 men against an army of 30,000 Romans and their allies. Although he could have continued the war in the eastern provinces, he renounced all claim to his conquests in Europe and in Asia Minor west of the Taurus at the peace Treaty of Apamea. He also was obliged to pay an indemnity of 15,000 talents over a period of 12 years, surrender his elephants and his fleet, and furnish hostages, including his son Antiochus IV. His kingdom was now reduced to Syria, Mesopotamia, and western Iran. In 187 Antiochus was murdered in a Baal temple near Susa, where he was exacting tribute in order to obtain much needed revenue.

BIBLIOGRAPHY. W.W. TARN, *The Cambridge Ancient History*, 7:723–726 (1928); E. WILL, *Histoire politique du monde hellénistique*, vol. 2 (1967), short biographical sketches; J. SEIBERT, *Historica-Einzelschrift*, 10:60 (1967), explains Antiochus' politics in marriage; H.H. SCHMITT, *Untersuchungen zur Geschichte Antiochos des Grossen und seiner Zeit* (1964), treats the first 25 years of his reign and his policies; E. BADIAN, "Rome and Antiochus the Great: A Study in Cold War," *Classical Philology*, 54:81–99 (1959), describes the negotiations between Rome and Antiochus before the war of 192–188 BC; H. BENGTSON, *Die Strategie in der hellenistischen Zeit*, vol. 2 (1944), delineates his administrative reforms; G.M.A. RICHTER, *The Portraits of the Greeks*, vol. 3 (1965), discusses the preserved portrait of Antiochus.

(H.V.)

Antiochus IV Epiphanes

Antiochus IV Epiphanes (meaning "god manifest"), also called Epimanes (meaning "the mad," because of his alleged emotional instability), ruled the Seleucid Kingdom from 175 to 164 BC. Antiochus is probably best remembered for his encouragement of the expansion of Greek culture and Greek institutions. As a result of his vigour in pursuing this Hellenistic policy, he quite inadvertently promoted the creation of the independent Jewish state of the Maccabees, or Hasmoneans (165–63 BC), which has been a source of influence for modern Zionism and the state of Israel in the 20th century.

Born *c.* 215 BC, Antiochus was the third son of Antiochus III the Great. After his father's defeat by the Romans in 190–189, he served as hostage for his father in Rome from 189 to 175, where he learned to admire Roman institutions and policies. His brother, King Seleucus IV, exchanged him for Demetrius, the son of Seleucus; and after Seleucus was murdered by Heliodorus, a usurper, Antiochus in turn ousted him. During this period of uncertainty in Syria, the guardians of Ptolemy VI, the Egyptian ruler, laid claim to Coele Syria, Palestine, and Phoenicia, which Antiochus III had conquered. Both the Syrian and Egyptian parties appealed to Rome for help, but the Senate refused to take sides. In 173 Antiochus paid the remainder of the war indemnity that had been imposed by the Romans on Antiochus III at the Treaty of Apamea (188).

Invasion of Egypt

Antiochus forestalled an Egyptian expedition to Palestine by invading Egypt. He defeated the Egyptians between Pelusium and Mount Kasion, conquered Pelusium, and in 169 occupied Egypt with the exception of Alexandria, the capital. Ptolemy VI was Antiochus' nephew—Antiochus' sister, Cleopatra, had married Ptolemy V—and Antiochus contented himself with ruling Egypt as Ptolemy's guardian, giving Rome no excuse for intervention. The citizens of Alexandria, however, appealed to Ptolemy VIII, Ptolemy's brother, and to his sister Cleopatra to form a rival government. Disturbances in Palestine forced Antiochus to return to Syria, but he safeguarded his access to Egypt with a strong garrison in Pelusium.

In the winter of 169/168 Perseus of Macedonia in vain begged Antiochus to join forces with him against the danger that Rome presented to all of the Hellenistic monarchs. In Egypt, Ptolemy VI made common cause with his brother and sister and sent a renewed request to Rome for aid, and Antiochus prepared for battle. The fleet of Antiochus won a victory at Cyprus, whose governor surrendered the island to him. Antiochus himself invaded Egypt again in 168, demanded that Cyprus and Pelusium be ceded to him, occupied Lower Egypt, and camped outside Alexandria. The cause of the Ptolemaeans seemed lost. But on June 22, 168, the Romans defeated Perseus and his Macedonians at Pydna, and there deprived Antiochus of the benefits of his victory. In Eleusis, a suburb of Alexandria, the Roman ambassador, Gaius Popillius Laenas, presented Antiochus with the ultimatum that he evacuate Egypt and Cyprus immediately. Antiochus, taken by surprise, asked for time to consider. Popillius, however, drew a circle in the earth around the king with his walking stick and demanded an unequivocal answer before Antiochus left the circle. Dismayed by this public humiliation, the king quickly agreed to comply. Roman intervention had re-established the status quo. By being allowed to retain southern Syria, to which Egypt had laid claim, Antiochus was able to preserve the territorial integrity of his realm.

Both economically and socially he made efforts to strengthen his kingdom—inhabited in the main by Orientals (non-Greeks of Asia Minor and Persia)—by founding and fostering Greek cities. Even before he had begun his reign he had contributed to the building of the temple of Zeus in Athens and to the adornment of the theatre. He enlarged Antioch on the Orontes by adding

Antiochus' efforts to hellenize his kingdom

a section to the city (named Epiphania after him). There he built an aqueduct, a council hall, a marketplace, and a temple to Jupiter Capitolinus. Babylon, which revered him as Soter ("Liberator," or "Saviour") of Asia, was given a Greek colony that was granted freedom of the city. Another Epiphania was founded in Armenia. Ecbatana (in Persia) was also renamed Epiphania and became a Greek city. Many of these cities were granted the right to coin their own municipal currency. The mint of Antioch on the Persian Gulf served the trade along the sea route between India and the district at the mouth of the great Mesopotamian rivers.

Antiochus' hellenizing policies brought him into conflict with the prosperous Oriental temple organizations, and particularly with the Jews. Since Antiochus III's reign the Jews had enjoyed extensive autonomy under their high priest. They were divided into two parties, the orthodox Hasideans ("Pious Ones") and a reform party that favoured Hellenism. For financial reasons Antiochus supported the reform party and, in return for a considerable sum, permitted the high priest, Jason, to build a gymnasium in Jerusalem and to introduce the Greek mode of educating young people. In 172, for an even bigger tribute, he appointed Menelaus in place of Jason. In 169, however, while Antiochus was campaigning in Egypt, Jason conquered Jerusalem—with the exception of the citadel—and murdered many adherents of his rival Menelaus. When Antiochus returned from Egypt in 167 he took Jerusalem by storm and enforced its hellenization. The city forfeited its privileges and was permanently garrisoned by Syrian soldiers.

The Greeks and those friendly toward them were united into the community of Antiochians; the worship of Yahweh and all of the Jewish rites were forbidden on pain of death. In the Temple an altar to Zeus Olympios was erected, and sacrifices were to be made at the feet of an idol in the image of the king. Against that desecration Judas Maccabeus, leader of the anti-Greek Jews, led the aroused Hasideans in a guerrilla war and several times defeated the generals Antiochus had commissioned to deal with the uprising. Judas refused a partial amnesty, conquered Judaea with the exception of the Acra in Jerusalem, and in December 164 was able to tear down the altar of Zeus and reconsecrate the Temple. Antiochus apparently had underestimated the strength of the Hasidean movement, which was behind the success in maintaining an independent Judaean state for about a century. The fighting spirit of the Jews is the more impressive because at the beginning of their rebellion in 166 Antiochus had just demonstrated his might to the world at Daphne, near Antioch, with a grand review of his army: 46,000 foot soldiers were on parade, among them a Macedonian phalanx of 20,000 men and 500 mercenaries equipped with Roman arms, followed by 8,500 horsemen and 306 armoured elephants.

The revolt of Judas Maccabeus

Antiochus then mounted a campaign against the Parthians who were threatening the empire in the east, recovered the income from that area, forced Artaxias of Armenia—who had defected—to recognize his suzerainty, founded the city of Antioch on the Persian Gulf, set out on an expedition to the Arabian coast, and, at the end of 164, died of an illness at Tabae (Isfahan) in Persis (in modern Iran). Many believers saw his death as a punishment for his attempt to loot the shrine of Nanaia in Elam (in modern Iran).

BIBLIOGRAPHY. O. MORKOLM, *Antiochus IV of Syria* (1966), a comprehensive biography with extensive literature; E. BICKERMANN, *Der Gott der Makkabäer* (1937); and S.K. EDDY, *The King is Dead: Studies in the Near Eastern Resistance to Hellenism 334–31 B. C.* (1961), discuss the response of the Jews toward Antiochus' demand of hellenization.

(H.V.)

Antiseptic and Germicide

Antiseptics and germicides include a large variety of chemical compounds and, in some instances, physical agents that are used to destroy micro-organisms or to prevent their development. Antiseptics and germicides may be applied either directly or as gases or vapours on living and nonliving objects. These agents are used in hospitals, in eating and drinking establishments, and in homes. This article describes antiseptics and germicides—their uses, methods for their evaluation, their modes of action, and their classification.

General features. *Antiseptics and germicides.* Although antiseptic literally means against putrefaction or prevention of sepsis (spread of bacteria), the term usually is used to refer to agents applied to the living tissues of animals (including man) and plants in order to destroy or inhibit the growth of infectious micro-organisms. In low concentrations, an antiseptic may inhibit microbial growth; in high concentrations, it may kill them. Substances (*e.g.*, salves, ointments, and dressings) that are in contact with external tissues, such as skin, for long periods of time are considered antiseptics if they inhibit microbial growth. On the other hand, substances (*e.g.*, mouthwashes, douches, gargles) that have only brief contact with internal tissues such as mucous membranes are considered antiseptics only if recommended amounts destroy micro-organisms. The term asepsis is used to describe a process by which micro-organisms are prevented from entering some object; surgical asepsis, for example, refers to preventing the introduction of disease-causing bacteria into wounds of animals, including man.

Definition of antiseptic

Germicides (disinfectants) destroy harmful micro-organisms and also may inactivate viruses and protozoans but ordinarily do not affect bacterial or fungal spores, which are walled, usually dormant, reproductive bodies. Disinfection has been officially defined by the governments of several countries, including the U.S. and Great Britain, as the killing of disease-causing (pathogenic) agents by direct application of chemical or physical methods. The term disinfectant is usually used to refer to destruction of micro-organisms on inanimate surfaces; *e.g.*, surgical instruments, floors, walls, linens.

Many antiseptics and germicides are specific regarding destruction of certain micro-organisms and not others. An ideal compound that will destroy all micro-organisms, without causing any residual toxic effect in high concentrations, has not yet been developed.

Various terms are associated with antiseptics and germicides; in general, the suffix "-cide" indicates death, the suffix "-stat" indicates inactivation or inhibition of growth (*i.e.*, increase in numbers) of micro-organisms.

Other agents. Sterilization, any process, physical or chemical, that destroys all forms of life, is used especially to destroy micro-organisms, spores, and viruses. Precisely defined, sterilization is the complete destruction of all micro-organisms by a suitable chemical agent or by heat, either wet steam under pressure at 120° C (250° F) or more, for at least 15 minutes, or dry heat at 160° to 180° C (320° to 360° F) for three hours. Objects that are to be sterilized must withstand treatment without suffering deleterious effects.

A sanitizer is an agent, usually chemical in nature, that is used to reduce the number of micro-organisms to a level that has been officially approved as safe. Sanitizers are commonly used to control bacterial levels in equipment and utensils found in dairies, other food-processing plants, eating and drinking establishments, and other places in which no specific pathogenic micro-organisms are known to be present and destruction of all micro-organisms may not be necessary.

Preservatives, usually chemical agents, are added to certain foods and medicines to prevent the growth of micro-organisms that may cause spoilage or disease. Prophylactics also are agents used to prevent infections and diseases. Although antibiotics usually are used to treat infectious diseases already in progress in man and other animals, they also have been used as antiseptics, especially on the skin or on mucous surfaces, to prevent the entry of pathogenic micro-organisms. Vaccination is the administration of harmless amounts of disease-causing micro-organisms into animals, including man, to prevent diseases.

Antibiotics as antiseptics

Evaluation. Some countries have exacting requirements that must be met by the manufacturers of antiseptics and germicides. Authentic data on the efficacy of the

Table 1: Examples of Various Types of Chemical Compounds with Antimicrobial Properties

Acids (inorganic and organic)	**Alkalies**
hydrochloric acid	calcium hydroxide
nitric acid	sodium hydroxide
sulfuric acid	potassium hydroxide
phosphoric acid	trisodium phosphate
acetic acid	sodium bisulfate
carbolic acid	sodium nitrite
boric acid	sodium sulfonate
benzoic acid	sodium borate
salicylic acid	sodium carbonate
Alcohols	**Aldehydes**
methyl alcoho	acetic aldehyde
ethyl alcohol	formaldehyde
propyl alcohol	glyceric aldehyde
isopropyl alcohol	glycolic aldehyde
benzyl alcohol	methoxybenzyl aldehyde
amyl alcohol	isovaleric aldehyde
Aromatic oils	**Coal-tar derivatives**
camphor	phenol (carbolic acid)
peppermint	cresol
pine	phenol bismuth
cinnamon	phenol diiodide
eucalyptus	*para*-Nitrophenol
Dyes	**Halogens**
acridine	bromine
brilliant green	chlorine
gentian violet	fluorine
malachite green	iodine
Metals (as salts)	**Oxidizing agents**
aluminum	chlorine
cobalt	oxygen
copper	ozone
iron	perchloric acid
mercury	sodium bromate
silver	sodium permanganate
zinc	sodium peroxide
Reducing agents	**Surface-active agents**
carbon dioxide	anionics (true soaps, synthetic soaps, sulfonates)
carbon monoxide	cationics (quaternary ammonium germicides)
hydrogen	non-ionics (alkylated aryl polyether alcohol; polyethylene, derivative of sorbitan mono-oleate)
ethylene oxide	
sodium thiosulfate	
sodium thioglycollate	
calcium chloride (anhydrous)	
Sulfonamides	**Antibiotics**
sulfanilamide	penicillin
sulfathiazole	streptomycin
sulfapyridine	chloramphenicol
sulfaguanidine	polymyxin
sulfadiazine	neomycin
sulfamerazine	terramycin
sulfasuxidine	tyrothricin (Gramicidin)
Physical agents	
dry heat	sterile filtration
boiling	X-rays (radium or artificial gamma rays)
pasteurization	
fractional sterilization	radioactive isotopes (gamma rays)
steam under pressure (autoclaving)	high-speed electrons
incineration	ultraviolet light

products must be provided before they are released for sale to the public. Some test methods used in the evaluation of antiseptics and germicides are discussed in the following paragraphs. The categories of chemical compounds with antimicrobial activity and several examples of each are presented in Table 1.

Phenol coefficient tests. These tests compare the antimicrobial activity of antiseptics and germicides with that of the compound phenol, which is used as the standard or control. After micro-organisms have been added to various known quantities, both of the chemical to be tested and of phenol, and have remained there for certain periods of time, a sample is removed from each solution and added to a suitable liquid growth medium. The sample in which no visible microbial growth occurs is considered the end point; *i.e.*, the inhibitory or killing point of the agent on micro-organisms. The same end point also is recorded for phenol. The ratio between the sample of the test agent at which no growth occurs and the sample of phenol at which no growth occurs is called the phenol coefficient of the compound being tested. Dilute solutions of some chemical agents are highly inhibitory to the growth of micro-organisms or viruses. If small amounts of these compounds are carried into the growth medium along with the micro-organisms, they prevent the growth of, but fail to kill, them. For this reason, a compound that inactivates or neutralizes the agent must be added to test solutions before transfers of micro-organisms are made from them to a growth medium.

Agar plate tests. These tests are used to evaluate ointments, salves, powders, and oils with antiseptic properties. Micro-organisms are added to a semisolid (agar) growth medium; if the substance being tested is a liquid or a powder, it is placed in a cuplike depression in the growth medium. Medium containing micro-organisms may be poured over bandages and similar materials and allowed to solidify. When antibiotics and certain liquid disinfectants are used, a sterile glass or stainless steel cylinder is placed on the surface of the growth medium, and the test agents are placed in them. After a suitable period of time, usually one or two days, any clear zone (indicating no growth of micro-organisms) around the agent indicates its antiseptic properties; *i.e.*, the larger the zone of inhibition of microbial growth, the more potent is the antiseptic activity of the test agent.

Tests for sanitizers and bacteriostatic substances. If a sanitized object is rubbed with a cotton-tipped swab or some other applicator and then rinsed thoroughly into a liquid growth medium, any micro-organisms on the object are transferred to the medium and will grow in it. Any growth of micro-organisms is detectable after several days. The effectiveness of the sanitizer or bacteriostatic substance may be determined by comparing the results with those of similar nonsanitized objects.

Infection-preventing or protective tests. If at specific time intervals after micro-organisms, including viruses, have been exposed to an antimicrobial agent, they are injected into an animal, usually mice or guinea pigs, susceptible to the micro-organisms, the effectiveness of the agent can be determined. Animals that do not die or become ill after several days are killed and examined for evidence of disease. If no evidence of disease is found, it is assumed that the antimicrobial agent killed the micro-organisms in the original test suspension. Sometimes pathogenic micro-organisms are placed on the skin of animals, the agent being tested is added to the skin, and the treated skin is removed and placed inside the same animal. Survival of an animal indicates that the micro-organisms were destroyed by the agent. The test can also be carried out on ten-day-old chick embryos by removing part of the eggshell and placing micro-organisms on the membrane of an embryo to which an antiseptic also has been added. The membrane is observed through a shield covering the egg opening to determine the occurrence of any abnormalities of the membrane or death of the embryo; either development indicates that the micro-organisms were unaffected by the antiseptic or germicide.

Use of chick embryos

Toxicity tests. Tests similar to protective tests are used. In toxicity tests, however, the chemical agent is used alone; *i.e.*, no infectious micro-organisms are added. An agent may be added to the drinking water or to the food of test animals; abnormal changes in eating habits or physical appearance of the animals indicate a toxic reaction to the agent.

Types of antimicrobial agents. Table 1 is not complete; volumes have been written in some instances on a single category of the chemical and physical agents listed. Sterile filtration usually removes large micro-organisms (*e.g.*, bacteria, fungi, and their spores) from heat-sensitive solutions, but this physical method does not effectively remove small infectious micro-organisms (*e.g.*, filterable viruses and rickettsiae). Surface-active quaternary ammonium germicides, effective against most bacteria, are not effective against bacterial spores, the tubercle bacillus, and some viruses. Although both X-rays and certain radioactive isotopes emit similar rays (gamma rays), X-rays usually are used to combat fungal infections of the skin; radioactive isotopes seldom are used in this way because they are difficult to confine to one area. Sulfonamides and antibiotic drugs are chemotherapeutic agents; *i.e.*, drugs that combat infectious diseases already in progress. The specificity of chemicals, even in a single class of antimicrobial agents, is often difficult to determine; one may affect some organisms more than others.

Table 2: Classification and Survey of Antiseptics and Germicides

agent	approximate concentration used (percent)	mode of action	uses
Alcohols	50–70	denaturation of proteins; interference with metabolism; lysis (dissolving of organism)	as skin disinfectants; to form tinctures of antiseptics (used with acetone)
Cationic, surface-active quaternary ammonium compounds	0.1–0.25	denaturation of proteins; inactivation of cellular metabolites; dissolving of cell wall	as skin disinfectants and antiseptics; in mechanical removal of foreign matter from wounds; in sanitizing eating and drinking utensils, food-processing equipment
Bisphenols (2 phenols linked together)	2–5	inhibition of cell growth	as surgical scrubs (used with soaps and detergents); as deodorants
Chlorine gas and chloride compounds with available chlorine	0.0000002–0.000002	liberation of cell constituents	in chlorination of water supplies; as food-plant sanitizer; in treating wounds and hospital equipment
Iodine and iodized compounds	2–16	precipitation of cell proteins	in ointment and salves as skin antiseptics; in surgical-instrument disinfection
Aldehydes	1–5	general micro-organism poison	in disinfection of dwellings, ships, storage houses, utensils, clothing; as fungicides for plants, in hospital-instrument sterilization
Mercurials (inorganic and organic)	0.001–1	precipitation of cell proteins; destruction of enzymes	as skin antiseptic in skin ointments and salves; as preservatives for drugs
Oligodynamic metals (silver, copper, mercury)	traces	cell-membrane destruction; coagulation of cell materilas	as disinfectants; in ointments and salves; in cement (*e.g.*, in shower rooms)
Heavy metals	0.1–1	precipitation of cell proteins	in cosmetics and deodorants; antiperspirants; skin antiseptics
Acids	0.1–5	precipitation of cell proteins; destruction of cell wall	as skin antiseptics (salicylic benzoic acids); in food preservatives (benzoic acid); in de-destruction of bacteria and viruses (inorganic acids)
Dyes	0.1–1	inhibition of cell functions; combination with essential metabolites	in dentistry as mucous antiseptics; in laboratory media to inhibit growth of unwanted bacteria
Antibiotics and chemotherapeutic drugs	0.001–1	interference with cell metabolism; synergistic action in body to counteract infection	in chemotherapy of disease; in ointment and salves as skin antiseptics
Coal-tar derivatives	0.1–5	cytoplasmic poisons; disruption of cell wall; precipitation of proteins; inactivation of enzymes	as skin antiseptics in dilute solutions; as floor and wall disinfectants, combined with soaps; as surgical-instrument sterilizers
Salicylanilides and carbanilides (similar to bisphenols; combine with salicylic acid and aniline)	0.1–5	catalyze oxidation of cell; bind cell proteins; inhibit cell growth	as skin antiseptics in soaps and cosmetics; in surgical and bath soaps and laundry detergents
Aromatic oils (especially pine oil)	0.1–5	effect on cell constituents; mechanical effect inhibits cell growth	as disinfectants with soaps for washing floors and walls; to mask unpleasant odours; as a deodorant on inanimate surfaces

Modes of action. As indicated in Table 2, some antimicrobial compounds act directly on microbial cells to dissolve them. Surface-active agents may not dissolve the micro-organisms; instead, they may penetrate the cells and cause the release of amino acids, nuclear material, and other important chemical constituents. Some compounds penetrate microbial cell walls and inactivate essential membrane transport systems, so that the cells can no longer obtain the nutrients necessary for them to survive and to reproduce. Others coagulate certain vital materials in cells, thereby destroying the micro-organisms. A few agents disrupt the metabolism of the cells, so that they can no longer assimilate nutrients; as a result, the cells starve and die. The antimicrobial effects of some of these compounds can be seen with an ordinary light microscope. The effects of many chemical agents on disease-causing micro-organisms (*e.g.*, viruses, rickettsiae, mycoplasma), however, must be studied with more elaborate techniques (*e.g.*, the electron microscope or injection into a susceptible host).

Factors that affect antimicrobial activity. *Moisture.* Although water is essential for growth and reproduction of all forms of life, the spores produced by some micro-organisms may remain in a dormant state, in the absence of moisture, for long periods of time. After exposure to moisture and a favourable environment, the spores develop into active micro-organisms. Quick-freezing followed by rapid drying also is used to maintain viable organisms in a dormant state. Under conditions of favourable temperature, adequate moisture, and a suitable medium, such freeze-dried cells revive rapidly. Antimicrobial agents destroy micro-organisms most effectively, therefore, if water is present.

Oxygen. Organisms that require oxygen are called aerobes. Some micro-organisms (called anaerobes), however, cannot grow in the presence of oxygen; members of a

group of anaerobic bacteria (*Clostridium*) infectious in man and animals, for example, are able to remain dormant as spores in soil or in other materials. If introduced into an animal body with debris such as a nail or a bullet, the spores may develop into active micro-organisms, and the toxins formed by them produce infections; *e.g.*, tetanus (lockjaw), war-wound infections, gas gangrene. Certain chemical and physical agents destroy aerobes by depriving them of oxygen; others destroy viable anaerobes (*e.g.*, *Clostridium*) by providing them with oxygen.

Aerobes and anaerobes

Extraneous organic matter. Because most antimicrobial agents may be inactivated after contact with congealed blood and damaged tissues, these materials are removed from areas of minor burns or lacerations before the area is treated with an antiseptic. This procedure applies not only to organic matter but also to inorganic materials; *e.g.*, oil, grease, dirt, bits of clothing. If infectious organisms are expected in excreta (*e.g.*, sputa from tuberculosis patients; feces from patients with typhoid, dysentery, or cholera; pus from infected wounds), either high concentrations of a suitable germicide or incineration may be used.

Surface tension. Surface tension is an elastic-like force at the surface of a liquid that tends to constrict it. Sometimes it is advantageous to lower the surface tension of certain antimicrobial agents, especially those used to treat specific infected areas of the body or to destroy infectious micro-organisms on inanimate materials. One class of antimicrobial compounds, known as surface-active quaternary ammonium germicides, that may be used with other agents to lower their surface tension deeply penetrates the surface areas of the skin and destroy micro-organisms there. These compounds, when added to some drugs, assist the penetration of the latter into certain inaccessible areas, especially the eye, while simultaneously providing antimicrobial action. Other surface-tension depressants are used in surgical medicine to remove organisms from the skin of surgeon and patient before a more effective agent is used. Although synthetic soaps and detergents have a few antimicrobial properties, they are used as surface-tension depressants in the mechanical removal of organisms from many surfaces; *e.g.*, skin, floors and walls, eating and drinking utensils.

Concentration. An optimal concentration of an antiseptic effectively destroys micro-organisms and yet is nontoxic to cells and tissues; similarly, germicides should not damage metals, paints, or plastics on inanimate surfaces and should be nontoxic to the skin. Instructions on container labels, provided by the manufacturer of the compounds, indicate the proper concentrations for use.

Oligodynamic action. Oligodynamic action is the ability of extremely small amounts of certain metals, particularly silver, to kill or to inhibit the growth of micro-organisms; other metals that exert this action are copper and mercury. Oligodynamic metals often are added to antimicrobial solutions, ointments, salves, and disinfectants to provide additional antimicrobial action. Small quantities of silver are often added to cement, plastics, and paints used on floors and walls of shower rooms and in food-processing plants; solutions of silver nitrate are used as antiseptics to prevent the growth of micro-organisms on human skin.

Temperature. Although not necessarily important to antiseptic action on skin or mucous membranes, temperature may be an important factor in germicidal action. Few germicides destroy bacterial or fungal spores without the application of heat. Materials commonly treated in hot disinfectants include dental and surgical instruments. Boiling spore-contaminated instruments for several hours usually does not kill all of the spores, however.

Acidity or alkalinity. Acidity or alkalinity (usually expressed as the negative logarithm of the hydrogen-ion concentration, pH) may be an important factor in the antimicrobial effects of chemical disinfectants because their activity may vary with pH. In addition, acidity or alkalinity may be important in maintaining a disinfectant in the solution or emulsion form in which it is used. If a germicide is too alkaline or too acid, it may corrode metals with which it comes in contact. The pH of antiseptics should approximate that of the body; an extremely acid or alkaline antiseptic or disinfectant may produce toxic effects.

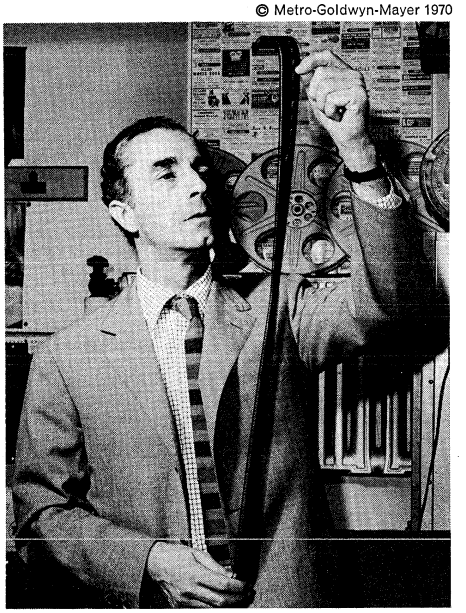
BIBLIOGRAPHY. C.A. LAWRENCE and S.S. BLOCK (eds.), *Disinfection, Sterilization, and Preservation*, pp. 109–233 (1968). Other information may be obtained from the United States Department of Agriculture, Disinfectants Evaluation Department, Pesticides Regulation Division, Washington, D.C.

(C.A.L.)

Antonioni, Michelangelo

One of the cinema's leading aesthetes, Michelangelo Antonioni spearheaded a critical and commercial reaction against the overwhelming acceptance, during post-World War II years, of unvarnished reality as a cinematic ideal. Under his direction the film became a metaphor for human experience, rather than a record of it. He never structured his films around a traditional plot or a character analysis; it was rather the visual image that became his fundamental vehicle of expression, most noticeably, the modern industrial landscape. After Antonioni, few people saw a water tower or a plastics dump with quite the same unappreciative eyes; in his films a new kind of beauty was revealed in the mechanical jungle of 20th-century urban society.

© Metro-Goldwyn-Mayer 1970



Antonioni.

Antonioni was born in Ferrara, Italy, on September 29, 1912. He is recognizably the product of the mild, uneventful plains of northern Italy that form the background for several of his films. Reserved and unexpansive in manner, he has said that the experience most important to his development as a filmmaker and as a man was his upbringing in a settled, bourgeois, provincial home, with a sufficiency of money; a traditional education; a code of reserve and self-discipline; and the leisure and ease necessary for a detached view of people and of life. He attended school in Ferrara and went to the university at Bologna, though he continued to live at home and commuted daily to his studies. As a child, his consuming interests were architecture and painting; at the university he studied classics, then economics and commerce. He also began haunting the cinemas and writing film criticism for a newspaper in the neighbouring city of Padova. In 1939 he decided to make the cinema his career.

Early enthusiasm for cinema

The obvious place to do this was Rome, where Antonioni soon became a staff member of the magazine *Cinema;* he also spent some months studying at a film school. What else he did in his 20s (apart from becoming an amateur tennis champion for northern Italy) remains obscure; his first credited film work dates from his 30th

year, when he collaborated on the scripts of some major feature films, one of them Roberto Rossellini's *Pilota ritorna*, and went to France to assist the director Marcel Carné on his wartime production *Les Visiteurs du soir*. In 1943 he began to direct his own first film, a short documentary called *Gente del Po*, but its completion was interrupted by the chaos of Italy's defeat in World War II. For a while Antonioni made his living by translating from the French, then became film critic of the underground paper *Italia libera*, and wrote some unproduced scripts. *Gente del Po* finally appeared in 1947; it was followed by six more shorts and then, in 1950, by his first feature, *Cronaca di un amore*, a rather bitter romance that immediately established him as a talent to be watched.

Since that time, Antonioni's life has, to the outside world, consisted of little more than the production of a steady succession of films. The landmarks are his first great film, *Le amiche* (based on a story by the Italian writer Cesare Pavese), in 1955; his first big international success, *L'avventura*, in 1959; his first film in colour, *Deserto rosso*, in 1964; his first full-length English-language film, *Blow-Up*, in 1966; and his first American film, *Zabriskie Point*, in 1970. He was responsible for shaping the career of the actress Monica Vitti, whose exquisite, mysterious presence provided that warming touch of human interest that assured *L'avventura*, despite its puzzling narrative structure and obscurity of motive, its breakthrough to a large international audience. Some mediation of this sort is necessary in Antonioni's films since without it his approach remains too cold and abstract to reach more than a select, esoteric audience. Human beings for him are depicted as moving parts in a pictorially exquisite pattern of moods and atmospheres; they wander around unable to engage in meaningful action, uncommunicative, lost in their dreams, while the outside world mirrors the feelings they never express, their alienation from one another and from the life around them.

Antonioni's critical reputation stood at its highest when the impact of his new, personal vision was most immediate, between *L'avventura* in 1959 and *L'eclisse* in 1962, in both of which a young woman, portrayed by Monica Vitti, unsuccessfully seeks in romance a means of coping with the emptiness of her life. *La notte*, the film he made in between, was similar in theme, but Jeanne Moreau, who played the central role, took less happily to Antonioni's methods of direction. In *Deserto rosso*, his attempts at overt symbolism, especially in his use of colour, were felt by some critics to be rather crude and intellectually pretentious. In his later films Antonioni still retained his reputation as a masterly technician in the face of some criticisms that he was attempting, but failing, to portray societies to which he was basically a stranger. The visual flair remains intact in *Blow-Up*, a film that surprised everyone by becoming a box-office success despite the frustration felt by many conventional moviegoers at its characteristic lack of structure and clear-cut action. *Zabriskie Point* re-emphasizes the gifts and their limitations; his portrayal of bourgeois American culture often appears trite, his affirmation of youth superficial. But however Antonioni's later films turn out, his place in the history of the cinema, and in the development of 20th-century sensibility, remains secure on the strength of *L'avventura* and *L'eclisse*.

Later films

MAJOR WORKS

SHORT FILMS: *Gente del Po* (1943–47); NU (1948); *Superstizione* (1948); *L'amorosa menzogna* (1948–49); *Sette canne un vestito* (1948); *La villa dei mostri* (1950); *La funivia del faloria* (1950).

FEATURE FILMS: *Cronaca di un amore* (1950); *I vinti* (1952); *La signora senza camelie* (1953); *Tentato suicidio* (1953; an episode in *L'amore in citta*); *Le amiche* (1955; *The Girl Friends*); *Il grido* (1957; *The Cry*); *L'avventura* (1959); *La notte* (1960; *The Night*); *L'eclisse* (1962; *Eclipse*); *Deserto rosso* (1964; *Red Desert*); *Prefazione* (1965; an episode in *I tre volti*); *Blow-Up* (1966); *Zabriskie Point* (1970).

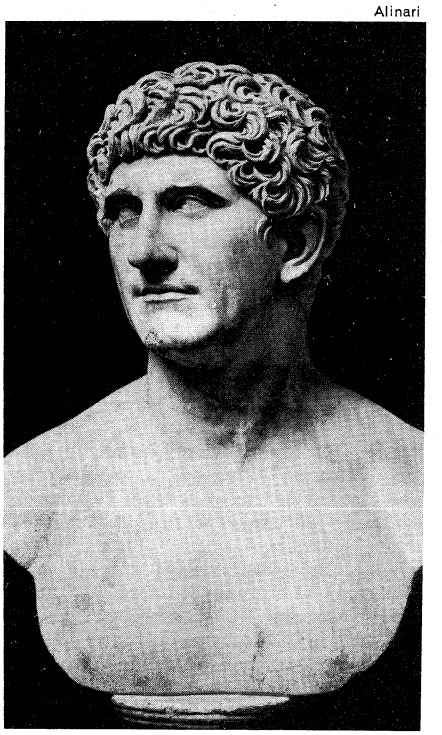
BIBLIOGRAPHY. PIERRE LEPROHON, *Michelangelo Antonioni* (1963), a critical study, with documents and extracts from Antonioni's writings and interviews; JOHN RUSSELL TAYLOR, *Cinema Eye, Cinema Ear* (1964), includes a critical study of Antonioni, with filmography and bibliography; IAN CAMERON and ROBIN WOOD, *Antonioni* (1969), a film-by-film analysis; MICHELANGELO ANTONIONI, *Screenplays: Il Grido, L'Avventura, La Notte, L'Eclisse* (Eng. trans. 1963), reading texts of four major works, with an introduction by Antonioni.

(J.R.T.)

Antonius, Marcus

An outstanding military and political leader and close associate of Julius Caesar, Marcus Antonius ensured the defeat of the Roman Republic by taking charge and manipulating the chaotic events that followed Caesar's assassination in 44 BC. He quelled the forces supporting the assassins Marcus Brutus and Gaius Cassius, who favoured a return to republicanism; personally distinguished himself at Philippi, where Brutus and Cassius were finally defeated; and joined the governing dictatorial trio, or triumvirate, in 43. In spite of his profound role in Roman politics, he is best remembered as the lover of Cleopatra and, later, as her ally against Rome.

Alinari



Marcus Antonius, marble bust. In the Vatican Museum.

Marcus Antonius was born in 82 or 81 BC, the son and grandson of men of the same name. His father was called Creticus because of his military operations in Crete, and his grandfather, one of the leading orators of his day, was vividly portrayed as a speaker in Cicero's *De oratore*. After a somewhat dissipated youth, the future triumvir served with distinction in 57–54 as a cavalry commander under Aulus Gabinius in Judaea and Egypt. He then joined the staff of Julius Caesar, to whom he was related on his mother's side, and served with him for much of the concluding phase of Caesar's conquest of central and northern Gaul and its aftermath (54–53, 52–50). In 51 Antonius held the minor office of quaestor, an office of financial administration that gave him a place in the senate, and he was subsequently elected to the politically influential priesthood of the augurs.

Participation in civil war and triumvirate. In 49, the year in which the Civil War broke out between Pompey and Caesar, Antonius became tribune of the people (an official with the function of protecting the plebeians from arbitrary actions of the patrician magistrates) and vigorously supported Caesar in the senate. He fled from Rome to his patron's headquarters after receiving threats of violence. After Antonius had fought in the brief Italian

campaign in which Pompey was forced to evacuate the Italian Peninsula, Caesar left him in charge of Italy, a post he again occupied in 48–47 as Master of the Horse (the dictator's assistant) after the decisive battle at Pharsalus (in Thessaly) in which he had commanded Caesar's left wing. Thereafter, because his methods as regent of Italy had displeased Caesar, he was removed from the post and was without employment until 44 when he became consul as the dictator's colleague. After Caesar's murder, he used a variety of methods, including the falsification of the dead man's papers, to control events and to arouse the people against Caesar's assassins, Marcus Brutus and Cassius. In June, Antonius was granted a five-year governorship of northern and central Transalpine Gaul (Gallia Comata) and Cisalpine Gaul (north Italy). Despite his growing power and popularity among the people, the orator Cicero attacked him fiercely in a series of speeches from September 44 to April 43 BC (he never tired of saying that Antonius should have been murdered also), and the 19-year-old Octavian, Caesar's great-nephew and adopted son, gradually emerged as a rival. In April 43 a coalition of Octavian, the two consuls of the year, and Decimus Brutus (another of the former conspirators against Caesar) defeated Antonius at Mutina (Modena) and compelled him to withdraw into the southern part (Narbonensis) of Transalpine Gaul. There, however, he was joined by a number of leading commanders including Marcus Aemilius Lepidus, who, after Antonius, had been Caesar's Master of the Horse. In early November Octavian met Antonius and Lepidus in Bononia (Bologna) and the three entered into an official five-year autocratic pact, the second triumvirate (November 43). The enemies of the triumvirs, including the orator Cicero, were proscribed and executed, and in the following year Marcus Brutus and Cassius killed themselves after their defeat at the battle of Philippi, in which Antonius greatly distinguished himself as a commander. The republican cause was now dead.

Administrator of eastern provinces

The triumvirs had agreed to divide the empire; so Antonius proceeded to take up the administration of the eastern provinces. He first summoned Cleopatra, the queen of Egypt, to Tarsus (southeast Asia Minor) to answer reports that she had assisted their enemies. She successfully exonerated herself, and Antonius spent the winter of 41–40 as her lover at Alexandria. In spite of the romantic accounts of ancient authors, however, she did not at this stage establish a permanent dominance over him, since he made no move to see her again for over three years.

Early in 40 he received two pieces of bad news: that his brother Lucius Antonius and his third wife, Fulvia, on their own initiative and without success, had revolted against Octavian, thus setting off the Perusine War (after the central point of the rising, Perusia, the modern Perugia); and that the Parthians, the eastern neighbours of the empire, had invaded Roman Syria. In spite of the latter information, Antonius first proceeded to Italy, where he became reconciled to Octavian at Brundusium (Brindisi), and, since Fulvia had died in the meantime, he married Octavian's sister Octavia. The two triumvirs agreed that Herod, who had fled from Judaea to escape the Parthians and their Jewish allies, should be encouraged to retake the country and become its king. In the following year they concluded the short-lived Treaty of Misenum with Pompey's son Sextus Pompeius, who because of his control of wide areas of the Mediterranean had been pirating Roman ships.

Accompanied by Octavia, Antonius then proceeded to Athens, where he was enthusiastically greeted and hailed as the New Dionysus, mystic god not only of wine but of happiness and immortality. In 38 Antonius' lieutenant Publius Ventidius won a decisive victory over the Parthians, and, in the following year, Herod was able to re-establish himself at Jerusalem. Meanwhile, however, further differences had arisen between Antonius and Octavian, and, although these were ostensibly settled by the Treaty of Tarentum, which prolonged the triumvirate for a further five years, Antonius sent Octavia back to Italy from Corcyra (Corfu; modern Kérkira) when he left again for the east and arranged for Cleopatra to join him in Syria. Henceforward, apart from his absences on land campaigns, they lived together for the remaining seven years of their lives.

Alliance with Cleopatra. Religious propaganda declared Cleopatra the New Isis, or Aphrodite, to his New Dionysus, and it is possible (but unlikely) that they contracted an Egyptian marriage: it would not have been valid in Roman law since Romans could not marry foreigners. Apart from their undoubted mutual affection, Cleopatra needed Antonius in order to revive the old boundaries of the Ptolemaic kingdom (though her efforts to convince him to give her Herod's Judaea failed), and Antonius needed Egypt as a source of supplies and funds for his planned attack on Parthia. His invasion, however, of Parthia's ally Media Atropatene (southwest of the Caspian) in 36 BC ended in a retreat involving heavy losses. On his return to Syria, Cleopatra met him with money and supplies. Octavian, exploiting the occasion and the contrast of Antonius' failure with the decisive victory he himself—or rather his admiral Agrippa—had won against Sextus Pompeius, sent Octavia to Antonius along with troops and provisions. But the soldiers fell far short of the numbers Antonius expected (and were owed by his fellow-triumvir), and he then made a future breach between the two leaders almost inevitable by ordering Octavia to return to Rome.

Break with Octavian

The break was accelerated in 34, when he celebrated a successful expedition to Armenia by appearing in a triumphal procession through the streets of Alexandria, a proceeding regarded by Romans as an impious parody of their traditional Triumph. A few days later he staged a ceremony at which Cleopatra was pronounced Queen of Kings, her son and joint monarch Ptolemy XV Caesar, or Caesarion (for Cleopatra, and now Antonius, claimed that Julius Caesar had fathered the boy), was declared King of Kings, and the two sons and a daughter that Cleopatra had borne to Antonius were also given imposing royal titles. The exact significance and substantiality of these Donations are disputable, but they could be interpreted as involving the transfer of Roman territories into alien hands. In the next year, 33, the two Roman leaders launched unprecedented, savage propaganda attacks upon one another, including the production by Octavian of a document (of dubious though possible authenticity) that purported to be a will of Antonius favouring the children of Cleopatra and providing for his own burial at Alexandria. In 32, the triumvirate had officially ended, although Antonius continued to call himself triumvir on his coins. Both consuls at Rome, however, happened to support Antonius, and now, threatened by Octavian, they left for his headquarters, bringing numerous, probably more than two hundred, Roman senators with them. After Antonius had officially divorced Octavia, her brother formally broke off the ties of personal friendship with him and declared war, not against him but against Cleopatra. Antonius successively established his headquarters at Ephesus (Selçuk), Athens, and Patras (Pátrai) and marshalled his principal fleet in the gulf of Ambracia (northwest Greece). More naval detachments occupied a long line of posts along the west coast of Greece. But Octavian's admiral Agrippa, and then Octavian himself, succeeded in sailing from Italy across the Ionian Sea and effecting landings, and Agrippa captured decisive points all along the line.

As Antonius lost more ground, the morale of his advisers and fighting forces deteriorated, a process aided by Cleopatra's insistence on being present at his headquarters against the wishes of many of his leading Roman supporters, thus providing Octavian with fresh propaganda fuel. Because of this lack of unity and the inexperience of Antonius' crews, the decisive battle was lost before it ever began. It took place off Actium, outside the Ambracian Gulf, on September 2, 31 BC. Antonius suffered the inevitable defeat, but Cleopatra, by prearranged plan rather than treachery, broke through the enemy line with her 60 ships (carrying her and Antonius' treasury) and, joined by her lover, made for Egypt. It was nearly a year before Octavian reached them there, but soon after his arrival, when resistance proved impossible, first An-

Defeat at Actium

tonius and then Cleopatra committed suicide (August 30 BC).

Personality. Antonius was a man of considerable ability and impressive appearance, far more genial than his adversary but not quite equal to Octavian's exceptional efficiency and energy and, in particular, unfit or unwilling to grasp the moment for action. Nevertheless, he was an outstanding leader of men and a competent general, though, in the end, not such a successful admiral as the experienced Agrippa. As a politician, he was astute enough—aided by a talent for florid oratory—but gradually lost touch with Roman feeling and fatally lacked the cold deliberateness of Octavian. Since the latter proved victorious in his struggle for power, it is his interpretation of events, rather than Antonius', that has remained lodged in the history books. Cicero had earlier depicted Antonius as a drunken, lustful debauchee—though his adulteries may have been less extensive than Octavian's. More significantly for history, the outcome of the battle off Actium made certain that Octavian's Roman-Italian policy prevailed throughout the empire, and the Antonian theme of Greco-Roman collaboration was not given a trial until the emperor Constantine captured Byzantium three centuries later.

BIBLIOGRAPHY. PLUTARCH, *Life of Antony;* APPIAN, *Civil Wars,* I–V; CASSIUS DIO XLI–LIII; CICERO, *Philippics, Letters to Atticus,* and *Letters to Friends;* JOSEPHUS, *Jewish War* and *Jewish Antiquities,* the basic ancient literary sources, all available in translation, but none of first-class reliability or sufficiently comprehensive; E.A. SYDENHAM, *The Coinage of the Roman Republic* (1952), a list, and in many cases photographs, of the principal Latin-language coins of Antonius; for commentaries on their designs, see H.A. GRUEBER, *Coins of the Roman Republic in the British Museum* (1910); for Greek-language coins, see M. GRANT, *From Imperium to Auctoritas* (1946, reprinted 1969); H.H. SCULLARD, *From the Gracchi to Nero,* 3rd ed. (1970), with a bibliography of recent additions to the extensive periodical literature; J.M. CARTER, *The Battle of Actium* (1970), an account of the events from 44 to 30 BC, with bibliography; W.W. TARN and M.P. CHARLESWORTH, *Octavian, Antony and Cleopatra,* reprint of *Cambridge Ancient History,* vol. 10, ch. 1–4 (1965), a sympathetic description, first published in 1934 and superseded in detail, of Antonius' regime in the eastern provinces; R.R. SYME, *The Roman Revolution* (1939), an account of the period based on analyses of party followings. The *Cambridge Ancient History,* vol. 9 (1932) and vol. 10 (1934), contain bibliographies up to those dates.

(M.Gr.)

Antwerp

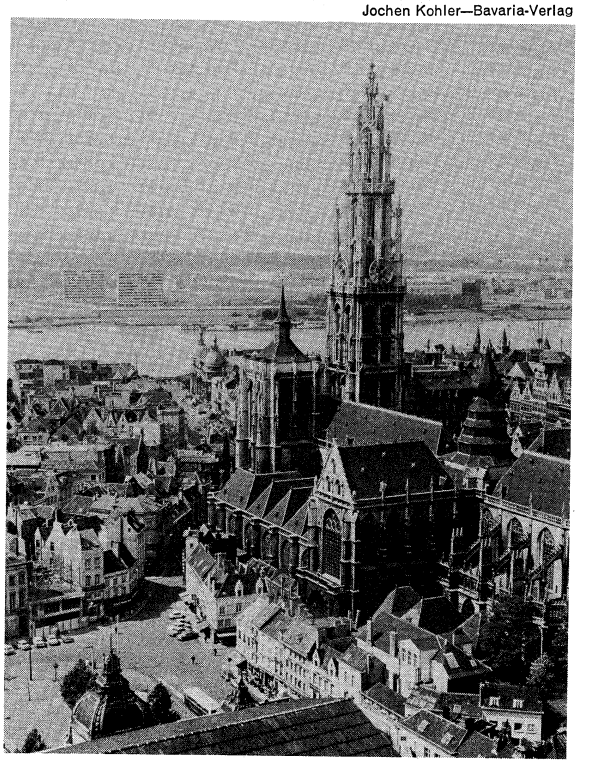
A city in the European Kingdom of Belgium, and the capital of a Belgian province of the same name, Antwerp (Antwerpen in Flemish, Anvers in French) is situated on the Schelde River. Greater Antwerp includes the city of Antwerp itself, with the surrounding municipalities of Ekeren, Merksem, Deurne, Borgerhout, Berchem, Wilrijk, Mortsel, and Hoboken. It had a total population of about 580,000 by the early 1970s.

Situated 55 miles (88 kilometres) from the North Sea, Antwerp is one of the biggest seaports in the world, ranking only after Rotterdam in The Netherlands and New York. As the Schelde, together with the Meuse and the Rhine, forms the biggest estuary in western Europe, Antwerp may be considered as an essential part of the greatest harbour complex not only in Europe but in the world, especially when seen in connection with growing European unity. The harbour installations of Antwerp have continued to grow, especially after World War II. Until the 1970s this expansion took place on the right bank of the Schelde only, but then development began on the left bank too, indicating that Antwerp's harbour facilities would ultimately extend on both banks of the river to the Dutch frontier, about 12.4 miles away. Because it lies in the Dutch- (Flemish-) speaking part of Belgium, the city often assumes, or tries to assume, the role of unofficial capital city of Flanders. Antwerpians generally take this role very seriously, conscious as they are of the great importance of their city in the past and present. The pride and competitive attitude thus exhibited by the inhabitants of the city has led to their being designated by the somewhat patrician nickname Sinjoren (from the Spanish *señores*).

Growth of the harbour

History. The site of Antwerp was probably already inhabited, as excavations on the right bank of the Schelde have proved, in Gallo-Roman times, in the 2nd and 3rd centuries. After the great Eurasian migrations of the 4th and 5th centuries, the region was occupied and Germanized by Franks and possibly Frisians, who gave it its present name, from the Germanic *anda* ("opposite") and *werpum* ("alluvium"). Christianity was introduced in the 7th century.

Jochen Kohler—Bavaria-Verlag



Cathedral of the Holy Virgin, Antwerp, on the east bank of the Schelde.

The city probably developed from two nuclei: a southern one called Chanelaus–Caloes–Callo and a more recent, northern nucleus grown around a fortification, which became the more important of the two. In the 9th century the region became a border county of the Holy Roman Empire. In 1124 the religious centre was transferred from Chanelaus to the northern nucleus. It was around the castle, built originally as a seat for the border county, that Antwerp developed, in the course of the centuries, in more or less concentric half circles.

Situated relatively far inland on the deep right bank of the Schelde, Antwerp was predestined to become a trade and shipping centre. This was already fully the case in the 13th century. At the end of that century and the beginning of the 14th, freedom of trade was given to the English, Venetians, and Geonese by the dukes of Brabant, who had made themselves masters of the county. The city of Antwerp became one of the duchy's capital cities, together with Louvain and Brussels, in Belgium, and the city of 's Hertogenbosh (Bois-le-Duc), in the Netherlands. In the first quarter of the 14th century, the Antwerp fairs began to flourish. These, together with the fairs of nearby Bergen op Zoom (now in The Netherlands) became one of the foundations of Antwerp's medieval economic growth. Antwerp succeeded, by gaining the Prince's favour, in becoming the successor of Bruges in Flanders, which until then had been the mercantile metropolis of western Europe. At the end of the 15th century, when nearly all the Low Countries were united under the Burgundian and Habsburg dynasties, the economic pre-eminence of Antwerp over Bruges is indicated by the fact that the majority of foreign merchants transferred their residence from the old Flemish town to

Commercial ascendancy

Antwerp. It now quickly became the leading commercial centre of western Europe, profiting from the beginnings of colonial trade, stimulated by the great discoveries of the Portuguese and Spaniards. Toward the middle of the 16th century, the population totalled nearly 100,000, whereas there had been about 20,000 people in the city at the end of the 14th century. Extensive urbanization plans were developed to lodge the increasing population, who earned their livelihood in trade, transport, and industry. The port underwent at this time its first northward extension. New industries established included breweries, malt factories, and bleaching works. Together with the already established finishing works of (English) cloth, tapestry and silk factories, sugar refineries, and the diamond industry, they made Antwerp one of the greatest industrial centres of western Europe. Antwerp also became a financial centre: its Stock Exchange (inaugurated 1531), a model for the younger London and Amsterdam exchanges, was the scene of dramatic and momentous events, in which financial agents and bankers of the Habsburg, Tudor, and Valois monarchs played their part, together with Antwerp, English, French, Portuguese, Italian, Spanish, and German merchants. Antwerp also became a great cultural centre: its school of painting began to flourish at the end of the 15th century; the city's printing houses became known throughout Europe; and humanism began to thrive in the city.

State bankruptcies in Spain, Portugal, and elsewhere, however, together with religious troubles and ensuing wars (Antwerp early in the 16th century became a centre of Protestant activity), brought about a decline. Antwerp became involved in the revolt of the Netherlands and was taken, in 1585, by Alessandro Farnese, duke of Parma, governor general for the Spanish king. The consequences of this strife with Spain were severe: from 1585 to 1589 the population diminished from 80,000 to 42,000. The Schelde, gateway to the sea, was closed by the Dutch, maintaining their positions against King Philip II of Spain. Capital and enterprise emigrated from Antwerp, mostly northward. The economic greatness of Amsterdam in the 17th century was due in part to Antwerp emigrants and their financial support, as was the case, to a lesser extent, for Hamburg, Frankfurt am Main, and other mercantile cities.

Yet Antwerp was not reduced to the status of a nonentity. The city remained the dynamic economic centre of the Spanish (later, in the 18th century, the Austrian) Netherlands. Antwerp became a more famous art centre than ever before: Peter Paul Rubens, Anthony van Dyck, Jacob Jordaens, and other major artists worked there. The French Revolution and the Napoleonic period, followed by Belgium's temporary union with The Netherlands (1815–30), reversed the general decline in trade of the 18th century. With the Schelde once more freely navigable (restored by the French in 1792), the attraction of Antwerp as a seaport became evident again. Since the second half of the 19th century, the growth of Antwerp's seaport has been interrupted only by the two world wars and associated German occupations. Immediately after liberation in 1944, Antwerp's nearly unharmed port played an immense part in the supply of the Allied armies aiming their final blow against Nazi Germany, despite devastating bombardment by German missiles. Since World War II, Antwerp, its suburbs, and seaport have grown apace, spurred by a commercial, industrial, and maritime boom. Port facilities, highways, and inland waterways have all been extended and improved. Some of this expansion has created ecological problems.

Revitalization of the port

The contemporary city. *Topography.* Antwerp's site on the right bank of the Schelde is a vast flat, alluvial plain. Since 1923, however, the city's territory also includes 4.7 square miles (12.2 square kilometres) on the left bank of the river (parts of Zwijndrecht and Burcht, municipalities that were formerly in the Belgian province of Oost-Vlaanderen [East Flanders]). Annexation of villages on the right bank northward of Antwerp (Oosterweel, Wilmarsdonk, and Oorderen in 1929 and Lillo, Berendrecht, and Zandvliet in 1958) extended the city's territory to the Dutch frontier.

The total area of contemporary Antwerp measures nearly 54 square miles (140 square kilometres) compared with seven square miles (19 square kilometres) before the beginning of the annexations. On this territory only a part, the original city and sections of the adjacent suburbs, is completely built upon. The extension of this agglomeration is continuing: outlying villages are quickly losing, or have already lost, their agricultural character and are growing in population as a result of emigration from the city. Population figures for the city proper illustrate this fact: whereas the number of inhabitants rose from 72,962 in 1829 to more than 320,000 in 1920, the population has declined steadily since then to about 230,000 in 1969. Only the left-bank territory annexed in 1923 is gaining in population.

Spread of the city

Until 1859–65 Antwerp was enclosed by fortifications erected in the 16th century. The territory outside the walls had an essentially agricultural character. With the economic boom of the second half of the 19th century, this farm and horticultural land was split up into building plots, while new fortifications were built in a larger half circle around the city and sections of the suburbs. In consequence three great parts can be discerned: the old city (within the 16th-century fortifications, which were transformed about 1860 into very broad boulevards), with mainly narrow, winding streets, and old buildings; the 19th-century city stretching beyond these boulevards and merging with some of the suburban extensions, with broader and substantially right-angled streets; and a third part, which largely came into existence after World War II, spreading beyond the 19th-century fortifications and flowing together with the surrounding municipalities, with many modern buildings.

The greatest part of Antwerp, however, is the essentially nonresidential northerly seaport complex, extending over the two great annexations of 1929 and 1958. The agricultural waterside villages then incorporated by the city have been eliminated, with the exception of Zandvliet and Berendrecht, to make room for expanding, if somewhat bleak, expanses of docks, industrial sites, and railway yards. Six locks connect this complex with the tidal Schelde River: Bonaparte (1811), Kattendijk (1860), Royers (1907), Van Cauwelaert (1928), Boudewijn (1955), and Zandvliet (1967).

The people. The inhabitants of Antwerp generally speak the local Brabantian-Antwerp dialect of Dutch. Dutch as spoken in The Netherlands and by Dutch-speaking cultured Belgians, however, is taught in the schools and is gaining ground in economic life. Although French is still much used in commercial and industrial circles, it is losing steadily as a means of expression. This does not mean that Dutch-speaking Antwerpians do not know French or other languages, for the international character of the city implies a readiness of many of its inhabitants to understand foreign languages. There were nearly 19,000 foreigners living in Antwerp by the early 1970s. The largest national group was the Dutch (about 4,500), followed by Moroccans (1,700), Spaniards (1,600), French (1,500), Germans (1,000); British and American colonies had, respectively, some 850 and 650 individuals, and the number of Israelis was 800. Unlike the other foreigners, the Moroccan and Spanish groups are largely unskilled workers who migrated to Belgium during periods of labour scarcity after World War II. There have been associated social tensions, notably concerning discrimination in housing.

Dialects and languages

The prevalent religion in Antwerp is Roman Catholicism. There are also small groups of various Protestant churches, and a sizable Jewish group of different tendencies. (By the 19th century, Antwerp already had a great number of Jewish residents, but many Jews died in World War II German concentration camps.) A large and growing part of the population is non-religious.

Economic life. Antwerp's economic life has for long been closely connected with its existence as a seaport. In the 16th century the city developed a great cane-sugar industry, in which partially worked raw material was brought in by ship, refined, and re-exported. The first petroleum refineries were established during the 1920s

and 1930s, and they were soon joined by automobile assembly plants. After World War II industrial expansion took on an explosive character. Larger petroleum refineries, closely followed by petrochemical industries, were established, together with chemical plants in the 1960s, and the automotive industry was restructured on a larger scale. In 1940 only about 250 acres of the port terrain were used for industry, but 6,200 acres were occupied by 1970. Besides these port-related activities, other industries include diamond cutting, already famous in the 16th century, and such newcomers as the photographic (1905) and electronics (1882) industries.

This complex of harbour and industrial activities is served by many commercial agencies, commission agents, import and export firms, banking establishments, insurance companies, road-transport enterprises, and railways (Antwerp has about 500 miles of rail).

Administration. Like all Belgian municipalities, Antwerp is governed by an elected city council (45 members), which in turn elects a board of ten aldermen, headed by a burgomaster, nominated by the king on recommendation of the council. Generally the burgomaster is a member of the elected council, but this is not obligatory. Each alderman has a specified department to manage, but decisions and resolutions are always taken by the board as a whole. Under the board of aldermen is the town clerk, the chief official to whom the different administrative services report. The city is also the residence of the government of the province of Antwerp, headed by a governor appointed by the king, a provincial council, elected every four years, and its board of deputies, who have, in regard to the province, the same task as the city aldermen. The provincial recorder is the counterpart of the town clerk. A third kind of public service is the Commission for Public Relief, consisting of a board of commissioners, presided over by a president, possessing their own administration under the supervision of their own chief official, the secretary. The members of the commission are appointed for six years by the city council.

Since World War I, Antwerp has been governed most often by a coalition of Social Democrats and members of the Catholic (Christian Democratic) Party. In the city council the opposition is formed by Volksunie (People's Union), a nationalist Flemish formation, Liberals, and Communists.

Education and health services. The Commission for Public Relief, responsible for public health and care of the aged and orphans, manages a series of institutions, among which are three great hospitals. Together with these official establishments, there are a number of independent, mostly Catholic institutions. This duality is also found in education. Besides the different schools of all grades (kindergarten, primary and secondary schools, technical institutions, etc.) managed by the city administration (and for a small part by the state), there exists an independent, essentially Catholic, network. This is also the case at the university level: from an official (state) and independent (Jesuit) institute, both founded in 1965, and both outgrowths of commercial high schools dating from 1852, a complete university is being developed. Higher art training is given in the Academy for Fine Arts, dating from 1663, whereas the Royal Conservatory provides advanced musical education.

Transportation. On a local (Greater Antwerp) level, transportation is in the hands of the MIVA company (Society for Inter-Municipal Transportation), which furnishes a network of bus lines and tramways, some of which are being transformed into subway lines. For the environs of Antwerp, the National Society for Local Railways provides a number of bus lines, which have replaced tramways. Antwerp is a railway centre for Belgian and international destinations, and its Deurne airfield (1928) has become important for freight and passenger flights, although pressure is being exerted by state authorities to divert Antwerp's air transportation to Brussels National Airport (at Zaventem near Brussels). Nearly 300 shipping lines assure more than 1,100 sailings a month to hundreds of ports of destination. An additional 55,000 to 60,000 dockings are made by inland barges that sail the rivers and canals not only to Belgian but also to Dutch, German, French, and Swiss destinations.

Cultural life. As the city where Rubens, van Dyck, and other famous artists were born, Antwerp seeks to maintain its reputation as a cultural centre. Among the city's museums are the Rubens' House, where the master lived and worked; the Museum Ridder Smidt van Gelder (porcelain, paintings, and furniture), the Museum Plantin-Moretus, house and workshop of the 16th-century printing master Christopher Plantin and his successors; Butchers' House (Vleeshuis) dating from 1501–03, now an archaeological museum with, among other features, a fine collection of musical instruments; the Steen, National Maritime Museum of Belgium; Brewers' House; the Museum Mayer van den Bergh, with a fine collection of paintings, sculpture, and furniture; a Folklore Museum; and an Ethnological Museum. There are two art museums: the Middelheim Open Air Museum for modern sculpture in a public park and the Royal (State) Museum of Fine Arts, where a great collection of paintings, for the most part the property of the city, is shown. A Royal Flemish Opera House and a Royal Dutch Theatre complete the list of official institutions. Besides these, there are a number of independently sponsored institutions, notable among which is the Antwerp Royal Society for Zoology (1843), owner of one of the oldest zoological gardens in the world, with its modern (1960) Queen Elisabeth Concert Hall.

Museums and galleries

BIBLIOGRAPHY

History: W.C. ROBINSON, *Antwerp: An Historical Sketch* (1904); LUDOVICO GUICCIARDINI, *Description de la cité d'Anvers* (1920); JERVIS WEGG, *Antwerp, 1477–1559, from the Battle of Nancy to the Treaty of Cateau Cambrésis* (1916), with good bibliography, and *The Decline of Antwerp Under Philip of Spain* (1924).

Economic aspects: HERMAN VAN DER WEE, *The Growth of the Antwerp Market and the European Economy (Fourteenth-Sixteenth Centuries)*, 3 vol. (1963); PORT OF ANTWERP, *Annual Report* (1927–).

Cultural aspects: L. DE BURBURE, *La Musique à Anvers aux XIVe, XVe, XVIe siècles* (1843); THEODORE L. DE VINNE, *The Plantin-Moretus Museum* (1929); JOHN J. MURRAY, *Antwerp in the Age of Plantin and Brueghel* (1970).

(J.V.R.)

Anura

Frogs and toads comprise the order Anura (Salientia of some authors) of the vertebrate class Amphibia and because of their wide distribution are known by most people around the world. The term frog is commonly applied to those forms with long legs and smooth, mucus-covered skins, toad being used for a variety of robust, short-legged anurans, especially those with rough skins. The name toad is applied so unevenly that one member of a family may be called a toad and a closely related member a frog. The familiar members of the family Bufonidae may be distinguished as "true toads." In this article, "frog" is applied generally to members of the Anura and "toad" to those for which it has traditionally been used. Many school children catch and keep frogs as pets and obtain one of their first biology lessons by rearing the larvae, known as tadpoles or polywogs, in science classes. College students become familiar with frog anatomy and embryology in biology courses. People in various parts of the world eat frog legs, and some kinds of toads are used in insect control on tropical plantations. Some South American Indians use the poisonous secretions of some kinds of frogs for poison arrows, and now biochemists are studying the possible medical uses of the constituents of the poison. The biologist interested in evolution finds a vast array of interesting, and often perplexing, problems in the study of frogs, a highly specialized group of amphibians.

GENERAL FEATURES

Although all frogs are readily recognizable, there are great varieties of sizes and of structural modifications. Many frogs are tiny animals; the smallest is the Cuban *Sminthillus limbatus*, adults of which do not exceed 12

millimetres (½ inch) in body length (with legs drawn in), whereas the African giant frog, *Conrana goliath*, has a body length of nearly 300 millimetres (slightly less than one foot). Many anurans have smooth, moist skins. Toads of the genus *Bufo* are familiar as "warty" amphibians, the skin being highly glandular and covered with tubercles. Frogs of many other families have rough, tubercular skins, usually an adaptation for life in the less humid environments. The opposite extreme is found in the small arboreal frogs of the tropical American family Centrolenidae, in which the skin on the under side is thin and transparent, and the heart and viscera can be seen through the skin.

Most frogs move by leaping. The long and powerful hind limbs are straightened rapidly from the crouching position, propelling the frog through the air into water or to a distant place on the ground. Many arboreal frogs, especially members of the family Hylidae, have adhesive discs on the ends of the fingers and toes and leap from branch to branch or from leaf to leaf. Members of the families Bufonidae, Rhinophrynidae, and Microhylidae and certain burrowing species in other families have relatively short hindlimbs and move forward by series of short hops. Some bufonids actually walk instead of hopping. Highly modified members of the hylid subfamily Phyllomedusinae have opposable digits on the hands and feet and walk slowly along branches, deliberately grasping the branch in the manner of tiny lemurs. Many kinds of frogs have membranous webbing between the fingers and toes; in the aquatic species the webbing on the feet aids in swimming. The extreme in this specialization is seen in the aquatic family Pipidae. Members of that family normally never leave the water. In regions of the earth subjected to long dry periods, frogs must seek cover in order to avoid desiccation and have behavioral and structural adaptations to conserve water.

Although many frogs are unimpressively coloured, some species are brilliantly marked. The most common colours are brown, gray, green, and yellow. Uniformly coloured frogs are the exception rather than the rule. The markings of a frog may seem bold when observed out of the natural habitat, but they usually are concealing or disruptive when the frog is in its environment.

Distribution and abundance

Because of their morphological and physiological adaptations, frogs are able to inhabit most regions of the world except the extremely cold land masses at high latitudes and some oceanic islands that they have been unable to colonize due to the barriers provided by saltwater. Frogs live in desert regions below sea level and in montane areas up to elevations above 4,560 metres (15,000 feet). Some members of the genus *Rana* live north of the Arctic Circle. Although widely distributed on the earth, frogs are most diverse and abundant in the tropics, and 7 of the 17 families are restricted to the tropics. In most temperate areas of the world, the number of species of frogs at any one locality is usually less than ten; but in the tropics, especially in rain forests, the number of species is much greater. At one locality in the upper Amazon Basin in eastern Ecuador, 83 species are known to occur, about the same number as is known for all of the United States.

In a complex environment, such as the tropical rain forest, the large number of species of frogs partition the environmental resources in a variety of ways. In the humid tropics, frogs can be active throughout the year, but many species are seasonal in their breeding activity. Various kinds of sites and different seasons are used for calling and egg laying; such temporal and spatial separation avoids interspecific competition. Frogs feed mostly on insects and other invertebrates, and the abundance of food in tropical rain forests probably places no competitive restrictions on this aspect of environmental resources. Some large species eat vertebrates, including small rodents and other frogs.

NATURAL HISTORY

Breeding behaviour. The breeding behaviour is one of the most distinctive attributes of the Anura. Because the eggs can develop only in moist conditions, most frogs place their eggs in bodies of fresh water. Many species congregate in large numbers at temporary pools for short breeding seasons. Mating calls serve to attract males and females to breeding sites. Other frogs breed along the mountain streams where they live year round. In the

Mating calls

H.V. Lacey—Annan Photo Features

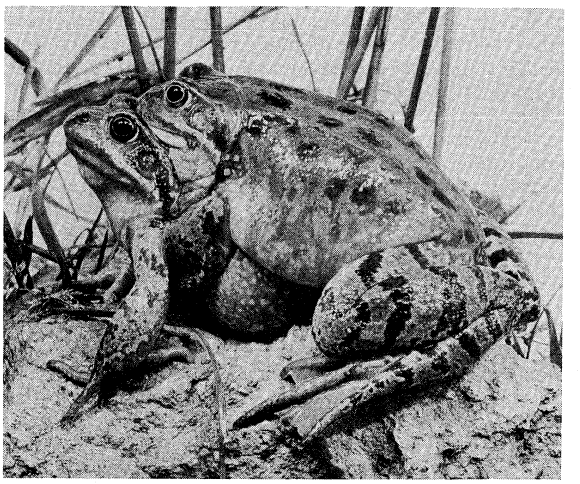


Figure 1: Frogs (*Rana*) in copulatory embrace.

latter species and in those that breed on the land, there is no great concentration of breeding individuals at one place. In all cases, the mating call produced by the male attracts the females. It has been observed in the field and experimentally in the laboratory that the females can discriminate between mating calls of their own species and those of other species. At a communal breeding site, such as a pond, swamp, or stream, differences in specific calling sites of the males help the frogs to maintain their identities. Differences in mating calls, however, constitute the principal premating isolating mechanism preventing hybridization of closely related species living in the same area and breeding at the same time and place. Frogs have rather simple vocal cords, in most species a pair of slits in the floor of the mouth opening into a vocal pouch. Air is forced from the lungs over the vocal cords, causing them to vibrate and thus produce sound of a given pitch and pulsation. The air passes into the vocal pouch, which, when inflated, acts as a resonating chamber emphasizing the same frequency or one of its harmonics. In this manner, different kinds of frogs produce different calls.

Females move toward and locate calling males. Once the male clasps the female, a union called amplexus, she selects the site for depositing the eggs. In the primitive frogs (superfamilies Discoglossoidea, Pipoidea, and Pelobatoidea), the male grasps the female from above and around the waist (inguinal amplexus); whereas in the advanced frogs (superfamilies Bufonoidea and Ranoidea), the position is shifted anteriorly to the armpits (axillary amplexus). The latter position brings the cloacae of the amplectic pair into closer proximity and presumably insures more efficient fertilization.

Most frogs deposit their eggs in quiet water as clumps, surface films, strings, or as individual eggs. The eggs may be freely suspended in the water or attached to sticks or submergent vegetation. Some frogs lay their eggs in streams, characteristically firmly attached to the lee sides or under sides of rocks where they are not subject to the current. The large pond-breeding frogs of the genus *Rana* and toads of the genus *Bufo* apparently produce more eggs than any other frogs. More than 10,000 eggs have been estimated in one clutch of the North American bullfrog, *Rana catesbeiana*. The habit of spreading the eggs as a film on the surface of the water apparently is an adaptation for oviposition in shallow temporary pools and allows the eggs to develop in the most highly oxygenated part of the pool. This type of egg deposition is characteristic of several groups of tree frogs, family Hylidae, in the

American tropics, one of which, *Smilisca baudinii*, is known to lay more than 3,000 eggs. Frogs breeding in cascading mountain streams lay far fewer eggs, usually no more than 200.

The problem of fertilization of eggs in rapidly flowing water has been overcome by various modifications. Some stream-breeding hylids have long cloacal tubes so that the semen can be directed onto the eggs as they emerge. Some other hylids have huge testes, which apparently produce vast quantities of sperm, helping to insure fertilization. Males of the North American tailed frog, *Ascaphus truei*, have an extension of the cloaca that functions as a copulatory organ (the "tail") to introduce sperm into the female's cloaca.

Aggressive behaviour

Most frogs are considered to be placid animals, but recent observations have shown that some species exhibit aggressive tendencies, especially at breeding time. Male bullfrogs (*Rana catesbeiana*) and green frogs (*R. clamitans*) defend calling territories against intrusion by other males by kicking, bumping, and biting. The South American nest-building hylid, *Hyla faber*, has a long sharp spine on the thumb with which males wound each other when wrestling. The small Central American *Dendrobates pumilio* calls from leaves of herbaceous plants. Intrusion into a territory of one calling male by another results in a wrestling match that terminates only after one male has been thrown off the leaf. Males of the Central American dendrobatid *Colostethus inguinalis* have calling sites on boulders in streams. The intrusion by another male results in the resident uttering a territorial call, and if the intruder does not leave, the resident charges him, attempting to butt him off the boulder. Females of the Venezuelan *Colostethus trinitatus* wrestle in defense of territories in a stream bed.

Males of at least three South American species of *Hyla* build basin-like nests, 25–30 centimetres (10–12 inches) wide and 2–5 centimetres (1–2 inches) deep, in the mud of riverbanks. Water seeps into the basin providing a medium for the eggs and young. Calling, mating, and oviposition take place in the nest, and the tadpoles undergo their development in the nests.

Some frogs in the families Leptodactylidae, Ranidae, and Microhylidae build froth nests. The small toadlike leptodactylids of the genus *Physalaemus* breed in small, shallow pools. Amplexus is axillary, and the pair floats on the water; as the female exudes the eggs, the male emits semen and kicks vigorously with his hind legs. The result is a frothy mixture of water, air, eggs, and semen, which floats on the water. This meringue-like nest is about 7.5–10.0 centimetres (3–4 inches) in diameter and about 5 centimetres (2 inches) deep. The outer surfaces exposed to the air harden and form a crust covering the moist interior, in which the eggs are randomly distributed. Upon hatching, the tadpoles wriggle down through the decaying froth into the water.

Metamorphosis

Most frogs have an aquatic, free-swimming larval stage (tadpole). After a period of growth, the tadpole undergoes a striking change (metamorphosis) in which the tail is lost and limbs appear. These are only two of the most obvious changes that take place. Tadpoles have a cartilaginous skeleton, thin nonglandular skin, and a long, coiled intestine; they lack jaws, lungs, and eyelids. Among the first changes that take place is the appearance of hindlimb buds, which grow and develop into differentiated hindlimbs, complete with toes, webbing, and tubercles. Much later, the front limbs emerge through the skin; in contrast to the hindlimbs, the forelimbs develop before they emerge. At the time of emergence of the forelimbs, the tail begins to shrink, being absorbed by the body. The tadpole mouth begins to change; as the horny denticles and papillae, if present, disappear, the jaws and true teeth develop. The eyelids develop, and mucous glands form in the skin. The vertebral column and limb bones ossify, and the adult digestive system differentiates as the long, coiled intestine shrinks to the short, thick-walled, folded intestine of the adult. These radical changes are equalled in the animal kingdom only by the metamorphosis found in insects.

Just how and where the changes from larva to adult take place is highly varied and is one of the most fascinating aspects of the study of frogs. The differences in modes of life history reflect varied environmental conditions. In various evolutionary lines in frogs, there is a strong tendency toward an escape from the ties that bind the less advanced frogs to the water for breeding.

The tadpoles of the pond breeders characteristically have rather large bodies and deep caudal (tail) fins, which in some have a terminal extension, as do the familiar swordtail fishes. The mouths are relatively small, either at the end of the snout or on the underside, and usually contain rather weak denticles (small toothlike projections). These tadpoles swim easily in the quiet water and feed on attached and free-floating vegetation, including algae. In contrast, the stream tadpoles have depressed bodies, long muscular tails, and shallow caudal fins. The mouth is relatively large and usually contains many rows of strong denticles. In highly modified stream tadpoles, the mouth is ventral and modified as an oral sucker, with which the tadpole anchors itself to stones in the stream. Such tadpoles move slowly across stones, gnawing off moss or algae as they move. It is conceivable that a tadpole can complete its larval life on a single large stone.

Drawing by Linda Trueb

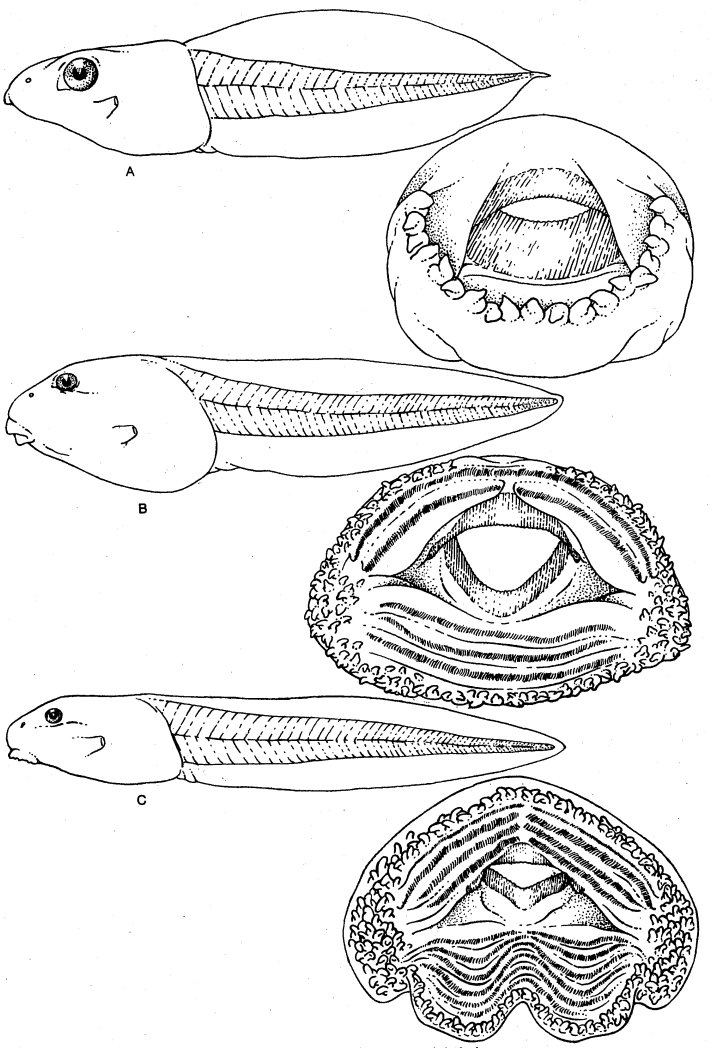


Figure 2: *Specializations of tail (left) and mouth (right) in tadpoles.*
(A) For pelagic life in ponds. (B) For bottom life in ponds. (C) For life in running water.

Most tadpoles complete their development in two or three months, but there are notable exceptions. Tadpoles of spadefoot toads, genus *Scaphiopus*, develop in temporary rain pools in arid parts of North America, where it is imperative for the tadpoles to complete their develop-

ment before the pools dry up. Some *Scaphiopus* tadpoles metamorphose about two weeks after hatching. In the northern part of its range in North America, the tadpoles of the bullfrog *Rana catesbeiana* require three years to undergo their development.

Some tree frogs of the family Hylidae deposit their eggs in water in trees. Several tropical species of *Hyla* lay their eggs in the water held in the overlapping bases of leaves of epiphytic bromeliads high in trees. Their tadpoles, which are slender with long muscular tails, develop in small quantities of water high above the ground. The Mexican hylid, *Anotheca spinosa*, lays its eggs in bromeliads or water-filled cavities in trees. The small tadpoles, like those of *Hyla*, feed on aquatic insect larvae, such as mosquitoes, but the larger tadpoles of *Anotheca* apparently feed only on the eggs of frogs.

A modification of the basic behaviour of aquatic eggs and aquatic tadpoles is the habit of depositing eggs on vegetation above water; this pattern occurs in some arboreal hylids and ranids and in all species in the family Centrolenidae. *Hyla ebraccata*, a small Central American tree frog, deposits its eggs in a single layer on the upper surfaces of horizontal leaves just a few inches above the pond. Upon hatching, the tadpoles wriggle to the edge of the leaf and drop into the water. The Mexican *Hyla thorectes* suspends 10–14 eggs on ferns overhanging cascading mountain streams. The phyllomedusine hylids in the American tropics suspend clutches of eggs from leaves or stems above ponds. Males call from trees; once a female has been attracted and amplexus takes place, the male placidly hangs onto the back of the female as she descends to the pond and absorbs water. This accomplished, she climbs into a tree, selects an oviposition site and deposits eggs until her water supply is depleted. She again descends to the pond and repeats the performance at a different site until the entire complement of eggs is deposited. Upon hatching, the tadpoles drop into the pond below. Most of the tree frogs of the family Centrolenidae are less than one inch long. Males call from leaves of trees or bushes over cascading mountain streams in the American tropics. Individuals return to the same leaf night after night. Attracted females are clasped on the leaf, and egg deposition takes place there immediately. A highly successful male may have three or four egg clutches on his leaf, each consisting of only about two dozen eggs. Upon hatching, the tadpoles drop into the stream bed; and if the tadpole lands on a stone, it flips about vigorously until it falls into the water, where it hides in the loose gravel on the bottom of the stream.

Parental care

Many kinds of frogs lay their eggs on land and subsequently transport the tadpoles to water. The ranid genus *Sooglossus* of the Seychelle Islands and all members of the family Dendrobatidae in the American tropics have terrestrial eggs. Upon hatching, the tadpoles adhere to the backs of adults, usually males. The exact means of attachment is not known. The frogs carry the tadpoles to streams, bromeliads, or pools of water in logs or stumps, where the tadpoles complete their development. The most unusual example of tadpole care known is that exhibited by the mouth-breeding frog, or Darwin's toad, *Rhinoderma darwinii* (family Rhinodermatidae) in southern South America. An amplectic pair deposits 20–30 eggs on moist ground. Several males gather around the eggs, and when the eggs are about ready to hatch, with the embryos moving, each male picks up some eggs with his tongue. The eggs pass through the vocal slits in the floor of the mouth and into the vocal sac. The eggs hatch, and the larvae complete their development in the large vocal sac. Upon metamorphosis, the young frogs emerge from the male's mouth.

The European midwife toad, *Alytes obstetricans* (Discoglossidae), displays a peculiar breeding habit. Inguinal amplexus takes place on land; at the time of oviposition the females extends her legs to form a receptacle for the string of 20–60 eggs. After fertilizing the eggs, the male moves forward on the back of the female and pushes his legs into the string of eggs until they are wound around his waist and legs. Then the female departs. The male carries the eggs with him on land until they are ready to hatch, at which time he moves to a pond, where the eggs hatch and complete their development.

The hylid *Gastrotheca marsupiata* lives in the high Andes of South America, one of several so-called marsupial frogs, so named because of a brood pouch in the back of the female. In amplexus the male exudes a quantity of semen, which flows into the pouch. The female extrudes eggs, a few at a time; these are pushed into the pouch by the male who uses the hindfeet to catch and push the eggs. The eggs are fertilized in the pouch, where they hatch and the tadpoles begin their development. Subsequently, the female moves to a pond, where the tadpoles emerge from the pouch and complete their development in the pond.

In each of the above instances of parental care, there is a trend away from the aquatic environment. Far fewer eggs (less than 50) are laid in comparison with those species depositing eggs in the water. The bonds with the aquatic environment have been partially broken, for, although the tadpoles must develop there, the eggs are effectively terrestrial, but not truly so, lacking the necessary embryonic membranes (allantois and amnion) to maintain physiological balance and lacking a shell. Consequently, if they are to survive and develop, the eggs must be maintained in moist places, such as damp soil or a part of the parental body. Water and waste products are transported through the membranes by osmosis.

The next evolutionary step in mode of life history is the elimination of the larval stage, thereby completely severing the ties with the aquatic environment. Direct development of the egg, in which the larvae undergo their development within the egg membranes and emerge as tiny froglets, occurs in many species, scattered in seven families (Leiopelmatidae, Pipidae, Leptodactylidae, Bufonidae, Brachycephalidae, Hylidae, and Microhylidae). Typical direct development of terrestrial eggs occurs in the many species of the leptodactylid genus *Eleutherodactylus* of Central and South America and the West Indies. During axillary amplexus, the female deposits a clutch of eggs in a moist place (beneath a log or stone, amidst leaf litter, in a rotting stump, in moss, or in a bromeliad). The parents depart, leaving the eggs to develop and subsequently hatch. In some *Eleutherodactylus* and in the New Zealand leiopelmatid *Leiopelma hochstetteri*, the hatching froglet still has a tail. In *Leiopelma*, at least, vigorous thrusts of the tail are used to rupture the egg membranes. Soon after hatching, the tail is completely absorbed.

Brooding of terrestrial eggs is known in a few species. Females of two species of *Eleutherodactylus* that lay their eggs on leaves of bushes or trees sit on the eggs. Apparently this brooding serves to prevent desiccation of the eggs by dry winds. Females of the Papuan microhylid *Sphenophryne* lay their few eggs beneath stones or logs and sit on top of them until they hatch.

Marsupial frogs

Direct development occurs in several genera of hylid marsupial frogs (*Gastrotheca* and *Amphignathodon*) living in mountain rain forests in northwestern South America. In these frogs, amplexus is axillary, and the female raises her cloaca so that the eggs, which are extruded one at a time, roll forward on her back and into the pouch. There the eggs develop into frogs. Large, external, gill-like structures envelop the developing embryo. These structures, which are attached to the throat of the embryo by a pair of cords, apparently function in respiration. These frogs live high in trees and complete their life cycle without descending to the ground. Thus they are rare in collections, and their biology is poorly known.

Some other South American genera of Hylidae also have direct development of eggs carried on the backs of the females. In *Flectonotus* and *Fritziana*, the eggs are in one large basin-like depression in the back, whereas in others each egg is in an individual depression. In *Hemiphractus*, gills and cords like those in *Gastrotheca* are present. At hatching, the expanded gill adheres to the modified skin of the maternal depression and is attached to the young by the pair of cords. The female carries the

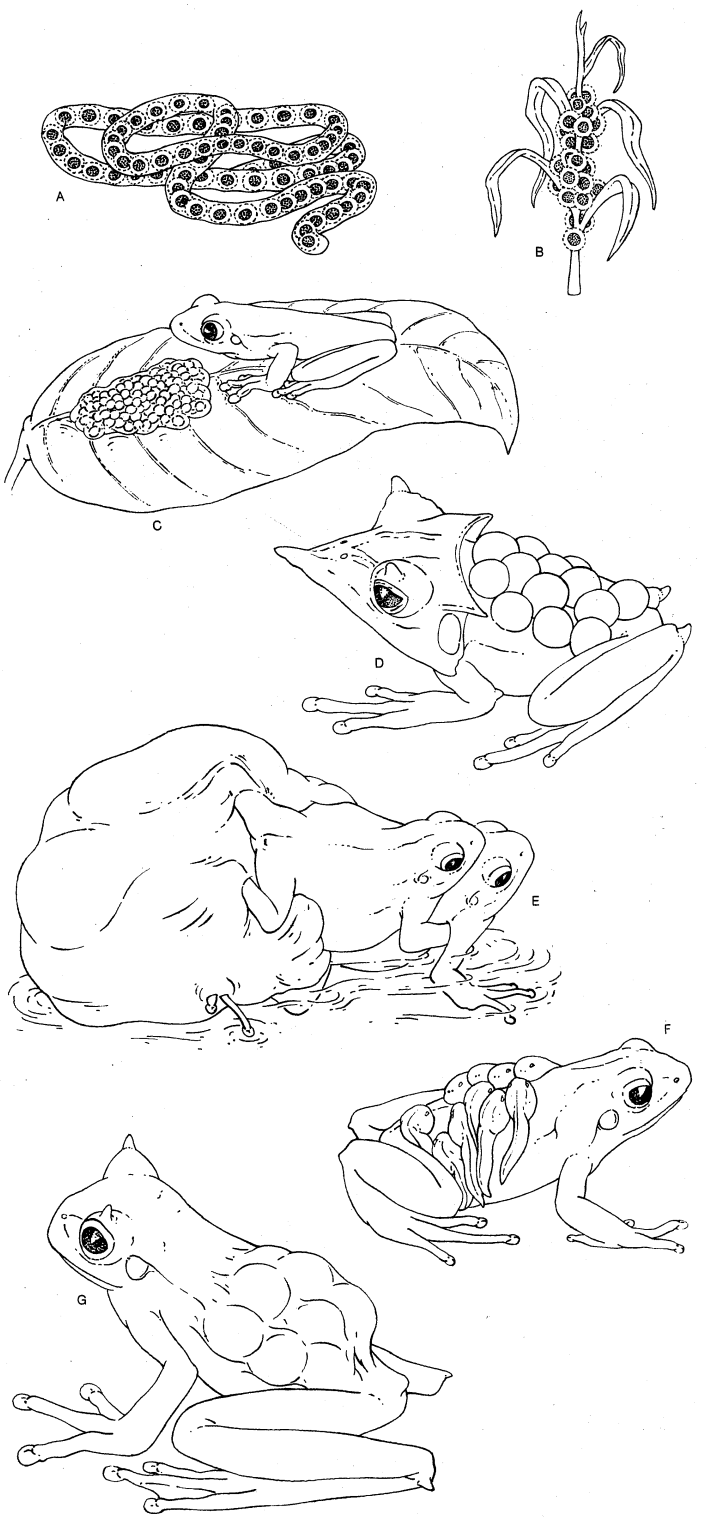


Figure 3: *Specialization for breeding in anurans.* (A) Aquatic eggs free-floating in a string. (B) Aquatic eggs attached to submerged vegetation. (C) Eggs of *Phyllomedusa* on leaf above water. (D) *Hemiphractus* with eggs on its back. (E) Froth nest made by amplexing pair of *Physalaemus*. (F) Tadpoles on the back of a male *Colostethus*. (G) Marsupial frog (*Gastrotheca*) with eggs in pouch.
Drawing by Linda Trueb

young until they are sufficiently well developed to care for themselves. The manner of detachment of gill from female and young is unknown.

The strictly aquatic Surinam toad *Pipa pipa* of northern South America also has direct development, in this case in the water. Eggs are carried in individual depressions in the back of the female. Amplexus is inguinal, and the pair rests on the bottom of the pond. The female initiates vertical circular turnovers, at the height of which she extrudes a few eggs. These are fertilized, fall against the belly of the then upside down male, and are pushed forward onto the female's back, where they adhere and subsequently become enclosed in tissue. Upon completion of development, the young frogs emerge from the skin of the female's back.

The most advanced form of reproduction known in frogs takes place in the small African bufonids of the genus *Nectophrynoides*. By some unknown means, fertilization is internal, and the young are born alive. It is noteworthy that the evolution of live birth has taken place independently in all three living orders of amphibians, for this phenomenon also occurs in salamanders and caecilians.

Feeding habits. The great majority of frogs feed on insects, other small arthropods, or worms, but some larger species eat vertebrates. The South American leptodactylid *Ceratophrys varia* and the large bufonid *Bufo marinus* eat other frogs and small rodents. The superficially similar Solomon Island ranid, *Ceratobatrachus guentheri*, and the South American hylids, *Hemiphractus*, eat other frogs. Large North American bullfrogs, *Rana catesbeiana*, have been reported to consume other frogs, mice, small snakes, and even small turtles.

FORM AND FUNCTION

Adult anurans are easily recognized, by layman and specialist alike, by the short body and elongated hindlimbs, the absence of a visible neck, and the absence of a tail. The compact body has been attained by a reduction of the number of trunk vertebrae and the fusion of tail vertebrae into a single rodlike bone, the coccyx or urastyle. The lengthening of the hindlimbs has been attained, in part, by the elongation of two bones (astragalus and calcaneum) in the foot. Considering the variety of habitats occupied by anurans, there is remarkably little gross variation in body plan. The female is usually larger than the male. In most frogs the tympanic membrane (eardrum) is visible as a prominent disc on each side of the head. Correlated with a sound-oriented existence, the larynx is also well developed, often accompanied by single or paired inflatable resonating sacs.

Poisons

All frogs have poison glands in the skin, well developed in many diverse groups. In the Dendrobatidae, the skin secretions are especially toxic. *Dendrobates* and *Phyllobates* are small, diurnal frogs living in Central and South America, brilliantly coloured, solid red, yellow, or orange, or patterned with bold stripes or cross bars. These bright patterns are believed to act as warning colours to ward off predators. One nonpoisonous South American leptodactylid, *Lithodytes lineatus*, mimics the dendrobatid *Phyllobates femoralis*, thus gaining protection from predation (see also MIMICRY).

The biochemical properties of amphibian skin toxins are highly varied, most being complex nitrogenous compounds. The toxically active ingredients are of various types, from local irritants to convulsants, hallucinogens, neurotoxins (nerve poisons), and vasoconstrictors (acting to narrow blood vessels). The medical importance of these ingredients is now being investigated. Although these skin secretions irritate human skin and mucous membranes, they do not cause warts.

Coloration

The skin toxins of most frogs do not provide security from predators; in fact, frogs are a basic food for many snakes, birds, and mammals. Edible anurans rely on modifications of shape, skin texture, and colour, supplemented by behaviour, to escape detection. These modifications may reach remarkable extremes. Hylids of the South American genus *Hemiphractus* live on the forest floor among leaf litter and have flattened bodies that enable them to blend well with dead leaves. Several tree frogs, rough skinned and greenish-gray, resemble lichens when flattened out on tree trunks. The coloration of many frogs changes from night to day. In most species the colour is darker and the pattern more distinct by day than by night, but the reverse is true for some tree frogs in-

habiting semi-arid regions. Colour change is brought about through the stimuli of light and moisture, which create a physiological change and result in contraction or expansion of the melanophores (pigment cells) in the skin.

More difficult to comprehend is the striking array of colours on the hidden surfaces of frogs. Many frogs that are rather dull or uniformly coloured when in a resting position have bright colours or patterns on the flanks, groin, posterior surfaces of the thighs, and belly. For example, the South and Central American hylid, *Agalychnis calcarifer*, when observed sleeping by day, is nothing more than a green bump on a leaf. The eyes are closed, the hindlimbs drawn in close to the body, and the hands folded beneath the chin. Upon moving, the frog creates a striking appearance, previously hidden surfaces showing a deep golden orange, interrupted by vertical black bars on the flanks and thighs. These so-called flash colours are common in frogs and are thought to serve in species recognition or in confusing predators. Some colour patterns obviously do confuse predators. The South American leptodactylids of the genus *Eupsophus* have a pair of brightly coloured "eyespots" on the rump. When approached by a potential predator, the frog lowers its head and elevates the rump, thus confronting the predator with a seemingly much larger head.

Structural modifications allow certain specialized frogs to survive dry periods. Some arboreal frogs hide in bromeliad plants (various genera of the family Bromeliaceae), which hold water in the axils of their leaves. Among the Hylidae are genera that have the head modified into a bony casque ("helmet") and the skin co-ossified with the underlying bone. The head is used by some species to plug the constricted base of the bromeliads and by others to plug up holes in trees, the frogs surviving the dry season by utilizing what little moisture is trapped in the cavity.

Most toads of the genus *Bufo* and many genera in the families Rhinophrynidae, Pelobatidae, Myobatrachidae, Leptodactylidae, Hylidae, Ranidae, and Microhylidae burrow in sand, soil, or mud. Many of these species have the tubercles on the middle (metatarsal) part of each foot modified into a spade-shaped digging organ. The animals are highly resistant to desiccation and conserve water in the body by the mucous skin secretions that tend to make the skin impermeable. This modification is carried to the extreme in some Australian myobatrachids, which secrete a cocoon formed of hardened mucus.

CLASSIFICATION

Features used in classification. The superfamilies and families of the Anura are based on anatomical, developmental, and behavioral characteristics. Important anatomical features include: the type of vertebrate present, especially with respect to the articulating surfaces, which may be concave on the anterior end of each vertebra (procoelus), on the posterior side (opisthocoelus), or concave on both ends (amphicoelus); the presence or absence of intercalary cartilages (between the terminal and penultimate bones of the digits); whether the pectoral girdle is firmisternal (*i.e.*, with the cartilages of the two epicoracoid bones fused together) or arciferal (with these cartilages separate and overlapping); the presence or absence of an anterior projection, the omosternum, on the pectoral girdle; the presence or absence of teeth on the maxillary bone; and the presence or absence of Bidder's organ (a rudimentary ovary) in the male. The type of egg and anatomy of the tadpole are often important. The key behavioral characteristics are primarily those involved with reproduction, such as the type of amplexus and method of egg deposition.

Annotated classification. The classification presented here is an amalgamation of recent views by various specialists, none of whom has surveyed the entire order. A scanty record of meaningful fossils and inadequate knowledge of the morphology and mode of life history of many kinds of frogs result in inconclusive evidence for any classification of the families; consequently, the following classification must be considered to be tentative.

ORDER ANURA

Amphibians lacking a tail in the adult stage; 6 to 9 presacral vertebrae; postsacral vertebrae (posterior to the pelvis) fused into a bony coccyx; hindlimbs elongated, modified for jumping; fertilization normally external; eggs laid in water or not; an aquatic larval stage (tadpole) present in most; males usually with vocal cords, vocal sac (resonating chamber), and a voice. About 2,660 species.

Superfamily Discoglossoidea

Vertebrae amphicoelous with labile centra (*i.e.*, joined by rings of cartilage); pectoral girdle arciferal; ribs present on anterior vertebrae; amplexus inguinal; larvae with single median spiracle (respiratory opening) and complex mouthparts or with direct development (froglet hatching directly from the egg).

Family Leiopelmatidae

No fossil record. Nine presacral vertebrae (*i.e.*, anterior to the pelvic girdle); parahyoid and caudalipuboischiotibialis ("tail-wagging") muscles present; direct development (*Liopelma*) or stream-adapted tadpoles (*Ascaphus*). New Zealand and northwestern North America; 2 genera, 4 species. Adult length about 5 cm.

Family Discoglossidae (fire-bellied and midwife toads)

Eocene to present. Usually 8 presacral vertebrae; parahyoid tongue muscle and caudalipuboischiotibialis muscle absent. Still-water tadpoles. Eurasia, North Africa, and Philippine Islands; 4 genera, 8 species. Adult length to about 10 cm.

Superfamily Pipoidea

Vertebrae opisthocoelous; pectoral girdle arciferal; ribs absent or fused to transverse processes of vertebrae; amplexus inguinal; larvae with paired spiracles and simple mouthparts or with direct development.

Family Rhinophrynidae (burrowing toad)

Oligocene to present. Eight presacral vertebrae; ribs absent; coccyx free, with 2 articulating surfaces; tongue free and protrusible; body robust; burrowing; aquatic larvae present. Mexico and Central America, 1 species; adult length to about 7 cm.

Family Pipidae (tongueless frogs)

Cretaceous to present; 5 to 8 presacral vertebrae; ribs present and free in larvae, but fused to transverse processes of vertebrae in adults; coccyx fused to sacrum or free and monocondylar (*i.e.*, with 1 articulation); tongue absent; body flattened. Aquatic, direct development or aquatic larvae present. Africa south of Sahara and tropical South America, east of Andes; 6 genera, 14 species; adult length 5–20 cm.

Superfamily Pelobatoidea

Vertebrae procoelous with labile centra; pectoral girdle arciferal; ribs absent; amplexus inguinal; larvae with single spiracle on the left, and complex mouthparts.

Family Pelobatidea (spadefoots)

Eocene to present. Eight presacral vertebrae; coccyx fused to sacrum or free and monocondylar. Ten genera, 59 species; adult length 4 to about 15 cm. Three subfamilies: Megophryinae (Southeast Asia, Indo-Australian Archipelago, Philippine Islands), Pelodytinae (southwestern Europe and southwestern Asia), Pelobatinae (Europe and North America).

Family Myobatrachidae

Eocene to present. Eight presacral vertebrae; coccyx free, bicondylar. Eighteen genera, 76 species; adult length to about 10 cm. Three subfamilies: Myobatrachinae (New Guinea and Australia), Cycloraninae (New Guinea and Australia), Heleophryninae (South Africa).

Superfamily Bufonoidea

Vertebrae procoelous; pectoral girdle arciferal (in some, secondarily firmisternal); ribs absent; amplexus axillary; larvae usually with single spiracle, on the left, and complex mouthparts, or with direct development.

Family Rhinodermatidae (mouth-breeding frog)

No fossil record. Eight presacral vertebrae, 1st and 2nd fused; pectoral girdle partly firmisternal; maxillary teeth, intercalary cartilages, and Bidder's organ absent; omosternum cartilaginous. Southern South America; 1 species; adult length 2.5 cm.

Family Leptodactylidae

Eocene to present. Eight presacral vertebrae; pectoral girdle arciferal; maxillary teeth present; Bidder's organ and intercalary cartilages absent; omosternum cartilaginous or ossified. Thirty-nine genera, about 625 species; adult length 2 to about 20 cm. Four subfamilies: Ceratophryinae (South America), Telmatobiinae (South and Central America, West Indies), Elosiinae (South America), Leptodactylinae (South and Central America).

Family Bufonidae (true toads)

Oligocene to present. Five to 8 presacral vertebrae; pectoral girdle arciferal, or partly or even completely firmisternal; intercalary cartilages and omosternum absent; Bidder's organ present; maxillary teeth present or absent; aquatic larvae, direct development, or live birth (*Nectophrynoides* only). Worldwide, except eastern part of Indo-Australian Archipelago, Polynesia, Madagascar; *Bufo marinus* introduced into Australia and some Pacific islands; 13 genera, about 234 species; adult size 2 to about 25 cm.

Family Brachycephalidae

No fossil record. Seven presacral vertebrae, pectoral girdle partly firmisternal; intercalary cartilages and omosternum absent; Bidder's organ present; maxillary teeth present; direct development. Southeastern Brazil; 1 species; adult length 2 cm.

Family Dendrobatidae (arrow-poison frogs)

No fossil record. Eight presacral vertebrae; pectoral girdle completely firmisternal; intercalary cartilages absent; omosternum present; Bidder's organ absent; maxillary teeth present or absent. Larvae carried on backs of adults. Central and South America; 3 genera, about 70 species; adult length 1.5–5 cm.

Family Pseudidae

No fossil record; 8 presacral vertebrae; sacral diapophyses round; pectoral girdle arciferal; intercalary cartilages present, ossified; omosternum present; Bidder's organ absent; maxillary teeth present; aquatic larvae (which grow to a much larger size than the adult). South America east of Andes; 2 genera, 5 species; adult length 2–5 cm, larval length to 25 cm.

Family Centrolenidae

No fossil record; 8 presacral vertebrae; pectoral girdle arciferal; intercalary cartilages present; omosternum absent; Bidder's organ absent; maxillary teeth present; terminal phalanges T-shaped; astragalus and calcaneum bones of the foot fused. Stream-adapted larvae. Central and South America; 2 genera, about 45 species; adult length 2–6 cm.

Family Hylidae (tree frogs)

Miocene to present; 8 presacral vertebrae; pectoral girdle arciferal; intercalary cartilages present; omosternum absent; Bidder's organ absent; maxillary teeth usually present; terminal phalanges claw-shaped; astragalus and calcaneum not fused; aquatic larvae or direct development; adult length 1.3 to about 12 cm. Four subfamilies: Phyllomedusinae (Central and South America), Hemiphractinae (Central and South America), Amphignathodontinae (Central and South America), Hylinae (North and South America, Europe, Asia except Indian subregion, Africa north of Sahara and Australasia).

Superfamily Ranoidea

Pectoral girdle firmisternal; ribs absent; amplexus axillary; larvae with single sinistral spiracle and complex mouthparts, single median spiracle and simple mouthparts, lacking spiracle or undergoing direct development.

Family Ranidae (true frogs)

Miocene to present. Vertebral column diplasiocoelous (mixed amphicoelous and procoelous); intercalary cartilages present or absent; larvae with single spiracle, on left, and complex mouthparts. About 55 genera and 842 species, adult length about 2–25 cm. Nine subfamilies: Raninae (worldwide except for southern South America, southern and central Australia, New Zealand, and eastern Polynesia), Astylosterninae (Africa), Scaphiophryninae (Madagascar), Rhacophorinae (Africa, Madagascar, southeastern Asia, Indo-Australian Archipelago), Arthroleptinae (Africa), Cacosterninae (Africa), Hyperoliinae (Africa, Madagascar, Seychelles), Hemisinae (Africa), Petropedatinae (Africa).

Family Sooglossidae

No fossil record. Vertebrae procoelous; sacral diapophyses dilated; intercalary cartilages absent; larvae lacking spiracle. Seychelle Islands; 2 genera, 3 species; length about 4 cm.

Family Microhylidae

No fossil record. Vertebrae procoelous or diplasiocoelous; intercalary cartilages usually absent; larvae with single median spiracle and simple mouthparts, or undergoing direct development. Eight subfamilies: Cophylinae (Madagascar), Dyscophinae (Malay Peninsula, western Indo-Australian Archipelago, Madagascar), Asterophryninae (New Guinea and Sulu Islands), Sphenophryninae (Philippine Islands, eastern Indo-Australian Archipelago, New Guinea, northern Australia), Brevicipitinae (Africa), Microhylinae (North and South America, Southeast Asia, Sri Lanka, western Indo-Australian Archipelago, Philippine and Riu Kiu Islands), Melanobatrachinae (east-central Africa, India), Phrynomerinae (Africa).

Critical appraisal. Modern authorities do not agree on all aspects of anuran classification, and further study is needed to clarify the relationships of certain groups. The most widely used classification above the family level is that established by G.K. Noble in 1931, in which five suborders were recognized, based on vertebral characteristics. The superfamilies given above are not exactly equivalent to the suborders of Noble, the following modifications being most noteworthy: (1) The names Ascaphidae and Leiopelmatidae are interchangeable, the first being used by most American authorities, the latter by most others. (2) The family Myobatrachidae includes what were formerly recognized as Old World members of the Leptodactylidae, a family now considered restricted to the New World. (3) The Brachycephalidae, as recognized by Noble, included the dendrobatids and the peculiar mouth-breeding frog (*Rhinoderma*), now considered distinct families, as well as several genera (*e.g.*, *Atelopus*) currently placed in the Bufonidae. Most authorities now restrict the family Brachycephalidae to the single genus *Brachycephalus*. (4) The family Atelopodidae (sometimes spelled Atelopidae) of many authors is now relegated to the Bufonidae. (5) *Rhinoderma*, which other authors have placed in the Atelopodidae, Leptodactylidae and Bufonidae, is now given family status. (6) The relationships of the Microhylidae are uncertain, and their placement in the suborder Ranoidea must be considered tentative. This family may have had an evolutionary history entirely independent from the ranoids.

BIBLIOGRAPHY. D.M. COCHRAN, *Living Amphibians of the World* (1961), a good general account abundantly illustrated with excellent black-and-white and colour photographs; R. CONANT, *A Field Guide to Reptiles and Amphibians* (1958), an excellent identification manual for frogs in the eastern United States, illustrated in colour and black and white; W.E. DUELLMAN, *The Hylid Frogs of Middle America*, 2 vol. (1970), a detailed study of the taxonomy, distribution, life history, and ecology of the tree frogs of Mexico and Central America; C.M. BOGERT, "The Influence of Sound on the Behavior of Amphibians and Reptiles," in W.E. LANYON and W.N. TAVOLGA, *Animal Sounds and Communication*, pp. 137–320 (1960), a thorough account of the calling behaviour of frogs; G.K. NOBLE, *The Biology of the Amphibia* (1931), a classic compilation of knowledge on frogs and other amphibians; H.W. PARKER, *A Monograph of the Frogs of the Family Microhylidae* (1934), a thorough account of one of the most interesting families of frogs; R. RUGH, *The Frog: Its Reproduction and Development* (1953), a good account of larval development; R.M. SAVAGE, *The Ecology and Life History of the Common Frog* (1961), a well-documented study of the biology of *Rana temporaria* in England; R.C. STEBBINS, *A Field Guide to Western Reptiles and Amphibians* (1966), an excellent identification manual for frogs in the western United States, illustrated in colour and in black and white; A.H. and A.A. WRIGHT, *Handbook of Frogs and Toads of the United States and Canada*, 3rd ed. (1949), a well-documented account of North American frogs and toads with many black-and-white photographs.

(W.E.D.)

An-yang

An-yang is a city of eastern China (36° 05′ N, 114° 21′ E; population [1970 est.] 100,000–300,000) in northern Honan Province, on the An-yang Ho (river), a tributary of the Wei Ho (river), which flows northeast to join the Grand Canal (Yün Ho). An-yang is a county- (hsien-) level municipality and the administrative centre of An-yang area (ti-ch'ü). Under the Chinese Empire, it was known as Chang-te or Changte-fu. The city is about 65 miles northwest of the Huang Ho (Yellow River) in the foothills of the T'ai-hang Shan (mountains); to the west towers the elevated massif of Shansi Province.

An-yang has been a regional agricultural and trade centre for centuries; in modern times, it became a station on the main north–south rail line from Peking to Wu-han. Coal mining is important in the area (mines at Liu-ho-kou, about 20 miles northwest of the city). An-yang's textile mills and food-processing plants have been supplemented by heavier manufactures since the Communists came to power (1949).

The district's significance, however, is pre-eminently historical and archaeological; west of the city is the site of the last capital of the Shang dynasty, from which the kings of Shang ruled China for 273 years (1384–1111 BC). The Shang palace stood about ten miles west of the site. Early historians gave the name Yin to the site of An-yang; and thus the Shang dynasty was also recorded by later historians as the Yin dynasty or the Yin-Shang dynasty. Shang was the second of the three earliest dynasties of China (the first was the Hsia dynasty; the third, the Western Chou). Modern archaeology has discovered in the An-yang district the historical remains of the Shang, fully illustrating the splendour and advanced stage of a previously totally unknown civilization that developed more than half a millennium before the beginning of authentic Chinese history in 776 BC. The discovery, including the earliest Chinese written records, has provided the most authentic evidence to date of the ancient civilization of China.

The site of the Shang capital at An-yang had been known to scholars since the beginning of this century through the accidental discovery of inscribed oracle bones (*i.e.*, bones used for divination: questions were engraved on the bones, which were then touched with a burning stick, and resulting cracks revealed the answers); but it was not until 1928 that the first organized scientific expedition started systematic excavation of these important remains under the auspices of Academia Sinica, organized by the Nationalist government of the Republic of China. Field work was carried out by Dr. Li Chi, later the director of the Institute of History and Philology. His party worked at the site for 15 seasons, from 1928 until the Japanese invasion in 1937.

The first organized scientific excavation

During this period, the party surveyed most of the early remains along the Huan River, which flows across An-yang. Two sites located near the bank of the river proved to be the most important: the archaeologists spent 12 seasons at one of the sites—locally known as Hsiao-t'un and situated on the southern bank—where they uncovered the remains of the Shang capital city, and 3 seasons at the other site—locally known as Hou-chia-chuang and located on the northern bank—where they discovered a number of royal tombs of the dynasty, together with more than 1,000 small burial pits.

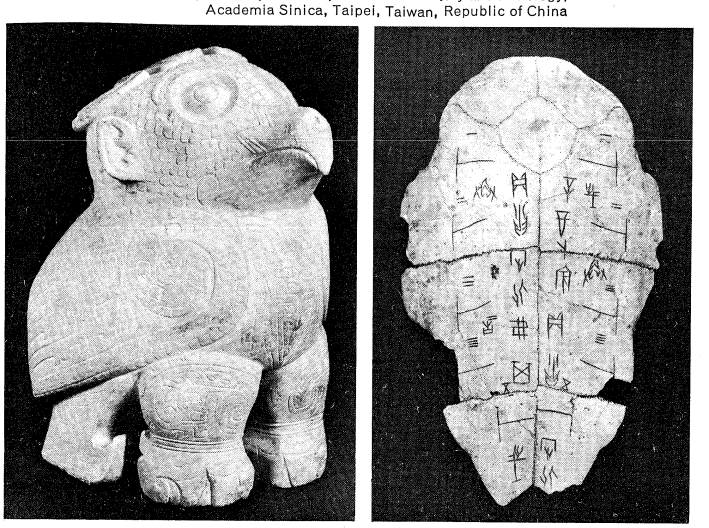
From their excavations at these two sites, the field workers made many significant discoveries, among which were many thousand pieces of inscribed tortoise shells and shoulder blades of oxen; this discovery confirmed the authenticity of the "oracle bone inscriptions" (as they have been known in the academic world), which for more than a quarter of a century had been assiduously studied by distinguished scholars but had been discredited as sheer forgeries by a number of others.

The written records inscribed on the oracle bones have added greatly to the knowledge of early Chinese history and have served as a check on the traditional accounts about ancient China handed down from before the Han dynasty (began 206 BC). The contents of the inscribed bones are rich and variegated. Although dates on the bones are only relative, they show several moon eclipses, which have been used to check traditional accounts of the calendar of the Yin people. On the basis of these records, the Yin calendar was reconstructed, and it is believed that the dynastic capital was moved to the site in the 15th year of King Pan Keng's era (*c.* 14th century BC).

Discoveries at An-yang

The field expedition of Academia Sinica discovered extensive architectural remains and thousands of tombs, most of which had been plundered ruthlessly in the 3,000 years since their construction. At the cemetery site, there were 11 large burials, each occupying a huge underground structure built of timber and stamped earth. Practically every one of the tombs had been entered by robbers more than once, so that the most valuable part of the royal treasuries had been taken away long before scientific digging started. Nevertheless, the archaeologists were still able to gather from both the royal tombs and the dwelling site a rich harvest of important and untouched artifacts, most of which, after decades of detailed studies by specialists, proved to be contemporaneous with the period of the oracle bone records.

By courtesy of Li Chi, Institute of History and Philology, Academia Sinica, Taipei, Taiwan, Republic of China



(Left) Owl, stone carving in white marble, and (right) oracle bone inscription on tortoise shell were excavated at An-yang. In the Institute of History and Philology, Academia Sinica, Taipei, Republic of China (Nationalist).

The important artifacts recovered from the two sites may be grouped as follows; (1) building foundations, (2) animal bones, (3) stone and jade, (4) pottery, (5) bronzes, (6) chariots, (7) human skeletons, (8) imprints of perishable objects such as a snakeskin drum and textile fabrics on the covering earth soil. Architectural remains are divisible into two subgroups: those found in Hsiao-t'un were mainly foundations of palaces, temples, and living quarters, of which all the superstructures had completely disappeared; those from Hou-chia-chuang were all tomb structures buried deep underground, the largest of which occupied a floor space measuring about 725 square metres and was several metres high. The area appears to have been rich in the quantity and variety of its fauna; certain animals, like buffalo and deer, whose bones were found in enormous quantity, are now extinct in the region. The most interesting specimens were the bones of animal and bird species now known only in tropical or subarctic habitats (evidently there was a royal zoo in which animals from all parts of the world were collected). Pottery shards were abundant, with nearly 250,000 pieces recorded in the field; from these, more than 1,700 complete pots have been reconstructed. The bronzes of the period are known to collectors all over the world, and scientific diggings uncovered hundreds of ritual bronzes in their original positions. These ritual vessels were found in every unplundered grave. Bronze weapons were also found in great number, consisting chiefly of Chinese halberds, spearheads, helmets, arrowheads, hatchets, knives, and swords. A third group of bronzes was composed mainly of decorative pieces and small tools. Chariots were found in both the cemetery sites and dwelling sites; some were decorated extensively with bronze ornaments and fittings. The archaeologists found extensive evidence of the practice of human sacrifice, which was characteristic of Bronze Age cultures.

Stone carvings were among the surprise finds; before these excavations, scholars had been unaware of the existence of the stone cutter's art in China during this early period. Although Yin-Shang China had reached an advanced stage of bronze casting, stone tools still appeared in great abundance. Genuine jades were also found, but most of them were small articles—decorative pieces or merely amulets; some larger ones may have borne important religious significance and may have been used only for ceremonial purposes.

BIBLIOGRAPHY. CHI LI, *The Beginnings of Chinese Civilization* (1957), three lectures dealing mainly with finds at An-

yang; TE-K'UN CHENG, *Archaeology in China*, vol. 2, *Shang China* (1959), materials covering many aspects of the An-yang discoveries that appeared in the original Chinese reports, translated by the author into English; EDITORS OF TIME-LIFE, INC., "The Shang, Molders of China," in *The Epic of Man*, ch. 11 (1961), a summary of the An-yang finds beautifully illustrated in colour; KWANG-CHIH CHANG, *The Archaeology of Ancient China*, ch. 6 (1963), mainly on the An-yang discoveries. The major reports in Chinese are all published in the two series of *Archaeologia Sinica;* the volumes of the later series are all appended with an extensive English summary.

(L.Ci.)

Apennine Range

The Apennine Range, bordered by narrow coastlands, forms the physical backbone of peninsular Italy and has had considerable influence on the human geography of that nation. From the Colle di Cadibona in the northwest, close to the maritime Alps, it forms a great arc, which extends as far as the Isole Egadi (islands) to the west of Sicily. Its total length is approximately 870 miles (1,400 kilometres), and its width ranges from 25 to 125 miles. Monte Corno, 9,560 feet (2,914 metres), is the highest point. The range, popularly known as the Apennines (Appennino), follows a northwest to southeast orientation as far as Calabria, at the southern tip of Italy; the regional trend then changes direction, first toward the south, and finally westward.

The Apennines are among the younger ranges of the Alpine system and, geologically speaking, are related to the coastal range of the Atlas Mountains of North Africa. Similarities have also been observed with the Dinaric Alps, which extend through Yugoslavia and Greece. The nearby islands of Sardinia and Corsica, on the other hand, are dissimilar to the Apennines, their granitic rock masses being linked to those of Spain. Geographically, the Apennines differ strikingly from the Alps. They have lower altitudes and lack glaciers, and the maritime influence on the valleys, and indeed overall on the exposed slopes, is much greater. Although several important agricultural and industrial activities are associated with the Apennines, their economic significance is as yet minor compared to that of the Alps.

Various aspects of the Apennines—their geology, hydrography, zoology, and botany—have been studied by the leading Italian universities, the Italian Geological Survey, and such bodies as the National Research Council of Italy and the Hydrographic Service of the Ministry of Public Works. In spite of these efforts, and of the current and past researches of a number of French, Dutch, Belgian, and German scientists, several sections of the Apennines—notably Calabria, Sicily, and other southern areas—remained little investigated at the start of the 1970s. (For related information see ITALY.)

The physical environment. Geologically, the most ancient sections of the Apennines are the granitic and metamorphic (heat-altered) rock masses, which outcrop in Calabria and Sicily. Represented by the La Sila, Aspromonte, and Peloritani mountains, these masses all date from the Hercynian mountain-building period, which occurred about 230,000,000 years ago, in the Upper Paleozoic Era. Ancient sedimentary rocks are virtually absent, whereas thick layers of sedimentary rocks of more recent geological origin are prominently represented. Notable among these are the shale, sand, or calcium-bearing deposits that were laid down in a great Mediterranean geosyncline (a downward warp in the earth's crust) during Mesozoic and Cenozoic times, from 200,000,000 to 35,000,000 years ago. The Apennine mountain-building period of Miocene and Pliocene times, about 25,000,000–10,000,000 years ago, uplifted these sediments, with much folding and faulting and also horizontal overthrusting of considerable length in the direction of the Adriatic and Ionian seas to the east. The deformation of the earth's crust on a continental scale is still uplifting the Apennines at a rate of approximately a millimetre a year. On the western, or Tyrrhenian, edge of the Apennines, severe faulting over the last 2,000,000 years gave birth to a line of volcanoes that runs from Mt. Amiata in Tuscany as far as Mt. Etna (*q.v.*) in Sicily, passing through the volcanic lakes of Bolsena, Vico, Bracciano, Albano, and Nemi, and continuing through the volcanoes of Roccamonfina (extinct), Vesuvius (*q.v.*), and the Isole Eolie. Seismic movements are frequent, especially within the central and southern Apennines and Sicily. Major earthquakes are caused by structural settling of the young mountain chain, movements along the abundant faults, and by volcanic activity.

Earthquake zone

The geological youth of the Apennines, and a great variety of rock type, are responsible for the rugged appearance of the range today. In the north, in Liguria, sandstones, marls, and greenstones occur. Landslides often occur in these brittle rocks. In Tuscany, Emilia, Marche, and Umbria, clay and sand characterize the rocks. In Lazio, Campania, Puglia, Calabria, and northern and eastern Sicily, there are large calcareous rock outcrops, separated by lowland areas of shale and sandstone. In Abruzzi e Molise, Basilicata, and southeastern Sicily extensive argillaceous (clayey) rock types occur. Here, the landscape has a thirsty and desolate appearance, with frequent erosion of the *calanchi*, or badlands, type.

Starting from the north, the main subdivisions of the Apennines are: the Appennino Tosco-Emiliano (maximum height 7,096 feet [2,163 metres]), the Appennino Umbro Marchigiano (8,130 feet [2,478 metres]), the Appennino Abruzzese (7,198 feet [2,914 metres]), the Appennino Campano–Lucano–Calabro (7,451 feet [2,271 metres]), and, finally, the northern Sicilian range (6,493 feet [1,979 metres]). Puglia and southeastern Sicily are formed by low, horizontal limestone plateaus, which remained unaffected by the Alpine mountain-building epoch.

The rivers of the Apennines have short courses: the principal ones are shown in the accompanying table.

River Lengths

name	length		name	length	
	mi	km		mi	km
Tevere (Tiber)	252	405	Basento	93	149
Arno	150	241	Pescara	90	145
Reno	131	211	Imera	89	144
Volturno	109	175	Ofanto	83	134
Ombrone	100	161	Flumendosa	79	127
Garigliano	97	158	Bradano	72	116
Tirso	93	150	Sangro	71	115
			Simeto	70	113

In spite of the limited length of the rivers, the action of running water is the chief agent of erosion responsible for molding the contemporary Apennine landscape. The character of the physical geography depends on the varying nature of the rocks in each region, and their resistance to water action. The overall aspect of relief, however, exhibits characteristics of an early, or juvenile, stage in the cycle of erosion. In limestone areas, karst erosion, with crevasses worn by water action, predominates. In the highest part of the Apennines there are traces of the erosive action of the glaciers of the last Ice Age, although, unlike the Alps, contemporary glaciers are lacking.

Landscape formed by erosion

Lakes—which today are small and scattered in distribution—were also much more abundant in Quaternary times. The alluvial lake of Trasimeno (49 square miles [128 square kilometres]) is the largest lake of the present range. Other natural lakes, of varying origin, are scattered throughout the range. There are more than 200 artificial lakes created for purposes of power and irrigation.

The climate of the highest section of the Apennines is continental (as found in the interior of Europe) but ameliorated by Mediterranean influences. Snowfalls are frequent, with cold winters and hot summers (average July temperature 75°–95° F [24°–35° C]). Average rainfall (at between 40 and 80 inches [1,000 and 2,000 millimetres] per year) is higher on the Tyrrhenian slopes than on the eastern, or Adriatic, side of the Apennines.

Vegetation, animal life, and human geography. The flora of the Apennines is Mediterranean in type and varies with both latitude and altitude. In the north, woodlands with oak, beech, chestnut, and pine predominate. To the south, ilexes, bays, lentisks, myrtles, and oleander (a flowery evergreen herb) abound. Prevailing crops are represented by the olive trees, growing to a height of about 1,300 feet above sea level; citrus fruits, which are well developed in Calabria and Sicily; and grapes, which are found in abundance in Tuscany, Lazio, and Puglia. Other products of the range include sugar beets (in the plain of Emilia), potatoes, vegetables, and fruit. The importance of corn is diminishing with the depopulation of hill farms that is characteristic of the region. In the highland areas, pasturing remains the main form of land utilization.

Apennine fauna has been little studied. There are some insects found exclusively within the area, and the brown bear, the chamois, the wolf, and the wild boar may also be noted. In Sicily the wild rabbit is important. Sheep, goats, asses, and mules are bred in the higher altitudes.

In the foothills of the Apennines, secondary industries are widespread, while primary industries are rapidly expanding in the adjacent coastal plain, often in association with important discoveries of natural gas. Such minerals as mercury, sulfur, boron, and potassic salts are also of significance, while the marble quarries of the Apennines have been famous for centuries.

Settlement patterns

Since prehistoric times the Apennines have been the home of Italic peoples. Today, the highest village settlement is found at about 4,500 to 5,000 feet above sea level, at the upper limit of cultivated land. More densely populated areas are the wide river valleys, which are rich in alluvial and cultivated land (*e.g.*, the valleys of Lunigiana, Garfagnana, Casentino, Val Tiberina). Internal basins (Foligno, Terni, Rieti, l'Aquila, Sulmona, Avezzano) are also well populated. Rural depopulation, resulting from the lack of development of the Italian south and the attraction of industrial areas in northern Italy and elsewhere in Europe, had, by the 1960s, reached major proportions. By the late 1960s, however, this emigration had slackened, mainly as a result of attempts to develop the local economy.

The Apennines are crossed by ten railway lines, some of them double-tracked. There are numerous roads providing access to the range, although the rugged terrain makes for difficulties. The Highway of the Sun (Autostrada del Sole, completed in 1964), which has overcome the barriers of relief with an imposing series of tunnels and embankments, is the main artery of peninsular Italy, and one of the great scenic routes of Europe. (B.A.)

Apodiformes

The order Apodiformes contains two groups of birds that are very different from one another in general appearance and way of life: the swifts and the hummingbirds. The two groups, considered suborders, are Apodi, which contains two families, Hemiprocnidae for the tree swifts (also called crested swifts) and Apodidae for the true swifts; and Trochili, which contains the single large family Trochilidae for the hummingbirds. The most obvious common characteristic of the two groups is a superb mastery of the air; both swifts and hummingbirds are agile fliers, and share, in this connection, some peculiarities of wing structure. Some ornithologists believe that these anatomical resemblances are convergent (the result of the same type of natural selection acting on unrelated organisms) and that the true relationships of hummingbirds and swifts lie with other groups, not with each other.

General features. *Importance to man.* Man has long been fascinated by hummingbirds. From the time the first specimens arrived in Europe from the newly discovered American tropics, the brilliant, iridescent colours and diminutive size of hummingbirds have made them favourites of both scientists and the public. Stuffed hummingbirds were in great demand for display in homes and museums and for decorating ladies' hats. South American Indians brought hummingbirds from hundreds of miles around to great market centres such as Bogotá, Colombia, and Cayenne, French Guiana. Even today, some species and subspecies of hummingbirds are known only from these "trade skins," and their actual range has never been discovered by ornithologists. The commerce in hummingbird skins has almost completely vanished, and collectors now capture hummingbirds alive for exhibition in zoos and aviaries. Although some species are surprisingly hardy in captivity, the proper care of hummingbirds is not simple, and few of the birds that leave the jungle survive.

In most parts of the world, swifts, little known to the average person, are often confused with swallows, which, although unrelated, are generally similar to swifts in size, proportions, and aerial habits. In Southeast Asia, however, one group of swifts is of major economic importance. The swiftlets of the genus *Collocalia* comprise a group of species, most of which live in caves. They build nests composed of varying amounts of plant and animal substances (*e.g.*, leaves, moss, hair, feathers) held together and fastened to the cave wall with a mucilaginous secretion of the salivary glands. The nest of one species, the edible-nest swiftlet (*C. fuciphaga*), is composed almost entirely of concentric layers of this salivary cement. These nests and, to a lesser extent, those of some other swiftlets are gathered commercially in the East Indies and form the base for the famous bird's-nest soup of the Orient.

Bird's nest soup

Size range and diversity of structure. The smallest of living birds are included among the Trochilidae, the smallest known species being the bee hummingbird (*Mellisuga helenae*) of Cuba, which is barely 62 millimetres (two and one-half inches) from bill tip to tail tip. Even the largest member of the family, the giant hummingbird (*Patagona gigas*) of western South America, has a body only about the size of that of a large sparrow, although the bill and tail extend this length by about 35 and 85 millimetres (one and one-half and three and one-half inches), respectively. There is relatively little structural diversity among hummingbirds except in the size and shape of the bill and the ornamental feathers. The bill may be very short or, as in the sword-billed hummingbird (*Ensifera ensifera*) of the Andes, may exceed the rest of the bird in length. The bill may also be straight or curved to varying degrees, usually downward, but, in a few species, upward. In most hummingbirds, the tail feathers are relatively unspecialized; but in some they are greatly elongated, either straight or curved, in others wirelike with spatulate tips. Brilliantly iridescent feathers are found somewhere on the body of most hummingbirds, often with an especially colourful gorget on the throat; a few genera have dull coloration. Sexual dimorphism (difference between sexes) varies, with females being either indistinguishable from males or less colourful or ornate.

Variety of swifts

The true swifts have a somewhat greater size range, from such tiny species as the five-gram (0.2-ounce) pigmy swiftlet (*Collocalia troglodytes*) of the Philippines to some of the large and powerful members of the Old World genus *Apus*, 30 times heavier. Beyond the size differences, the most obvious morphological variation among swifts is in the conformation of the tail. The spine-tailed swifts, a group that includes the familiar American chimney swift (*Chaetura pelagica*), have the central shaft of the tail feathers extended beyond the vanes (flat, expanded part) into pointed spines. In other swifts, with less specialized tail feathers, the tail itself may be square ended, moderately forked, or deeply forked (swallow-tailed). Less conspicuous structural variations that are used in swift taxonomy include the position of the toes and the amount of feathering on the legs. The four species of tree swifts that comprise the family Hemiprocnidae are rather uniform, differing mainly in size, proportions, and plumage pattern.

Distribution and abundance. Although there are unrelated Old World birds, such as the sunbirds (Nectariniidae), that superficially resemble hummingbirds, the family Trochilidae is confined entirely to the Americas. Of approximately 320 known species, most are tropical, and

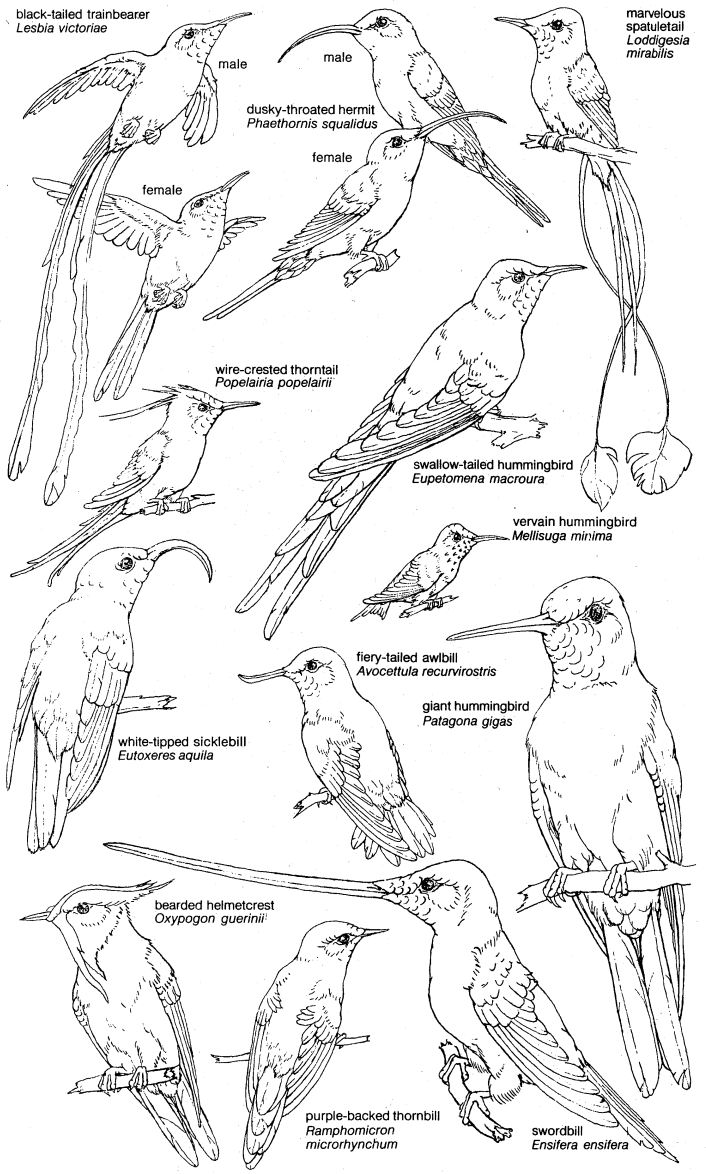


Figure 1: Body plans of hummingbirds.
Drawing by Christian D. Olsen

the greatest numbers of these are found in northwestern South America (Colombia and Ecuador). Several species have become adapted to temperate conditions, some by having extended their distributions north or south of the tropics, others by their occupation of high-altitude habitats in South and Central America. In North America only one species, the ruby-throated hummingbird (*Archilochus colubris*), is normally found east of the Great Plains. About a dozen occur regularly in the western United States; one, the rufous hummingbird (*Selasphorus rufus*), reaches southern Alaska.

The family Hemiprocnidae is confined to Asia, from India south and east through the East Indies, with one species reaching as far as the Solomon Islands. The Apodidae are much more widely distributed, occurring in both the Old and New Worlds. There are approximately 70 to 80 species, but the features that distinguish closely related species are often so subtle that there is little agreement about swift classification at either the genus or species level. Virtually nothing is known about some species other than their existence as evidenced by museum specimens; such birds may be rare, shy, or both. Others, such as the chimney swift, the European swift (*Apus apus*), and the house swift (*Apus affinis*) of India, are familiar inhabitants of towns and have been extensively studied by biologists.

Natural history. *Behaviour and locomotion.* In such overwhelmingly aerial birds as swifts and hummingbirds, locomotion is an inseparable part of behaviour. Although both groups depend on their flying ability to obtain food, the actual feeding methods are quite different. Swifts feed exclusively on flying insects, sharing with the unrelated swallows several morphological adaptations for this habit (*e.g.*, the short, weak bill, large mouth, and long, narrow wings). The feet of the true swifts are much weaker in proportion to body size than are those of swallows or hummingbirds, and the former are unable to perch on twigs, wires, or other similar landing places. In their normal resting posture the true swifts cling to vertical surfaces except when sitting on the nest. The tree swifts are somewhat less specialized and are fully capable of perching; in fact, their usual foraging method involves launching after flying prey from the vantage point of a high perch. Swifts drink by swooping down at the surface of a body of water, and nesting material may be picked in midair, as in the case of drifting feathers or plant material, or snatched in full flight, as are dead twigs and grass stems. Most courtship displays that have been described for swifts are wholly aerial, and copulation is known to take place in full flight.

Some species of swiftlets (*Collocalia*) that nest in total darkness deep in caves have the ability to orient themselves in flight by echolocation (reflection of sound waves back to the emitter), a phenomenon otherwise known, among birds, only in the oilbird (*Steatornis caripensis*, of the order Caprimulgiformes). The swiftlets emit a rattling call consisting of a succession of clicks, which, unlike the calls used by bats for the same purpose, are fully audible to the human ear. A few species of swiftlets lack this ability and nest only where some light penetrates the caves.

Feeding behaviour of hummingbirds

Hummingbirds are less constantly aerial than swifts and often spend long periods quietly perched. In foraging, however, most species depend upon their unique ability to dash from place to place and then to hover, motionless except for beating wings and darting tongue. They can even fly backward or upside down for short distances. The basic food is flower nectar, and the shapes of hummingbird bills are sometimes closely correlated with the shapes of the flowers from which they feed. The long, tubular tongue serves also for sucking up dew or rainwater as needed. Hummingbirds also eat small insects and spiders, which usually appear to be captured with the bill rather than with the tongue. Although hummingbirds perch, they share with swifts the lack of walking ability or other nonflying locomotion. They may sometimes sidle on a perch, just as swifts may crawl up a wall, but normal locomotion is by flying, even for short distances.

Prenesting, or courtship, behaviour is varied in hummingbirds; territoriality is completely absent in some species; others fiercely defend nesting or feeding territories. Hummingbirds in general tend to be aggressive and may attack vastly larger potential predators, such as crows and hawks. Leks, areas of communal male displays, have been described in some tropical species. Most displays are aerial, and some are spectacular; in other species, such as the plainly coloured hermits of the genus *Phaethornis*, singing forms an important part of display behaviour.

Reproduction and life cycle. The nests of tree swifts are tiny cups, barely large enough to hold the single egg. They are placed on small limbs, high up in trees, where they are subject to being moved about by the wind. The egg is glued to the bottom of the nest with the sticky saliva of the parent; this prevents the egg from being turned during incubation (as is done by virtually all other birds), and it is thought that the waving of the nest branch in the wind may in some way act as a substitute, moving the egg's contents within the shell. One of the true swifts, the palm swift (*Cypsiurus parvus*) of the Old World tropics, also glues its egg into the nest, but the nest is very different. This species utilizes feathers and plant material and attaches the nest, little more than a flat pad, to the underside of a hanging palm leaf, where it, too, is subject to agitation by the wind.

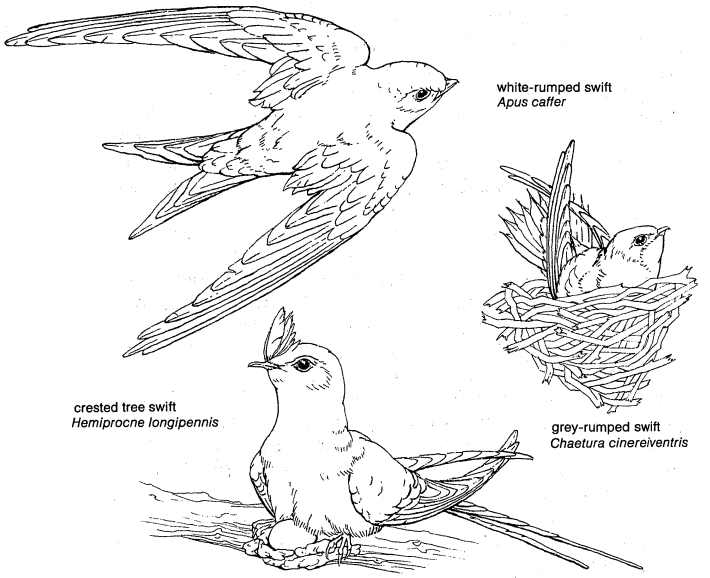


Figure 2: Body plans of swifts.
Drawing by Christian D. Olsen

Variations in swift nesting habits

For such a relatively uniform group in a structural sense, the true swifts show a surprising variation in nesting habits. The New World palm swifts (*Tachornis*), like those of the Old World, place their nests under palm leaves, but the nest itself is more elaborate, being a sack entered from below through a tubular tunnel. Most other swifts nest either inside a crevice or hole (*e.g.*, in the ground, in a tree, between rocks), where they build a shallow cup of feathers and plant material as a flooring, or inside hollow trees, caves, or such man-made structures as buildings, wells, and tunnels. The nests may be circular cups placed on ledges or semicircular brackets attached to the walls. The bracket nests are fastened to vertical surfaces with a sticky secretion from the large salivary glands; in the swiftlet genus *Collocalia*, this substance may comprise most or all of the nest material itself. Once species, the white-naped swift of Mexico, builds no nest at all but lays its eggs in a depression on bare sand on ledges deep inside caves.

Clutches in the true swifts vary from one to about six white eggs, with the higher numbers being found among some of the more northern species of the genera *Chaetura* and *Aeronautes*. Incubation is by both sexes. The young are hatched completely naked. Young swifts are left unattended for longer periods than is true of most altricial (helpless) nestlings; the adults spend a substantial amount of time foraging for the relatively small insects on which they feed, in order to capture a sufficient number to warrant a trip back to the nest. In bad weather, when few insects are flying, the young may be left unbrooded for hours. In some birds, this would normally lead to chilling and, perhaps, death, but young swifts maintain an essentially poikilothermic (cold-blooded) condition for the first few weeks of life, the body temperature dropping to near that of the environment without harming the young bird. In one genus, *Cypseloides*, the young develop a coat of downlike semiplumes (fluffy feathers) prior to the appearance of the juvenile pennaceous (smooth) contour feathers, and this fluffy coat also appears to aid in thermoregulation.

Many species of swifts are colonial nesters, and even those that are relatively solitary during the breeding season appear to be highly gregarious at other times of year. Night roosts of some species may number hundreds of individuals, and foraging flocks may include several species. The northernmost species, such as the European swift and the chimney swift, are long-distance migrants. The winter home of the latter species was a mystery for many years, until Indians in the upper Amazonian forests of Peru were found to have in their possession numbered aluminum bands that had been placed on the legs of chimney swifts by the United States Bureau of Biological Survey. Migration is also known to occur in some of the tropical species, notably in Africa.

The Indian population of the house swift has two breeding seasons per year and is one of the few birds in the world in which this phenomenon has been demonstrated. Adult males are in full breeding condition in late January and again in May and June; eggs are laid in January and February and again in June to September. The molt cycle, however, appears to be independent of the breeding cycle, and the flight feathers are renewed only once a year.

Nest design among hummingbirds

Unlike swifts, hummingbirds are largely polygamous. Males may build nests, but they are not used by the females. Most hummingbird nests, saddled on a branch, are composed of fine vegetable fibres, spider webs, fragments of bark, and lichens. Among the most interesting exceptions are the nests constructed by the hermits. They are long, slender, woven baskets of plant fibres and down, suspended from a strip of vegetation attached at only one side of the nest. The weight of the incubating female and, later, the weight of the young is counterbalanced by pebbles or clods of hard earth at the bottom of the dangling nest. The usual hummingbird clutch is two white eggs, and the female alone incubates and cares for the young in most species, although this is difficult to determine in the species that are not sexually dimorphic. In the sexually dimorphic species, the juveniles usually resemble the adult females, although some genera have a distinctive juvenile plumage that resembles neither parent. The young are fed by the parent thrusting its bill deep within the young bird's throat and regurgitating food.

Like the northern swifts, the hummingbirds that breed farthest north are the most highly migratory. The ruby-throated hummingbird was known to make annual nonstop flights across the Gulf of Mexico long before physiologists understood avian metabolism well enough to determine how this could be possible for such a tiny bird. Migration takes place among some Southern Hemisphere hummingbirds, such as the giant hummingbird and the green-backed fire crown (*Sephanoides sephaniodes*); the latter reaches Tierra del Fuego and is thus the southernmost hummingbird, but there is little information about the nature and extent of these movements.

Ecology. The distribution of both swifts and hummingbirds is strongly influenced by the specialized nature of their foods. Swifts are found virtually worldwide but are limited to areas where flying insects are found in sufficient numbers to sustain adults and young. Food of this kind is presumably more abundant in the tropics, and the family is primarily a tropical one; it is chiefly in the tropics that sympatric species (*i.e.*, those whose ranges overlap), sometimes closely related, may be found. Although there may be some altitudinal and other ecological separation of sympatric species, virtually nothing is known about possible partition of environmental resources among swifts; as many as three closely similar species of *Chaetura* can be seen foraging together in Trinidad. Swifts also forage with swallows.

Hummingbirds are more specialized in their feeding habits, in that their bill shapes are, in many cases, correlated with corolla (collective-petals) shapes of the most commonly utilized flowers. Because hummingbirds are restricted almost entirely to feeding at flowers, their distribution is affected by flowering periods, and the virtually seasonless tropics, where flowers of some kind may be found the year round, are, not surprisingly, the home of the most species. Despite the degree of specialization found, a flowering tree in the tropics may attract six species of hummingbirds simultaneously. In some cases feeding areas are strongly defended territories; in others the birds feed side by side, apparently ignoring each other completely.

Lowered metabolic levels in resting birds

Torpor. Swifts and hummingbirds are among the few groups of birds that have been shown to be capable of energy conservation through a reduction of body temperature and entry into a torpid (sluggish) condition. The extremely high metabolic rate of active hummingbirds suggests that the value of torpor to the birds lies in energy conservation while roosting at night. In the swifts,

whose normal metabolic rate approaches that of other birds, the lowered body temperature is an energy-conserving mechanism during what may sometimes be long periods of fasting, when foraging for flying insects is not possible.

Evolution, paleontology, and classification. *Fossil history.* As might be expected of such fragile birds, no fossil species of hummingbirds are known, although there are two living species of which remains of Pleistocene origin (less than 2,000,000 years old) have been reported. Five fossil swifts have been described; two from the lower Miocene age (about 20,000,000 years ago) and one from the upper Miocene (about 10,000,000 years ago) have been assigned to modern genera. The extinct genus *Cypselavus* is known from two fossil deposits in France: one species from the lower Miocene, the other from the upper Eocene or lower Oligocene (35,000,000–40,000,000 years ago). Their identification as swifts does not seem to be questioned, but they throw no light on the origins of the family. The fossil family Aegialornithidae, sometimes provisionally assigned to the Apodiformes, is now usually considered to belong to the order Caprimulgiformes.

Classification. As presently constituted, the subdivisions of the order Apodiformes are summarized below.

ORDER APODIFORMES

Characterized primarily by the shape and proportions of the wing. The humerus (bone of the upper arm) is short, the hand is strikingly long, and the primary flight feathers (attached to the hand) are much longer than the secondaries (attached to the forearm). Other anatomical characteristics shared by swifts and hummingbirds include the holorhinal (the posterior border of the internal openings rounded) and impervious (without external openings) condition of the nostrils; single (left) carotid artery (except in swifts of the genus *Cypseloides*, which have 2); sternum without notches; and oil gland unfeathered.

Suborder Apodi

Bill short and gape (mouth) deeply cleft; tongue short; salivary glands large; crop absent; nostrils without opercula (coverings); 8–11 secondaries; 6–7 pairs of ribs.

Family Hemiprocnidae (tree swifts)

Hallux (hindtoe) directed backward, not reversible, foot capable of perching; no claw on manus (hand). Hatched with natal down; a distinct juvenile plumage unlike that of adults; adult plumage with crests or plumes on head. One genus, 4 species.

Family Apodidae (true swifts)

Approximately 70–80 species; species limits, especially in *Collocalia*, highly controversial. Hallux lateral or nearly so, reversible in some; foot incapable of true perching; claw present on hand. Young hatched naked; juvenile little different from adult in plumage; no crests or plumes. One fossil and 8 to 14 Recent genera; 70–80 species.

Suborder Trochili

Bill slender, usually long, gape not deeply cleft; tongue long, tubular, and extensible; crop present in nestling; nostrils with opercula; 6–7 secondaries; 8 pairs of ribs.

Family Trochilidae (hummingbirds)

Characters of suborder. More than 120 genera usually recognized, but a further study will probably result in reduction. Approximately 320 species, nearly 30 of which are known only from trade skins; some may be hybrids.

Critical appraisal. The most fundamental question in the classification of the Apodiformes is whether the swifts and hummingbirds are, in fact, more closely related to each other than to any other birds. Some ornithologists believe that the resemblances in wing structure are convergent and that the other shared characters are relatively widespread among several orders of birds and of little significance in indicating relationships. These authorities argue that the differences between swifts and hummingbirds are so fundamental that the two groups could not have shared a common ancestor recently enough to warrant placement in the same order. Ornithologists holding this opinion would recognize a separate order, Trochiliformes, for the hummingbirds. The Caprimulgiformes (nightjars and allies) are usually considered to be the closest relatives of the swifts, but the placement of the hummingbirds is more difficult. Their closest living relatives, if not the swifts, must almost certainly be among the group of orders usually placed nearest to the Passeriformes (perching birds).

The tree swifts have sometimes been subordinated to the level of a subfamily of the Apodidae. Their many relatively nonspecialized anatomical characters indicate that, although indubitably swifts, they merit separate family status. Two subfamilies, usually Apodinae and Chaeturinae, are customarily recognized in the Apodidae but probably represent an exaggerated dichotomy. The African species *Schoudetenapus myoptilus* has been shown to combine certain characteristics of both subfamilies and may be a relict of an ancient and relatively primitive group of Apodidae. Additional studies may enable ornithologists to redefine subfamilies of swifts.

The classification of swifts at the genus and species level is also fraught with controversy. Part of this represents the traditional battle between advocates of large, inclusive genera and those who prefer to use generic names to call attention to differences; in other classifications, one difficult species is transferred between what are otherwise generally accepted genera. Species limits are especially difficult to determine among the swiftlets (*Collocalia*); there is absolutely no agreement about how many species exist nor what names they should bear. The characters that have long been used to define species of *Collocalia* are mixed, and some have proved to be of no taxonomic importance. Important morphological characters have been overlooked; one swiftlet, first described as a subspecies, was found 26 years later to be the only three-toed form in the genus and was quite properly given species status. The type of nest construction and the presence or absence of the ability to navigate with echolocation appear to be fundamental characters within *Collocalia*, but this information is not available for many forms, and the taxonomy of the genus will long be in a state of flux.

Among the hummingbirds there are obvious groupings of related genera, but no discontinuities are marked enough to warrant the recognition of subfamilies. There are generally conceded to be too many hummingbird genera that are based entirely on characters found only in males, involving ornamentation used in species-specific displays. As yet, however, there has been no research that will permit the redefinition of hummingbird genera.

BIBLIOGRAPHY

Swifts: RICHARD B. FISCHER, *The Breeding Biology of the Chimney Swift "Chaetura pelagica" (Linnaeus)* (1958), the most thorough (excellently illustrated) single-species study of any New World swift; DAVID L. LACK, *Swifts in a Tower* (1956), a popular summary based on numerous scattered scientific papers on *Apus apus*.

Hummingbirds: CRAWFORD H. GREENEWALT, *Hummingbirds* (1960), a beautifully illustrated, popular book, with emphasis on the physics of plumage iridescence and of hummingbird flight; EUGENE L. SIMON, *Histoire naturelle des "Trochilidae" (Synopsis et catalogue)* (1921), the basis of the current classification of hummingbirds, with detailed descriptions, keys, and nomenclatorial synonymies.

(K.C.P.)

Appalachian Mountains

The Appalachian Mountains are one of the great highland systems of North America, the eastern counterpart of the Rocky Mountains. Extending for over 1,200 miles from the Canadian province of Quebec to central Alabama in the United States, they form a natural barrier between the Eastern seaboard and the vast lowlands of the continental interior. As a result, they have played a vital role in the settlement and development of the entire continent. They combine a heritage of natural beauty and a distinctive regional culture with contemporary problems of economic deprivation and environmental deterioration. (For the relationship of the Appalachian Mountains to the geological and ecological development of North America as a whole, see NORTH AMERICA.)

Regional division. The system may be divided into three large regions: northern, central, and southern Appalachia. These include such mountains as, in the northern area, the Shickshocks and the Notre Dame ranges in

Quebec; the Long Range in Newfoundland; the great monadnock of Katahdin in Maine; the White Mountains of New Hampshire; and Vermont's Green Mountains, which become the Berkshires in Massachusetts, Connecticut, and eastern New York. New York's Catskills are in central Appalachia, as are the beginnings of the Blue Ridge range in southern Pennsylvania and the Alleghenies, which rise in southwestern New York and cover parts of western Pennsylvania, western Maryland, and eastern Ohio before merging into the third, or southern region. This area includes the Alleghenies of West Virginia and Virginia; the Blue Ridge range, extending across Virginia and western North Carolina, the northwestern tip of South Carolina, and the northeastern corner of Georgia; the Unakas in southwestern Virginia, eastern Tennessee, and western North Carolina (of which the Great Smoky Mountains are a part); and the Cumberland Mountains of eastern Kentucky, southwestern West Virginia, southwestern Virginia, eastern Tennessee, and northern Alabama.

Altitudes

The highest altitudes in the Appalachians are in the northern division, with Maine's Mt. Katahdin (5,268 feet [1,606 metres]), New Hampshire's Mt. Washington (6,288 feet [1,917 metres]), and other pinnacles in the White Mountains rising above 5,000 feet, and in the southern region, where peaks of the North Carolina Black Mountains and the Tennessee–North Carolina Great Smoky Mountains rise above 6,000 feet, and the entire system reaches its highest summit on Mt. Mitchell (6,684 feet [2,037 metres]).

A distinctive feature of the system is the Great Appalachian Valley. It includes the St. Lawrence Valley in Canada, the Kittatinny, Cumberland, Shenandoah, and Tennessee valleys—the latter the site of the world famous Tennessee Valley Authority, a government agency for natural resource conservation, power production, and regional development.

Exploration and scientific study. The ruggedness of the Appalachians, the transverse ranges by which they are crossed, their maze of streams and rivers, and their lack of natural passes created a formidable barrier to early explorers and settlers. The Spanish conquistador, Hernando de Soto, was probably the first white man to enter southern Appalachia, in 1539–40. In 1716 Governor Alexander Spotswood of Virginia led the first organized body of Englishmen across the Blue Ridge Mountains. During the 1760s and 1770s Daniel Boone became America's frontier folk hero through his exploits in exploring and settling the Blue Ridge and Cumberland Mountain country. Historical figures associated with the opening of northern Appalachia include Samuel de Champlain, who sighted the mountains in 1605 as he sailed along the Maine coast; Darby Field, who made the first climb up Mt. Washington (1642); Timothy Nash, discoverer of the important Crawford Notch (1771); and Sir William Logan, founder of Canada's geological survey, who made a cross section of the geological formation of Gaspé peninsula in 1844 and became the first white man to cross the Shickshock Mountains. During the mid-19th century the first extensive scientific studies of Appalachia began when, in 1849, the Swiss geographer Arnold Guyot commenced mapping the eastern mountains. Starting with the White Mountains, he spent five years in northern Appalachia, then moved south to the Great Smoky Mountains area. He mapped, measured elevation, and made the first methodical effort to name mountains. The highest peak in the Eastern U.S. was named for another pioneer explorer-scientist, Professor Elisha Mitchell, who lost his life in 1857 while establishing the fact of this balsam-covered mountain's preeminence. It remained for Horace Kephart, a St. Louis librarian turned naturalist who came to the southern mountains in 1904, to bring the isolated and scenic region to national attention. From his writings grew the interest and impetus that led to the establishment of the Great Smoky Mountains National Park.

Geological development. The Appalachians are among the oldest mountains on earth, born of powerful upheavals within the earth's crust and sculptured by the ceaseless action of water upon the earth's surface. The two types of rock that characterize the present Appalachian ranges tell much of the story of the mountains' long existence.

First there are the most ancient crystallines. During the Precambrian, between 570,000,000 and 1,100,000,000 years ago, long periods of sedimentation and violent eruptions alternated to create rocks and then subject them to such extreme heat and pressure that they were changed into sequences of metamorphic rocks. Among the oldest of these are the gneisses. Limestone changed into marble, shales became slate and schist, sandstones were transformed into quartzite, and intrusions of magma formed bodies of granite. These ancient rocks antedated most plant or animal life; in addition, the intense pressures and heat destroyed any traces of primitive life —so that the precambrian crystallines contain no trace of fossils. They make up what is known as "Old" Appalachia in Canada, New England, and a belt east of the Great Valley with the Blue Ridge at its heart. To the west, the Great Valley, the Valley Ridges, and the Appalachian plateau (including the Alleghenies) are characterized by the second type of rocks, the Paleozoic sediments. These make up "New" Appalachia—the shales, sandstones, and coals that were created by sediments deposited, stratified, and solidified over a vast range of geological time. During the Mississippian and Pennsylvanian epochs of the Carboniferous Period, this long process included the formation of some of the richest coal beds in the world. During the Permian Period (around 250,000,000 years ago), near the end of the Paleozoic Era, a great mountain folding occurred. This was the Appalachian Revolution, a vast interior crumpling resulting from the stress placed on huge masses of subterranean rock. As parts of the earth buckled into folds, cracked, and faulted, other parts were uplifted—sometimes in the parallel ridges distinctive to the Appalachians—and thrust faults served to move one rock mass atop another. Thus the ancient crystallines were lifted in places above more recent sedimentary rock deposits.

The Appalachian Revolution

Agencies of erosion. In addition to the massive folding of the Appalachian Revolution, however, two other agents—ice and water—have carved the steep ridges and pinnacles and gouged out the deep ravines and valleys of the Appalachians. This building, eroding, uplifting, and shaping of the Appalachians has been a continuous process through the ages. Many of the major rivers are older than the mountains. This accounts for the fact that northeast of the New River in Virginia the major Appalachian rivers flow into the Atlantic Ocean, often through dramatic passages called water gaps, while southwest of the New, the rivers, with few exceptions, flow to the Ohio River. When the mountains were thrust up, blocking their westward course to the ancient sea that once covered mid-America, these old rivers cut out their own routes, creating those spectacular canyons, gorges, and "narrows" that are part of Appalachian scenery.

The glacial legacy

The northern Appalachians were also affected by glacial forces. During the Pleistocene (10,000 to 2,500,000 years ago), continental ice sheets flowed down over North America, covering New England but reaching no nearer the southern Appalachians than the Ohio Valley. These moving tongues of ice stripped topsoil, ground and polished certain peaks, and elsewhere scattered rock debris and random boulders, all the while driving plants and animals farther south in search of survival. Thus the southern Appalachians became the refuge for northern life forms, a giant bed for reseeding when the glaciers retreated and the plants moved slowly north again, leaving behind a rich botanical variety thriving in northern altitudes and southern latitudes.

The New River, rising on the Blue Ridge in North Carolina, runs northward and then turns westward across the Appalachian Valley and the Alleghenies (where it becomes the Kanawha) and empties into the Mississippi Basin. The larger mountain streams to the south, dominated by the Tennessee River, follow this example. Exceptions are the rivers rising southeastward on the Blue Ridge, which flow into the Atlantic, and the Chattahoo-

chee, rising in the northeastern corner of Georgia, which runs southwestward into the Gulf of Mexico.

The entire Appalachian system is laced with an intricate network of springs, streams, waterfalls, and rivers. Abundance of water reaches its zenith in the southern Appalachian mountains. Certain areas of the Blue Ridge receive an annual rainfall of 69 inches during an average year. Elsewhere precipitation is even higher, not exceeded elsewhere in the United States except along the northwest Pacific Coast. Much of this rainfall comes in extremely heavy downpours during short periods of time. Since this region does not have the natural storehouses of numerous lakes and glacial deposits of sand and gravel spread over hills and valleys such as are found in the northern Appalachian region, sudden rainfalls bring rapid rises and falls in the southern Appalachian stream flows. Under certain conditions (as when the forest cover, which serves as a storage sponge, has been destroyed) destructive floods have characterized much of the hydrologic history of this part of the Appalachians. To contain these floods and harness the might of an entire river system for purposes of navigation, power production, land reclamation, and watershed development, the Tennessee Valley Authority was established in 1933 as one of the chief factors influencing the ecology of the Southern Appalachian region. Its system of dams turned a river that rampaged and often destroyed into a river that flows gently and productively. The TVA created a series of spacious reservoirs (the majority of which are in or adjoining the Appalachian region) called "the Great Lakes of the South." These lakes, in turn, have altered the natural and human resources of the region, using Appalachian water power to produce electrical and nuclear power, enlarging industrial and agricultural and recreational opportunities.

The Tennessee Valley Authority

Waterfalls are common throughout much of the Appalachian system. Most of those in the northern Appalachians, especially from New York to Maine, were created when glacial moraine or debris, scraped from surrounding peaks by the melting ice cap, solidified into shelves along creeks or river valleys over which the water must plunge as over a terrace. Southern Appalachian waterfalls were generally formed by the action of water on alternating layers of soft and hard rock.

In the area known geologically as "New" Appalachia, especially where there are the softer limestone rocks to yield to the constant erosion by water, numerous caves are a distinctive feature of the physiography. The chief caverns lie within or bordering the Great Valley region of Pennsylvania, Maryland, West Virginia, Virginia, and Tennessee. Caverns of the Shenandoah Valley in Virginia and the Mammoth Cave in Kentucky provide well-known and dramatic examples of underground passages, rooms, watercourses, formations, and decorations that honeycomb much of the earth below central and southern Appalachia.

A unique man-made feature of a large portion of the Appalachian system is the 2,000-mile Appalachian Trail. This footpath, stretching from Mt. Katahdin, Maine, to Springer Mountain, Georgia, provides a hiker's grandstand on the varied ranges of the Appalachians. Overnight shelters are scattered along the way. A noncommercial motor route, the Blue Ridge Parkway, stretches 469 miles from the Shenandoah National Park in northern Virginia to the Great Smoky Mountains National Park and is the most popular area administered by the U.S. National Park Service.

The Appalachian Trail

Plant and animal life. From Maine to Georgia, the Appalachian Mountain system was once almost totally covered with forest. Today the best and most extensive broadleaved deciduous forests in the world still flourish in the Appalachian Mountains and bordering areas, notably in southern Appalachia. To the north are the conifers (red spruce and Fraser fir, which grow at the highest elevations and distinguish the Canadian and Maine woods) and the northern hardwoods (sugar maple, buckeye, beech, ash, birch, red and white oak). Farther south are hickory, poplar, walnut, sycamore, and at one time the important and, before they were destroyed by blight, plentiful chestnuts. All of these, plus other of the 140 species of trees of Appalachia, are found in the southern mountain region. Lofty elevations nurture representatives of the Canadian forest, while the western slopes of the Great Smokies, receiving an annual rainfall often as high as 90 inches, produce trees that have reached record maximum height and diameter. Among these are the tulip tree, buckeye, Canadian hemlock, and chestnut oak.

Forests

The interdependent system of southern plant growth known as the "Appalachian forest" is highly complex. It forms one of the great floral provinces of the earth. There are the trees that bear luxuriant bloom, such as serviceberry, redbud, hawthorne, tulip poplar, dogwood, locust, sourwood, and many others. Among the numerous shrubs with particularly showy flowers are the rhododendron, azalea, and mountain laurel. Certain summits of the southern Appalachians are "heath-balds"—open meadows or grasslands interspersed with thick growths of heath. Roan Mountain in the North Carolina–Tennessee Unakas is one of the most extensive of these, with some 1,200 acres of natural gardens sprawling vivid rose and pink and purple rhododendron across its high pinnacle and down its slopes. It is estimated that of some 2,000 species of Appalachian flora, perhaps 200 are native to and wholly confined to the southern Appalachians. Ferns, mosses, and mushrooms of many species are also part of the complex Appalachian plant life.

Bison, elk, and wolves, once common to the Appalachians, disappeared long ago; caribou and moose are still found in the northernmost corners of the region. Scattered through other areas are the black bear, white-tailed deer, wild boar, fox, raccoon, beaver, and numerous other small animals. All areas of Appalachia, from Gaspé to Georgia, support an abundant bird life. In the Great Smoky Mountains alone some 200 varieties of game and songbirds have been recorded.

Resources and settlement. The diversity that characterizes Appalachian plant and animal life also exists in the system's mineral resources. As the immense stands of timber throughout the Appalachians brought into existence important lumber and pulp industries, so rich coal beds, veins of iron ore, salt springs and licks, and deposits of granite and marble created important American industries in the region. Each of these has been attended by its own peculiar problems, especially in the field of conservation. There is, for instance, pollution of Appalachian waterways by pulp and chemical industries, and the devastation of land and human resources brought on by certain coal-mine operations.

The long blue wall of the Appalachians thrust up a mighty barrier to early westward expansion. Exploration and settlement was further discouraged by the size and complexity of the lateral mountain ranges, the rugged courses of many of the streams and rivers, and the ubiquitous dense forest. The central Appalachians, with their more spacious water gaps affording easy passage, attracted the largest number of early settlers. Many of these were Germans and Scotch-Irish who went into the interior of Pennsylvania and subsequently migrated down the Great Appalachian Valley in Virginia and Tennessee. In the New England Appalachians, the narrow notches, often blocked by glacial debris, and the steepness of the mountains discouraged early settlement—as did the massive ruggedness of the successive ranges in the southern Appalachians—so that each of these regions remained a wilderness long after Daniel Boone had forged his pioneer route through the Cumberland Gap. During the French and Indian War (1756–1763), trails to the Ohio country and Great Lakes led through central Appalachia, and these, with their scattered lonely forts of the interior, became brutal battlegrounds of border warfare. Indian allies of the British waged bloody skirmishes against Appalachian settlements during the American Revolution, but the Appalachian Mountains confined most of that struggle for independence to the eastern seaboard.

Obstacles to settlement

Despite the early arrival of the lumber industry and opening of the coal mines, some areas of Appalachia remained isolated until early in the 20th century, notably those mountain areas of the southern region where rough terrain hindered building of roads. As a consequence,

Southern Appalachian highlanders developed a distinctive culture characterized by handicrafts, ballads, folklore, and mores that reflect both the massive problems and the rich potential of the region.

BIBLIOGRAPHY. There are numerous works on specific subjects and scope confined to single portions of the Appalachian mountains, but there is no definitive overall view. Information may be found in *Maps and Geologic Studies of the U.S. Geological Survey; T.V.A. Topographical Quadrangle Maps; T.V.A. Annual Reports;* N.M. FENNEMAN, *Physiography of the Eastern United States* (1938); and in the *Message from the President of the U.S. Transmitting a Report of the Secretary of Agriculture in Relation to the Forests, Rivers and Mountains of the Southern Appalachian Region* (1902), a thorough survey of the natural resources of southern Appalachia and some of the conservation problems emerging early in this century. M. BROOKS, *The Appalachians* (1965), while in no sense definitive, is the most accurate and comprehensive recent book on the Appalachian Mountains, with good coverage of the physical environment. Other works include: T.L. CONNELLY, *Discovering the Appalachians* (1968), a highly readable account of each part of the Appalachians in their natural setting, with simple helpful maps; M. FROME, *Strangers in High Places* (1966), a long and detailed account of the natural and human history of the Great Smoky Mountains; *Biographical Memoirs of the National Academy of Science*, vol. 2, *Memoir of Arnold Guyot, 1807–1884* (1886), Smithsonian Publication No. 707, a memoir written by one of the explorers of the Southern Appalachians, whose work is commemorated by a peak named for him in the Great Smokies; A.W. ECKERT, *The Frontiersmen* (1967), a compelling, documented narrative history of the unique men who pushed America's frontiers across the Appalachians to the Ohio country; A. and M. SUTTON, *The Appalachian Trail* (1967), an account of the history and natural surroundings of major portions of the Trail, with suggestions on hiking and camping and a bibliography; and H.E. JOLLEY, *The Blue Ridge Parkway* (1969), a scholarly study of the establishment of the Blue Ridge Parkway, with photographs and a Log of the Parkway in the Appendix.

(W.D.)

Apprenticeship and Employee Training

Apprenticeship means training in an art, trade, or craft, under a legal agreement defining the relationship between master and learner and the duration and conditions of their relationship. As a form of professional initiation, it most often represents an intermediate phase in the evolution from traditional to industrial society. In the traditional society social position and occupation were determined by birth and inheritance; craftsmanship remained within particular families, as in India, where occupation was determined by caste membership and occupational specialization arose from hereditary caste callings.

From the earliest times, in Egypt and Babylon, training in craft skills was organized to maintain craftsmen in adequate numbers. The laws of Hammurabi of Babylon, which date from the 18th century BC, required artisans to teach their crafts to the young. In some ancient societies, Rome for example, many craftsmen were slaves, but in the later years of the Roman Empire, craftsmen began to organize themselves into *collegia* to maintain the standards of their trades.

By the 13th century a similar practice reappeared in western Europe with the emergence of the craft guilds, which supervised quality and methods of production and regulated conditions of employment for each occupational group in a town. The guilds were controlled by the master craftsmen, and the recruit entered after a period of training as an apprentice, commonly lasting seven years. It was a system suited to domestic industry. The master operated on his own premises, where his assistants resided as well as worked. An artificial family relationship was thus created, with the articles of apprenticeship taking the place of kinship.

Impact of Industrial Revolution

The Industrial Revolution, with its incipient mass production, destroyed the personal relationship. Early machines were easy to mind, there was plenty of labour available to mind them, and the main inducement to productivity was fear of unemployment. Training was a simple matter; the worker achieved proficiency for economic survival. Little attention was paid to the training of employees in the 19th century, though the crafts themselves survived, maintained their apprenticeships, and even extended their scope. Nor were managers consciously trained. Management was empirical; the margin of profit was so great that there was no need for refined organization. Experience, learning "the hard way," was the essential qualification for supervision and management, and so there arose the tradition of the self-developed industrial tycoon who began life in humble circumstances.

Social and industrial change gradually altered attitudes to work and training. The worker asserted himself against the rigid control of the entrepreneurs and managers. New techniques in machine design and in electronics and synthetics require technological training and technical education. Modern industries and communications demand sophisticated management. Engineering and scientific graduates now enter industry in large numbers. As technology has advanced, management has become more complex and a variety of schemes for management training have been developed.

In recent times governments have introduced economic policies aimed at full employment, making labour a relatively scarce commodity. Employers must actively recruit staff and must encourage new entrants to develop skills by effective training. Labour unions have become vigorous supporters of training programs. Professional societies also have taken an increasing interest in the standards required of their members.

Government itself is involved in new activities, such as state insurance, public welfare, communications, public utilities, and in some countries industrial production. This has led to programs of public-service training, both in established industrial societies and, with aid from the United Nations and other technical assistance programs, in the new countries seeking to improve living standards through technological development.

HISTORICAL BACKGROUND

The medieval guild made its members responsible for ensuring the skill and capability of each newcomer, initiating him into the customs of society, and bringing him up in ways of godliness. The tools and methods of work were prescribed by the guild, and the master was not allowed to take more apprentices than he could effectively train. The apprentice was accepted as a master craftsman after he had satisfactorily completed the full term of training and had demonstrated his ability by examination or by preparing a *masterpiece*. Even then, he did not necessarily become a master immediately. There might be an intermediate period during which he worked for days' wages as a journeyman (derived from the French *journée*).

Traditionally the apprentice slept in the shop, turned out from time to time for a political demonstration, and, if he was lucky, married the master's daughter and succeeded to the business. It was a system primarily suited to small working units, though it was capable of large-scale operations. The cathedrals of Europe, for example, were built by master masons whose organizations frequently left some characteristic mark, like the lions of the Comacine Guild in Italy, which date from the 11th century.

The exclusivity of the guilds. As time went on, governments had to contend with the exclusiveness of the guilds, who, controlled by their wealthiest members and frequently abusing their privileges, monopolized their trades in each town. Outsiders were prevented from entering by heavy fees, and apprenticeship was restricted to sons of guild members or of their wealthy neighbours.

The English government tried to define conditions of apprenticeship by the Statute of Artificers of 1563, which attempted to arrest these undesirable practices and to ensure adequate labour for industrial and agricultural production. Under its terms, no one might exercise a craft unless he had been apprenticed, and the nature of employment was determined by the boy's parentage. The act also established a ratio between the number of journey-

men and the number of apprentices; each master with three apprentices was compelled to keep at least one journeyman. Justices of the peace were required to administer the statute, which also gave them authority to assess wages.

Apprenticeship declined as the Industrial Revolution progressed during the 18th century, but attempts were yet made to maintain traditional standards. In 1756 the city guilds in Glasgow refused to recognize the inventor James Watt because he had not served his full term as an apprentice. In France, the legislation under which the craft *corporations* administered their trades was abolished by the National Assembly in 1791, and in England the Statute of Artificers was repealed in 1814.

The notion of individual training extended beyond the craft guilds in the Middle Ages. The university accepted the same principle with its *master's* degree, and the religious orders insisted on the newcomer passing through a novitiate. In medicine, the guild system applied to the surgeon, who also performed the function of barber and was regarded as a craftsman, with less prestige than the physician. The lawyer served an apprenticeship by working in close association with a master of the profession. This has persisted in England, with the articles of clerkship that bind the pupil to a solicitor or barrister until he is qualified to practice for himself, but in the United States and in most European countries, the law degree is now the main avenue to the profession.

Sons of the nobility needed training for their duties in the feudal community, and they usually were attached as pages or squires to the court of some prominent nobleman, thus serving an apprenticeship in the code of chivalry, by which they gained admission to knighthood and full membership of the aristocratic circle. The squire had to learn the use of weapons and to ride *the great horse*, the heavy charger, necessary for fighting in armour. He also had to understand the innumerable intricacies of the world of chivalry, with its orders and its rank and precedence, and he had to behave with elegance, courtesy, and other marks of good breeding.

The medieval universities were closely associated with the teaching of theology, and the student at Bologna, Paris, or Oxford frequently took holy orders; but this was the main avenue to what would now be a senior post in government service. Ministers of the European rulers frequently were ecclesiastics. They had the education necessary to deal with the intricacies of statecraft and the subtleties of diplomatic documents, and they could write and speak Latin, then the language of diplomacy.

From the 16th century the lay nobility and the new middle class assumed much of the routine work in government. In France lawyers became prominent among the king's intimate advisers, and in England the close association between the universities and government began in the days of Elizabeth I. Training for the minor posts in government service was conspicuously lacking. Appointments were made by patronage with no test of ability, until the entrance examinations and other reforms of the 19th century.

The coming of the Industrial Revolution. The Industrial Revolution altered attitudes toward training. Machines created a need for unskilled labour for which formal training was unnecessary, and employees who showed aptitude moved on to semiskilled jobs. Yet craftsmen remained an essential feature of industrial society, both to build machines and to keep them running and to maintain trades still dependent on individual skill; and so apprenticeship continued, and, indeed, a new development strengthened its position. The trade unions became directly interested in maintaining standards and in controlling recruitment to protect skilled workers. As the factory system replaced domestic industry, a new kind of apprenticeship developed in which the employer was the factory owner, and, after his training, the apprentice became a factory worker.

In France there was an early reaction against free admission to all trades; legislation concerning apprenticeship was reintroduced in 1803, to be greatly strengthened in 1851. In the reconstruction of Prussia, which began during the Napoleonic Wars, apprenticeship became an important feature of industrial training. In England apprenticeship was maintained by the craft industries and even extended to analogous fields. In education, for example, there were various schemes for pupil teachers, which were in effect a kind of apprenticeship, and there was a comparable system of pupil training for young farmers.

In the United States apprenticeship existed in colonial times. Indentured apprentices were brought there from England in the 17th century. Benjamin Franklin served as apprentice to his brother in the printing trade. But apprenticeship was less important than in Europe because of the high proportion of skilled workers among immigrants. Legal regulation, at both the federal and the state level, belongs to the present century.

Technical education

The development of large-scale machine production increased the demand for semiskilled workers whose skills were confined to a particular specialty, created by an ever-increasing division of labour. The more ambitious among them sought to increase their effectiveness and potential for advancement by voluntary study. To meet this need, mechanics' institutes were established, such as the one founded in London in 1823 by George Birkbeck, which still exists in modernized form as Birkbeck College, and Cooper Union for the Advancement of Science and Art in New York City, established in 1859. In France, technical education on a national scale dates from 1880.

Preparation for clerical work has been more closely associated with the regular educational structure. In the late 19th century evening schools were established to prepare students for employment as clerks, cashiers, and bookkeepers, but apart from the secretarial schools made necessary by the invention of the typewriter and the widespread use of shorthand, training for office work, after recruitment, has been a relatively recent development.

The new technology also demanded more theoretical understanding and skills, for which the universities made no provision, and the earliest attempt to meet this requirement was inspired by the French Revolution, although it is commonly associated with Napoleon. This was the Polytechnic (*École Polytechnique*) established in 1794, which soon came to be a great scientific and technological institution. It was the first of the great professional schools (*grandes écoles*), peculiar to France. Each profession has its school; for example, the School of Mines, the School of Roads and Bridges, the National School of Agriculture and, among more recent creations, the School of Aeronautics and the National School of Administration. Admission to each is by a nationwide entrance examination, which requires special preparation.

The armed services also have long felt the need for professional training for officers. The Royal Military Academy was set up in 1741 at Woolwich to train officers for the artillery, the most technical arm then in use. The French Military School at Saint-Cyr dates from the Napoleonic period, and the U.S. Military Academy at West Point from 1802. The U.S. Naval Academy at Annapolis, founded in 1845, is similar in character to West Point; and in France the École Navale at Brest complements Saint-Cyr. From the 17th century the British Navy trained its officers at sea as midshipmen on active duty, in effect a form of apprenticeship, although many future officers also served a preliminary period of training at a naval college.

CURRENT STATUS OF APPRENTICESHIP

Apprenticeship in Europe has maintained remarkable continuity from medieval times, but early in the 20th century assembly-line methods created repetitive work of an unskilled or semiskilled character, which made the long period of apprenticeship unattractive. This led many countries to attempt planned organization of the labour market, designed to help juveniles and unskilled and semiskilled workers to find suitable employment.

The need for apprenticeship has remained in craft industries, in spite of the fact that mechanization has progressively increased the number of jobs for which little formal instruction is necessary. After World War I a new pattern of recruitment emerged. Apprenticeship on traditional lines was maintained for skilled craftsmanship; for less skilled work learnership became a common practice —the newcomer was given facilities to learn by working with others. Some industries introduced a system of upgrading; labourers and unskilled workers were allowed to undertake skilled work after having served as assistants to other skilled workers. These training schemes were supplemented by pupil apprenticeship, whereby the recruit learned working skills with the intention of qualifying for a responsible position in the industry, and student apprenticeship, offering entrants with a university or technical education working experience to qualify for staff appointments.

Apprenticeships under the trade unions

Trade unions emerged, like the medieval guilds, as guardians of the principle of restricting entry to skilled trades by means of rigorous apprenticeship, but with the opportunity for semiskilled labour to move gradually into skilled jobs, the craft unions themselves began to change. From the 1920s they began to accept for membership workers who had not entered as apprentices. The ratio of apprentices to journeymen ceased then to be an issue.

Apprentice selection rarely was based on exacting criteria. If a choice had to be made, preference was given to sons of employees. Training consisted of little more than working with a craftsman. It was haphazard and frequently not well done. The period of apprenticeship was standardized, normally at five years in western Europe, with entry at age 16. Wages were low. It has been estimated that in Britain during the 1920s, pay for the last year of apprenticeship was about 40–50% of the journeyman's rate in many crafts, with a further short period as an "improver" at 76% of full pay. Parents frequently were willing to support their sons for the long-term benefits and attrition was low, from 1.3% in printing to 10% in shipbuilding; but at that time apprentices represented only 20% of workers under the age of 21.

Apprenticeship schemes in modern Europe. After World War II the situation changed, with considerable variations between different countries. Perhaps the most radical changes took place in West Germany, which had the advantage of a practically new start. There, a distinction is made between skilled trades needing apprentices, semiskilled trades needing trainees, and handicrafts employing artisans. Trades are grouped under local chambers of industry, handicrafts under chambers of handicrafts. The apprentice's contract is registered with the appropriate chamber. He keeps a workbook, which its officials inspect from time to time, and there are regular tests at monthly or three-monthly intervals. The school-leaving age is 14, but part-time attendance at a vocational school, included in the hours of employment, is compulsory until 18. The apprentice can apply to take his school-leaving examination when he is ready, sometimes resulting in a training period of as little as three years.

Apprenticeship in The Netherlands is controlled by foundations sponsored jointly by employers and trade unions. Each foundation is responsible for craft training in a group of related trades in a particular industry, such as the metal industries (the Bemetel Foundation); the chemical industry; the electrical industry; gas, water, and sanitary fitters; automobile servicing; and the hotel and restaurant business. Training lasts from two to three years according to trade. One large corporation, the Royal Dutch Blast Furnaces and Steelworks at Ijmuiden, has a training school of its own with a capacity of 400 apprentices.

In France, vocational training is supervised by the Directorate of Technical and Vocational Education of the Ministry of Education. There are 24 national, professional, consultative commissions on which employers, government, and trade unions are represented, and since 1930 the government has developed technical colleges, which train about one-third of all skilled workers, about one-third of them girls. There are also a number of works schools, the most famous being the Renault School. Apprenticeship may be by contract with a private employer, by attendance at school beyond the normal leaving age of 16 (called the complimentary course), or by apprenticeship to an artisan trade. There is always a passing-out examination; those at the colleges and schools supervised by the Directorate, those for apprentices by the local chambers of crafts.

In Italy, where apprenticeship is not well developed, the works schools are of great importance, notably the Giovanni Agnelli School of Fiat in Turin and those of the Pirelli, Olivetti, and Alfa Romeo corporations. There is a system of apprenticeship for tourist-industry workers, supported by the government. A particularly interesting training scheme belongs to the Salesian Order initiated by Giovanni Bosco, a priest in Turin who in 1841 started evening training for boys from the poorer parts of the city to equip them for better jobs. His mother established a dormitory for homeless pupils and thus began the first Salesian Home, so called because Bosco, himself later canonized, dedicated the work to St. Francis de Sales. Thus began a widespread movement, developed by the order and approved by the Pope in 1874. When Bosco died in 1888, there were 250 Salesian homes with 130,000 pupils in all parts of the world. The mother home in Turin trained the best of its pupils as instructors. The homes later became technical institutes, perhaps the most famous being the Institute of San Giorgio in Venice, which specializes in printing and furniture making.

In Switzerland high-quality manufacture for exports and excellence in the hotel industry are national necessities to maintain adequate balance of trade, so that although education and training are primarily cantonal matters, traditional attitudes have given way to the need for uniform national standards, prescribed by the Federal Vocational Training Act of 1930. Apprentices must be registered, and craft training, lasting from one to four years, must be accompanied by study one day a week in a vocational school. Some apprentices, about 5 percent of the total, attend cantonal vocational schools full-time. After passing a final examination, the apprentice is certified as a "skilled worker." An adult worker also may take the skilled worker examination if he has worked in the trade for twice the period of apprenticeship.

Great Britain and the Commonwealth. In Britain, the Employment and Training Act of 1948 created a Central Youth Employment Executive and led to proposals for a National Joint Apprenticeship and Training Council to be set up in each industry. The printing trade, usually conservative in these matters, introduced selection by objective testing of suitability and aptitude. The motor-vehicle-repair industry developed an apprenticeship scheme with a final passing-out test, an examination for a national craftsman's certificate. The Post Office Engineering Department, which never accepted the traditional pattern of apprenticeship in the training of its telephone and telegraph technicians, has a three-year course for recruits, accepted between the ages of 16 and 20, with the first two years spent in general training and the third year, after a selective interview, in one of several fields of specialization. But the most significant break with the past is the module system in the engineering industry. This provides a year's training in a wide selection of skills, followed by selected training in specialized skills. There are performance tests and appropriate further education. Apprentices are allowed to qualify as quickly as possible; the traditional five-year period has disappeared. Because some small firms are unwilling to take apprentices and to afford them part-time day release, it is the large corporations who have in fact made the best use of apprenticeship.

A scheme of apprenticeship of historical significance was introduced in 1843 in the Dockyard Schools of the Royal Navy and lasted until recently. The primary intention was to train candidates for the School of Naval Architecture. From the beginning, apprentices were given

release one day a week for studies in general and technical subjects and were required to attend evening classes, a pattern of technical education not generally accepted until a century later. The seven-year training reached a final standard approaching that of an engineering degree. There were exacting annual examinations, but the School of Naval Architecture did not absorb all the successful students; many went on to distinguished careers in the technical and scientific services of government.

Apprenticeship has extended to other Commonwealth countries. In Australia it is regulated by laws that vary from state to state. In New South Wales, for example, conditions are fixed by the Apprenticeship Council of New South Wales; in Victoria they are prescribed by regulations on the recommendation of the Apprenticeship Commission, following advice by the Trade Committees and Advisory Committees or by the determinations made by Wages Boards. The National Australian Apprenticeship Advisory Committee, established in 1957, consists of representatives of the State Departments of Labour and Technical Education and Apprenticeship Authorities.

The employment of apprentices in New Zealand is regulated either by special order of the Industrial Arbitration Court or by apprenticeship committees appointed by the court.

The United States. In the United States, Wisconsin was the first state to pass legislation (1915) for organized apprenticeship. By the Smith-Hughes Act of 1917, the federal government allocates funds to the states for vocational education in agriculture, trade, and industry, applied primarily to vocational education through the school system, but there is close collaboration between the schools, labour, and management.

Conditions of apprenticeship are more flexible than in Europe. The lowest age for entry is 16, but in many trades the requirement of a high school diploma makes the effective entry age 18. The upper limit normally is 24 years of age and in some trades it is higher. The period varies from two to five years but is usually three or four, and much of the training is at technical and vocational schools.

The Bureau of Apprenticeship was set up in 1937 in the Department of Labor, and the National Apprenticeship and Training Council promotes schemes for individual industries. Apprenticeship normally is phased in periods of 1,000 credit hours, each equivalent to six months and including 144 hours as the minimum of related classroom instruction. The trainee advances from phase to phase after passing qualifying examinations. Pay, varying from 60 to 90 percent of the journeyman's rate, increases at each phase. Apprentices can be indentured to the Local Apprenticeship Committee of the craft, and thus, in effect, to the industry not to an employer. This has been particularly successful in the building industry.

The Fitzgerald Act of 1937 authorizes the secretary of labor to safeguard the welfare of apprentices, to encourage employers to develop programs, and to cooperate with state agencies in formulating and promoting standards of apprenticeship, but variations in practice exist among the states. In New York, for example, contracts of apprenticeship must be registered with the State Department of Labor, and the yearly number of registrations is about 15,000.

Japan. In Japan, apprenticeship and employee training often have a personal orientation rarely found in other industrial nations. The Japanese concept of apprenticeship differs from elsewhere because of a difference in the relationship between employer and employee. Employment with a Japanese firm tends to become a lifetime relationship. Social obligations require the employer to provide work for his employee and require the employee to continue with the same employer until death or retirement in any capacity to which he may be assigned. The apprenticeship is thus oriented toward employment with a particular company, in which the employee is likely to spend his whole working life, expecting permanent employment even if his specific assignment may vary from time to time. The close relationship between the apprenticeship and a specific trade, common elsewhere, is thus missing in Japan. Since passage of the Vocational Training Law of 1958, however, publicly subsidized programs have been set up by private companies or associations of private companies, and a great many general vocational training centres have been established and operated by the 46 prefectural governments. These cater to junior high school graduates, workers wishing or obliged to change their occupations (*e.g.*, those workers displaced from coal mining by reorganization), and the physically handicapped.

EMPLOYEE TRAINING IN INDUSTRY

During and after World War II, in-service training by the employer became a much more common practice. The rapid changeover in industry from peace to war led to training schemes for semiskilled workers, for workers transferred to new jobs, and for women brought into industry on a wartime basis. At the end of hostilities the armed forces were reduced in numbers, and the manufacture of their equipment and supplies declined rapidly. Many people thus changed jobs, and many changed from war to peace production in the same trades.

All of this brought about a new approach to training. Great emphasis is now placed on a good start through initial job training, supplemented by orientation sessions or by attractively produced printed material describing the nature and objectives of the employment and the conditions of work. Since changes are frequent with technological advances, refresher training has become common in clerical as well as in industrial work.

A whole range of training has been developed for the training of supervisors. This began during the war, with group training in the basic elements of supervision, under a plan first developed in the United States under the title of *Training Within Industry*. It was designed for mass application under wartime pressures, and while the principles on which it was based were sound, more sophisticated techniques have greatly developed the training of supervisors under more favourable conditions. Another important feature of any successful training organization is a well-designed scheme for training the instructors.

For the more technical skills, it is quite common in the United States for the large employer to make arrangements with a university to set up special courses; in Great Britain it is more usual to encourage employees to attend regular class facilities to obtain technical certificates. Sometimes this is achieved by *sandwich* training, periods on the job alternating with periods at a technical institute. Many employers encourage further education by paying tuition fees or by allowing free time to attend classes. Some very large corporations have developed their own systems of technical classes, supplementary to direct job training.

New methods and new needs

This widespread interest in training has led to considerable innovation in method. Formal lectures have given way to group discussion. The case-study method has become popular; a problem situation is presented in considerable detail and trainees are asked to make suggestions for its solution. Another new technique is role playing. Members of the training staff create a situation by playacting, and the trainees either comment on what is taking place or participate in the attempt to find a solution, or they perform functions or services in conditions that simulate their working environment. Attention also has been given to audiovisual aids. Sensitivity training recently has been introduced to help individuals to study their own behaviour and reactions to one another by means of group discussion in which there is frank analysis of interrelationships between members of the group.

New industries have created new needs. The instruction of airline stewardesses has become a highly developed operation for the major airlines. Television and soundbroadcasting organizations have introduced training schemes to improve the quality of their program services. At United Nations headquarters a training scheme has

been developed for the guides, who conduct visitors round the building, including daily briefing on the international events on which they may be questioned.

Initiative in training lies with the organization rather than with industry, and the large corporation tends to develop a variety of training projects and adequate administration of training. In a large organization, the individual employee needs to be introduced to his task and to identify himself with it, so as not to be lost in its complexity; and he needs to have subsequent periods of training to keep abreast with developments. In small- and medium-sized concerns some managements are interested in training; some tend to regard it as a luxury. It is impossible to determine which industries give the most attention to formal training; it depends on the initiative of the management rather than on the nature of the work.

Corporation schools in the United States date from the 1890s; the National Association of Corporate Training was created in 1919. Training schemes also have been supported by professional groups, such as the International City Managers' Association, the Public Personnel Association, and the Council of State Governments.

Government initiative has been necessary in the United States in certain fields of training. Title VIII of the National Defense Education Act (1958) is designed to increase the available supply of technicians, and the Manpower Development Training Act of 1962 makes the secretary of labor responsible for programs of occupational training for employed workers.

The Industrial Training Act, which came into force in Britain in 1964, is perhaps the most elaborately structured government initiative in training so far undertaken. It provides for the establishment of an Industrial Training Board for each industry to make specific recommendations concerning the form and content of training courses and the standards to be set, and to recommend appropriate further education. By 1960, 26 boards had been set up, covering some 15,000,000 employees. They may impose a financial levy on firms in the industry and use the money to make grants in support of training schemes. There is a Central Council on Industrial Training and Further Education to give direction to the scheme.

Management training

With the rapid advances in technology and the growing complexity of business and industry, management training has become accepted as a necessity in both the public and private sectors. In the United States, graduate business education and senior executive training schemes, such as the advanced management program for senior executives at the Harvard Graduate School of Business Administration, were already well established in the immediate postwar period. In Great Britain the Administrative Staff College was set up at Henley-on-Thames in 1944 to offer short courses in problems of advanced management. It employs a novel technique of training by group initiative, drawing its inspiration from the professional experience of the participants. It has been copied successfully at Mount Eliza, Australia; Hyderabad, India; Lahore, Pakistan; and Baguio, the Philippines. Individual corporations have established institutions of their own for the development of general management techniques at senior levels, and two international corporations, Unilever and Royal Dutch-Shell, have introduced management training for groups of senior staff drawn from a variety of countries.

TRAINING FOR GOVERNMENT SERVICE

The problems of management training in the private sector and among public enterprises operating on commercial lines do not differ greatly between countries with free-market economies. The training of government staff varies considerably in terms of different national traditions of administration and education. An important recent development has been the establishment, in 1945, of the National School of Administration in Paris. It follows the tradition of the great French professional schools and is also the recruitment agency for the French administrative and diplomatic services. Its entrance examination is rigorous and its standards are high. The period of instruction includes 12 months at the office of a Prefect, followed by 17 months in Paris.

In Britain full-scale plans for civil service training began with the *Report on the Training of Civil Servants,* submitted to Parliament in 1944, and a special committee on training was set up by the National Whitley Council. Training suffered a setback through economic stringency in 1952, but a review of the situation by the Civil Service National Whitley Council in 1964 specifically criticized the retrenchment of 1952 as false economy and urged better use of available facilities. A Centre for Administrative Studies was established in London in 1964, and as a result of the Report of the Fulton Committee on the Civil Service, this has developed into the Civil Service College with three centres: one nonresidential, in London and two residential, one close to London and the other in Edinburgh, to provide major training in administration and management and short courses in general management and vocational subjects.

In the United States, government agencies tended to follow their own initiatives for training until Congress passed the Government Employees Training Act of 1958, designed to promote efficiency, economy, and high standards of service by departmental initiative, supported by interdepartmental efforts under the guidance of the Civil Service Commission. There is no general administrative staff, as in France or Britain, and management training for senior civil servants is therefore aimed at giving senior staff—who have spent their earlier careers in more specialized areas of administration—opportunities to study the effective coordination of government. This need was first met by a series of conferences in Williamsburg, Virginia, arranged for the federal government by the Brookings Institution, over a period of eight years from 1957 to 1965. In 1969 the government itself took a direct initiative with the establishment of the Federal Executive Institute at Charlottesville, Virginia. In 1970 New York State set up an Institute for Government Executives, the first training institution for senior staff in a state administration.

In Ireland the initiative in public-service training came largely from the staff associations. The Association of Higher Civil Servants formed a discussion group in 1952, which inspired the publication of a Journal of Administration in 1954. The Institute of Public Administration was established in 1957, with strong support from the Association of County Managers. It began with training in local government and went on to training for the employees of the state corporations. A School of Public Administration, associated with the Institute and supported by government grants, was set up in 1964.

In the U.S.S.R., where industrial and commercial operations are under government control, higher education is related directly to professional development. The student undertakes specialized studies at a university or professional institute appropriate to the profession or specialization which he intends to make a career. Management itself is approached on an institutional basis through the Laboratory for Industrial Management Problems of Moscow State University.

In Yugoslavia the drastic governmental decentralization of 1953 set a unique problem of retraining to meet new working conditions, and in 1956 seven administrative high schools were created, offering two-year courses for officials in the middle ranks in the republics and municipalities and in the federal social and economic institutions. An advanced school of administration was set up in Zagreb, with a three-year residential course. In 1962, 21 administrative secondary schools were established, each with a four-year curriculum, thus providing training for the public service by specialized education at the high school level.

When India became independent in 1948, its civil service already had a long tradition of professional training, shared by Indians as well as Europeans, from the days of the East India Company. A national school at Delhi offered a one-year course for new entrants to the Indian

Administrative Service; in 1959 the school was transferred to Mussoorie as part of a National Academy of Administration, which provides associated schemes of training for the Administrative Service and the specialized federal services.

Problems of the developing countries

The newly developed countries have unique problems of training, their economic advance depending largely on the introduction of new and unfamiliar techniques. Training organization is needed in basic skills, both industrial and clerical, and for the provision of adequate quantities of trained technicians, supervisors, and competent managers. To achieve planned progress these nations need skilled administrators in large numbers, and above all they require educators and instructors. In some, the primary and higher educational structure is inadequate for current needs, no vocational training is built into the school system, and little or no science and technology are offered in the universities.

Inevitably they must send some of their key personnel for training abroad and call in foreign experts under one or another of the technical-assistance programs. But foreign experts cannot train workers for a whole industry or instruct the staff necessary to organize an entire national development program. They must concentrate first on building up local groups of indigenous experts, with aptitude for training others, organized if possible on an institutional basis.

The United Nations and its specialized agencies contribute to the development training schemes in these countries, emphasizing development of institutional training. The Regional Economic Commissions have sponsored the establishment of regional Institutes of Economic Development in Africa, Latin America, Asia, and the Far East and have encouraged schemes for training in statistics to provide a sound basis for development planning. A particularly successful project is the Institute of Public Administration in Costa Rica, whose task is to train the staff necessary to administer the coordinated regional economic development of Central America. Eleven former French-African territories have established national schools of administration, noticeably influenced by the school in Paris.

In Turkey a Civil Service School was established in Istanbul in 1859. It was moved to Ankara by Atatürk in 1936, ultimately to become the Faculty of Political Science at the University. In 1952, with United Nations assistance, an Institute of Public Administration for Turkey and the Middle East was developed in close association with the university, but this institution was subsequently moved to the office of the prime minister. The Philippines, with United States aid, developed within the university an institutional project that provided direct vocational training for civil servants for a limited period; this later became a graduate College of Public Administration.

In support of such projects, of which these are but a few examples, the United Nations has sponsored a series of international studies of training problems. An outstanding example was the interregional seminar on the development of senior administrators in the public service of developing countries, held in Geneva in 1968.

Measuring the success of training programs

Perhaps the most intractable problem associated with training is its evaluation. Its actual cost may be calculated in terms of the expenses of its administration and the salary costs of both the training staff and of the trainees while they are on nonproductive work. But the quality and ultimate success can be determined only by a value judgment on whether the effort seems justified as conducive to greater overall efficiency and to more successful operations. Craft skills and routine occupational skills can be measured by tests based on agreed standards, but occupations measurable in this way account only for a limited range of training activities. In the office, typing and shorthand work can be tested for speed and accuracy, but a great deal of clerical work cannot be analyzed statistically. Supervision, management, and administrative tasks depend on personal capability as much as they do on knowledge and experience. Knowledge can be imparted and experience acquired; the guided development of personality is more difficult. How far the training opportunities offered to a senior executive during his career have helped his professional development cannot be mathematically assessed.

Future prospects for training. The rapid contemporary advance of technological change makes training a necessity in almost all walks of life. At the operating level in industry and in public utilities, new techniques, new methods, new tools, new synthetics, new sources of power, and increased uses of automation have brought extensive changes in the past two decades, and the rate of change tends to increase as time goes on. Comparable changes are taking place in the office with the extended use of computers, which provide for the storing and recall of information in amounts unknown 20 years ago. The cost and maintenance expenses of this complex mechanization of industry, commerce, and the public services demand more critical analysis of development plans and proposals. In the modern world the training of managers—and particularly of men and women on the threshold of management—will be especially important because they must be equipped to meet such change.

BIBLIOGRAPHY. G. WILLIAMS, *Recruitment to Skilled Trades* (1957), and *Apprenticeship in Europe: The Lesson for Britain* (1963), offer a broad study of apprenticeship, including a survey of the position in western Europe following World War II. F.R. MARSHALL and V.M. BRIGGS, JR., *The Negro and Apprenticeship* (1967), deals with a contemporary problem of apprenticeship in the U.S. There is no fully comprehensive survey of staff training. G.D. HALSEY, *Training Employees* (1949); and W. MCGEHEE and P.W. THAYER, *Training in Business and Industry* (1961), are studies made in the U.S.; and F.J. TICKNER, *Modern Staff Training: A Survey of Training Needs and Methods of Today* (1952), a study made in Britain. W.Y. ELLIOT (ed.), *Education and Training in the Developing Countries: The Role of U.S. Foreign Aid* (1966), deals with training problems of the "new nations." A wide range of studies, too numerous to be listed, has been made of training for specific vocations and professions. N.F. HALL, *The Making of Higher Executives: The Modern Challenges* (1958); F.C. PEARSON *et al.*, *The Education of American Businessmen: A Study of University-College Programs in Business Administration* (1959); and S.B. SWEENEY and T.J. DAVY (eds.), *Education for Administrative Careers in Government Service* (1958), form a substantial survey of management training. The UNITED NATIONS, *Report of the Interregional Seminar on the Development of Senior Administrators in the Public Service of Developing Countries*, 2 vol. (1969), is a unique world survey that discusses the experiences of developed countries as well as the needs of developing countries. Another useful UN survey is *Local Government Training* (1968). See also R. WILKINSON (ed.), *Governing Elites: Studies in Training and Selection* (1969).

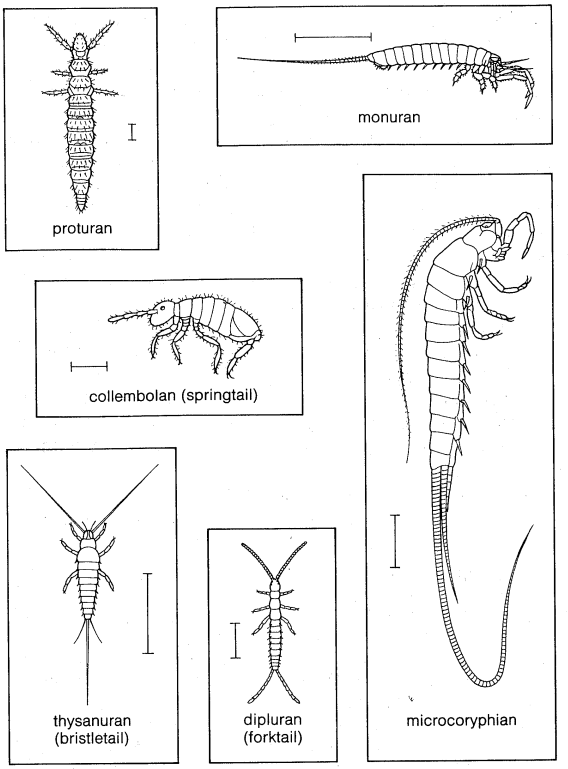
(F.J.T.)

Apterygote

Apterygotes have long been defined as primitive wingless insects of the subclass Apterygota (class Insecta), distinct from the subclass Pterygota or winged insects. Used in this sense, the term apterygotes commonly includes four groups of primitive insects: proturans, collembolans, diplurans, and thysanurans. The taxonomic status of these groups, however, remains unsettled.

A typical apterygote is wingless and has six legs. Although the presence of six legs was once an important feature in the identification of true insects, three groups commonly known as apterygotes (*i.e.*, proturans, collembolans, diplurans) are now considered by some entomologists to be offshoots from the main insectan stem of evolution and have been given independent taxonomic status as classes equivalent to the class Insecta. The term apterygotes, therefore, has been redefined to include only those groups thought to be ancestors of pterygotes *i.e.*, the thysanurans (*e.g.*, silverfish, firebrats, and bristletails), together with the Microcoryphia (a group closely related to the Thysanura) and the extinct Monura.

For completeness of discussion, and because of the similarities of these primitive hexapods, the proturans, collembolans, and diplurans, as well as the thysanurans. microcoryphians, and monurans are included in this article.



Types of hexapods. The line scales indicate an average length (in millimetres).

From (monuran) A.G. Sharov, *Basic Arthropodan Stock with Special References to Insects* (1966), Pergamon Press; (microcoryphian) R.J. Tillyard, "Primitive Wingless Insects," *New Zealand Journal of Science and Technology* 7(4) (1924); (others) *Invertebrate Identification Manual* by R.A. Pimental © 1967 by Litton Educational Publishing, Inc., reprinted by permission of Van Nostrand Reinhold Company

General features. Protura are minute (to two millimetres in length), elongated, white, and lack antennae. Distributed throughout the world in soil and leaf litter, they number about 152 species. Collembola are diverse in form, coloration, and habitat. Most species are less than three millimetres in length, but some range to ten millimetres. They are elongated or globular insects with antennae and may have a furcula (ventral abdominal springing organ). Collembolans occur in soil and leaf litter throughout the world including Antarctica. There are over 3,500 known species. Diplura are white or yellowish, blind, elongated with long antennae, and less than 10 millimetres in length although one group attains 50 millimetres. Their two tail filaments (or threadlike structures) can be long and thin, short and thick, or in the form of pincers. Diplurans are widely distributed in soil, leaf litter, and rotting logs. Over 400 species have been described. Thysanura and Microcoryphia are mainly long with three elongated tail filaments. Mostly 5 to 20 millimetres in length when fully grown, these insects are widely distributed in leaf litter; some live in ant and termite nests. Microcoryphians have large compound eyes and styli (*i.e.*, bristlelike processes) on legs and abdomen; thysanurans have small compound eyes and styli on abdomen only. More than 350 species of Thysanura and Microcoryphia are known. The extinct Monura were about 11 millimetres in length. Only two species are known from fossils.

Natural history. *Life cycle.* After hatching, the immature primitive wingless hexapods (called nymphs) change to adults with only slight metamorphosis. The young are similar to adults, changing little from molt to molt until sexual maturity is attained. Molting may continue throughout adult life. The greatest changes occur in the Protura, the only anamorphic hexapod group (*i.e.*, increase in number of body segments occurs at time of molting); the complete number of segments is present only after the third molt. There are at least six stages between molts (instars), and the last is the adult. Little is known about the postembryonic development of Diplura. They feed on living and dead vegetable matter and fungi; one group preys on other small invertebrates.

The life cycles of the Collembola are diverse although generally there are three stages (juvenile, postmaturity growth, senile molts) with no significant growth. Females may lay up to 800 eggs that hatch in 2 to 40 days. Three to 12 juvenile molts occur in intervals ranging from 11 days to a year; up to 50 molts occur in a lifetime (from 4–18 months). Collembola feed on living or decaying plant material, including fungi, algae, and spores; a few feed on carrion or are predatory.

In Thysanura there may be more than 40 instars; the adult one occurs after about 12 molts. Some silverfish live for seven years; they reach sexual maturity in two or three years and molt three to five times each year. In Microcoryphia there are six instars including adults. Both groups feed on decaying or dried vegetable material. Domestic silverfish eat plant and animal remains, paper, and artificial silk.

Reproduction

Reproduction in nonpterygote groups is mainly sexual, but parthenogenesis (reproduction without fertilization) can occur. Males deposit sperm packets, often haphazardly, that are taken up by females and stored until the time for fertilization. In proturans small external genitalia of unusual form surround the gonopore in both sexes; collembolans lack external genitalia, but the gonopore and surrounding area differ; in diplurans external genitalia are absent or vestigial. Thysanurans and microcoryphians have external genitalia similar to those of the pterygotes; however, the aedeagus in males is used to deposit sperm drops, and not as a copulative organ. The deposition and pickup of sperm drops in thysanurans and microcoryphians must take place during each adult stage if young are to be produced since the contents of the female sperm receptacle are lost with each molt.

Ecology. Most primitive hexapods live in soil and leaf litter found near the ground or associated with plants and rotting logs in moist regions. Collembolans are also found in aquatic environments, caves, permanent ice fields and snow, and insect and animal nests. Moisture is an important ecological factor in habitat selection. Most primitive wingless hexapods have a thin external covering and must guard against dehydration.

The condition and nature of soil are important factors in the distribution and nature of collembolan populations. Fungi in soil are especially important as collembolan food. Physical factors known to affect collembolans often do so secondarily; the primary effect is on the soil microflora. Changes in macroflora also affect collembolan populations. Direct cultivation (*e.g.*, plowing) has an immediate harmful effect on collembolan populations but does little long-range damage. Fertilizers increase the numbers of collembolans in soil, and herbicides have no effect. Some insecticides are lethal, while others are not. Increases in soil collembolans following insecticide applications are probably due to lethal effects on predators. Estimates of soil collembolans average about 100,000 per cubic metre. Soil collembolans usually form aggregations in soil. Certain species are more prevalent in a given layer (*e.g.*, leaf litter, the fermentation layer, the humus-forming layer, or deep in the soil). Typically small species and young insects occur in deeper layers, although daily and seasonal vertical movements may occur. Occasional swarming of large numbers of collembolans occurs after unusually wet conditions.

The most important collembolan predators are mites. Others are pseudoscorpions, staphylinid and carabid beetles, an empid fly, and dacytine ants. Occasional collembolan feeders include spiders, fish, frogs, miscellaneous ants, and pulmonate snails. Collembolans are parasitized by gregarines, nematodes, virus, bacteria, fungi, and gordian worms.

Form and function. The bodies of proturans, collembolans, diplurans, thysanurans, and microcoryphians are divided into three segments: head, thorax, and abdomen. The three-segmented, leg-bearing thorax has one pair of legs on each segment. The extinct monurans, considered an offshoot of the Thysanura–Microcoryphia stem, have

no separate abdomen or thorax; three pairs of legs occur on the first three of 14 segments. The monuran body, although similar to present day Thysanura, has one tail filament or telson (the median tail filament in present-day Thysanura–Microcoryphia and winged mayflies). Many proturans and collembolans have entirely cutaneous respiration (breathing through the skin). In some proturans, collembolans, diplurans, and microcoryphians, a breathing system of tracheae and spiracles (external openings) occurs in the trunk, while in thysanurans the tracheal system is like that in pterygote insects.

Head. Proturans have anterior mouthparts (they are prognathous) and a pair of sense organs. Collembola are either prognathous or hypognathous (mouthparts are ventral) and may or may not have a pair of postantennal sense organs and lateral ocelli (simple eyes). The diplurans lack eyes and sense organs, while the thysanurans have simple lateral eyes and are hypognathous. Microcoryphia have both compound eyes and ocelli. Monurans have large compound eyes. Mouthparts are important taxonomically in primitive wingless hexapods. Movable mouthparts are generally recognized as walking limbs that were attached to ancestral body segments that became fused to form the posterior part of the head. The Protura, Collembola, and Diplura are entognathous (*i.e.*, the mouthparts are withdrawn inside the head capsule); Monura, Thysanura, and Microcoryphia are ectognathous (*i.e.*, the mouthparts are external, as in pterygote insects). Mandibles are present in the biting mouthparts of Diplura, Thysanura, Microcoryphia, Monura, and most Collembola. Although proturan mouthparts are modified for sucking, mandibles are retained. However, in collembolans with sucking mouthparts, mandibles may or may not be retained.

Mouthparts

Body. In general the three segments of the thorax are distinct, and each bears one pair of legs. Each leg has a terminal claw. Proturan legs are simple except that the anterior legs are longer, serving as tactile organs; the anterior tarsus has sensory organs. The three pairs of legs are similar to each other in the nonpterygote groups.

The abdomen of Protura undergoes anamorphosis: the first and second instars have 9 segments, the third 10, and the rest 12. Collembola have a maximum of six abdominal segments; fusion occurs between segments in some groups. Diplura, Thysanura, Microcoryphia, and the extinct Monura have 11 abdominal segments. The final abdominal tailpiece is the telson (see ARTHROPODA).

Modified abdominal structures. Primitive hexapods have abdominal structures that represent modified remnants of ancestral walking limbs. Many hexapods have cerci (sensory appendages) on the eleventh abdominal segment; these aid in identification of the telson. The Protura, Collembola, and Monura lack cerci. In Diplura a pair of cerci arise from the small terminal segment. Cerci can be long with numerous segments, short with a central duct and terminal pore, or modified into a pair of pincers for holding prey. They are important taxonomic criteria in separating hexapod groups. The extinct monurans have pairs of abdominal structures, called ventral styli, that are modified limbs. Similar styli are present in Thysanura and Microcoryphia; these hexapods also have one or two pairs of vesicles on some abdominal segments, as do proturans. Collembolans have a single tube containing a pair of vesicles, a single median tenaculum (catchlike process that holds furcula in place), and a single median furcula (springing organ); each organ is a modified pair of ancestral limbs.

Sensory appendages

Locomotion. The primitive wingless hexapods are walking terrestrial invertebrates. Proturans, however, use only the middle and posterior legs for locomotion; the anterior pair of legs are held above the head and used as tactile organs. Collembolans have a furcula. This ventral abdominal springing organ enables them to spring rapidly into the air. Apparently thysanurans and microcoryphians have developed the median tail filament (telson) to assist in supporting the abdomen during movement. Microcoryphians can also spring with considerable agility by using the body and tail filaments.

Evolution and paleontology. Evolution of the hexapods has occurred in conjunction with that of the myriapods. Both apparently had a common origin from early crustaceans. Among the myriapods, the Symphyla bear the closest relationship to hexapods, and both may have arisen from an ancestral symphylan stem. The evolution of Protura, Collembola, Diplura, and Thysanura plus Pterygota has usually been considered linear, with offshoots at successive intervals. The hexapodous state may have arisen several times; if this were the case, Protura, Collembola, and Diplura could have arisen separately. The extinct Monura and the Microcoryphia are offshoots from a thysanuran stem; earlier specialization may have led to the winged insects (pterygotes).

Fossil species

There are few fossil species of the primitive wingless hexapods. One extinct collembolan family (Protentomobryidae) contains a species (*Protentombrya walkeri*) of the Cretaceous (approximately 100,000,000 years) Period of Canada. The oldest fossil collembolan species, *Rhyniella praecursor* (family Neanuridae), is found in the Middle Devonian (approximately 370,000,000 years) sandstone of Scotland and has specialized sucking mouthparts. Other species known from the Baltic amber include one campodeid dipluran. The extinct order Monura includes two species, *Dasyleptus lucasi* of the Upper Carboniferous (approximately 280,000,000 years) of France and *D. brongniarti* of the Siberian Permian (approximately 252,000,000 years) deposits. The extinct Family Triassomachilidae (order Microcoryphia) includes *Triassomachilis uralensis* of the Triassic (approximately 210,000,000 years) deposits of Russia. Other extinct species occur in the genera *Machilis*, *Praemachilis*, and *Parastylus*. One thysanuran (family Lepidotrichidae), thought to be extinct, has now been found in California.

Classification. *Distinguishing taxonomic features.* Apterygotes differ from pterygotes in lacking wings and undergoing simple metamorphosis. They differ also in the structure of the thorax and in the development of abdominal appendages.

Mouthparts are an important criterion in separating these hexapods into two groups, those with entognathous mouthparts (Protura, Collembola, and Diplura) and those with ectognathous mouthparts (Thysanura, Microcoryphia, and the extinct Monura). These groups are further differentiated on the basis of antennae (number of segments) if present, placement of eyes (simple and compound) if present, segmentation and specialization of legs, number and modifications of abdominal appendages, and number of abdominal segments.

Annotated classification. Three previous orders (Protura, Collembola, and Diplura) are treated here as separate classes. The Thysanura, Microcoryphia, and Monura are placed in the subclass Apterygota (class Insecta).

CLASS PROTURA

Entognathous mouthparts.

Order Protura

Antennae and eyes absent; anterior pair of legs elongated and sensory; 8th abdominal segment with paired openings to large abdominal glands; 3 pairs of ventral abdominal appendages; sparse setae (hairs or bristles) important for classification; scales absent.

Suborder Eosentomoidea

Tracheal system present; claw of middle and posterior legs claw-shaped; 8th abdominal segment with striate band; lids to gland openings small, unornamented.

Family Eosentomidae. Pairs of abdominal appendages alike, 2-segmented with 5 setae.

Suborder Acerentomoidea

Tracheal system absent; claw of middle and hindlegs broadly boat-shaped; lids to gland openings large and with or without teeth on posterior border.

Family Protentomidae. Abdominal appendages 2-segmented with up to 4 setae, 2 pairs with terminal vesicles.

Family Acerentomidae. Abdominal appendages 2-segmented with up to 4 setae, anterior pair with terminal vesicle.

CLASS COLLEMBOLA

Entognathous mouthparts.

Order Collembola

1 pair of 4- to 6-segmented antennae; a pair of postantennal organs; 0 to 8 simple eyes (ocelli) on each side of head; abdo-

men of 6 or fewer segments including telson, a ventral tube on 1st abdominal segment, usually a ventral furcula on 4th abdominal segment, tenaculum on 3rd abdominal segment; hairs, setae, bristles, or scales present.

Suborder Arthropleona

Head prognathous (except Poduridae, Actaletidae); postantennal organ present or absent; trunk elongated, most segments distinct.

Section Poduromorpha. Trunk segments essentially similar; 1st thoracic segment distinct with dorsal setae; antennae short; scales absent.

Family Onychiuridae. Mouthparts mandibular; complicated sensory organs on antennae; ocelli absent.

Family Hypogastruridae. Mouthparts mandibular; sensory organs of antennae simple; ocelli usually present.

Family Neanuridae. Mouthparts modified for sucking, mandibles present or absent; sensory organs of antennae simple; ocelli usually present.

Family Poduridae. 1 species, *Podura aquatica;* hypognathous head; mouthparts mandibular; ocelli present.

Section Entomobryomorpha. Trunk segments usually dissimilar; 1st thoracic segment reduced dorsally, without dorsal setae; antennae usually long; scales present or absent.

Family Actaletidae. Head hypognathous; mouthparts mandibular.

Family Isotomidae. Postantennal organ usually present; scales absent.

Family Oncopoduridae. Postantennal organ present or absent; scales present.

Family Tomoceridae. Postantennal organ absent; scales present, distinctly ribbed or fluted.

Family Protentomobryidae. Extinct.

Family Entomobryidae. Postantennal organ generally absent; antennae and furcula long; scales present or absent.

Suborder Symphypleona

Head hypognathous; postantennal organ absent; thorax and abdomen fused, segmentation lost or indistinct.

Family Neelidae. Antennae small, shorter than head, situated midway or on anterior half; ocelli absent.

Family Sminthuridae. Antennae longer than head, situated on posterior half; ocelli usually present.

CLASS DIPLURA

Entognathous mouthparts.

Order Diplura

Antennae multijointed; compound eyes and ocelli absent; abdomen of 11 segments, telson included or absent; paired ventral styli; paired cerci of variable form; setae, rarely scales.

Family Campodeidae. Antennae longer than head plus thorax; 1st abdominal segment with lobelike appendages; cerci as long as abdomen, multijointed.

Family Projapygidae. Antennae shorter than head plus thorax; 1st abdominal segment with styli; cerci less than half abdomen length, fewer than 10 joints; scales absent.

Family Japygidae. Antennae shorter than head plus thorax; 1st abdominal segment with styli; cerci paired terminal forceps; scales absent.

CLASS INSECTA

Ectognathous mouthparts.

Subclass Apterygota

Primitive wingless insects.

Order Monura

Extinct; compound eyes present; trunk not obviously separated from thorax; telson a long median tail filament.

Family Dasyleptidae. Fine hairs; abdominal styli 2-jointed.

Order Microcoryphia

Compound eyes present; 3 ocelli; telson a long median tail filament; cerci as lateral tail filaments; styli on thoracic legs and abdomen; setae and scales.

Family Triassomachilidae. Extinct.

Family Machilidae. Abdominal segments with 1 or 2 pairs of vesicles, median ventral sclerites visible.

Family Meinertellidae. Abdominal segments with 1 pair of vesicles, median ventral sclerites almost invisible.

Order Thysanura

Small compound eyes present or absent; trunk segments flattened; telson a long median tail filament; telson and cerci long or short; abdominal styli unjointed.

Family Lepismatidae. Eyes present; scales usually present.

Family Nicoletiidae. Eyes absent; scales present or absent.

Critical appraisal. Detailed studies of morphology, physiology, embryology, and life histories are required to further elucidate relationships of these hexapods.

The standard system of classification has been class Insecta; subclass Apterygota; orders Protura, Collembola, Diplura (or Aptera), and Thysanura; subclass Pterygota. This system is retained in many textbooks, but it is no longer accepted by most specialists, who have shifted the entognathous orders to separate subclasses or classes, and the culmination is the classification used above. Another classification is superclass Hexapoda, class Protura, class Collembola, class Diplura, class Thysanura, class Insecta. Under this system apterygotes are primitive wingless classes of Hexapoda.

BIBLIOGRAPHY. K. CHRISTIANSEN, "Bionomics of Collembola," *A. Rev. Ent.*, 9:147–178 (1964), a concise review of the early literature; A.D. IMMS, *A General Textbook of Entomology*, 9th ed. rev. by O.W. RICHARDS and R.G. DAVIES (1957), the standard textbook; S.M. MANTON, "Introduction to Classification of Arthropoda," in R.C. MOORE (ed.), *Treatise on Invertebrate Paleontology*, pt. R, *Arthropoda 4* (1969), a recent brief coverage of the subject; Z. MASSOUD, "Monographie des Neanuridae, Collemboles Poduromorphes à pièces buccales modifiées," *Biologie de l'Amérique Australe*, 3:7–399 (1967), an important revision and classification of Collembola with sucking mouthparts; J. PACLT, *Biologie der primär flügellosen Insekten* (1956), an extensive review of earlier literature and research; J.T. SALMON, *An Index to the Collembola*, 3 vol. (1964–65); A.G. SHAROV, *Basic Arthropodan Stock: With Special Reference to Insects* (1966), an excellent work that is based on personal palaeontological research and emphasizes phylogenetic relationships from worms to winged insects; S.L. TUXEN, *The Protura* (1964), a taxonomic coverage of this group.

(K.A.J.W.)

Aquarium

The term aquarium refers to a receptacle for maintaining aquatic organisms or to a facility in which a collection of aquatic organisms is displayed or studied.

Historical background. There is no way of knowing who the first aquarists were. The earliest known ones were the Sumerians, who kept fishes in artificial ponds at least 4,500 years ago; records of fish keeping also date from ancient Egypt and Assyria. The Chinese, who raised carp for food as early as 1000 BC, were probably the first to breed fish with any degree of success. Their selective breeding of ornamental goldfish was later introduced to Japan, where the breeding of ornamental carp was perfected. The ancient Romans, who kept fish for food and entertainment, were the first known marine aquarists; they constructed ponds that were supplied with fresh seawater through channels cut to the ocean. Although goldfish were successfully kept in glass vessels in England during the middle 1700s, aquarium keeping did not become well established until the relationship between oxygen, animals, and plants became known, a century later.

By 1850 the keeping of fishes, amphibians, and reptiles had become useful in the study of nature. It was in the works of Philip Gosse, a British ornithologist, that the term aquarium first appeared. His work and the work of others aroused increased public interest in aquatic life.

The first display aquarium was opened to the public in 1853 at Regent's Park in England. It was followed by aquariums in Berlin, Naples, and Paris. P.T. Barnum, the circus entrepreneur, recognized the commercial possibilities of living aquatic animals and, in 1856, opened the first display aquarium at the American Museum in New York as a private enterprise. By 1928 there were 45 public or commercial aquariums throughout the world, but growth then slowed and few new large aquariums appeared until after World War II.

Many of the world's principal cities now have public aquariums as well as commercial ones. Another category encompasses those aquariums that serve chiefly as research institutions. Among the best known of the latter are those at Naples; L'Institut Oceanographique, Monaco; Plymouth Aquarium, Marine Biological Station, Plymouth, England; and Scripps Institution of Oceanography, La Jolla, California. Still another category includes temporary aquariums that have served as exhibits at world's fairs and expositions.

The first oceanarium, or large marine aquarium, Marineland of Florida, was built in 1956 as a private enterprise; it featured a giant community fish tank and trained dolphins. Marineland of the Pacific, at Palos Verdes, California, and the Seaquarium, Miami, Florida, are similar. The emphasis in this type of aquarium is on very large tanks, up to 1,000,000 gallons each, in which a great variety of fishes is placed with no attempt to separate them. In the formal aquarium (*e.g.*, the John G. Shedd Aquarium, Chicago) the kinds and types of fishes are separated. The 1970 *Directory of Public Aquaria* listed 468 aquariums throughout the world.

Function and purpose. Early aquariums were considered to be special museums, but their recreational appeal was soon recognized and developed. Publicly owned aquariums are primarily educational institutions. Privately owned ones, varying in size from small one-man operations to huge oceanariums, however, also frequently provide educational programs and exhibits; for example, the oceanariums pioneering in the capture, care, and training of dolphins and whales have increased public awareness of their importance. Some aquariums built for research are not always open to the public.

Closely related to aquariums is the raising of fishes and invertebrates to provide food or stock for sport fishing. This is now usually termed aquaculture or aquiculture (see FISHING, COMMERCIAL). The management of ornamental ponds may also serve a useful purpose, particularly since filtration and water treatment are sometimes used to improve pond-water quality. Such ponds may be small garden pools, primarily for water lilies, or larger ponds on farms for the culture of food fishes.

Design and architecture. The first containers specifically designed for aquatic specimens were the strictly functional open-air tanks used by the Romans to preserve and fatten fish for market. It was not until the 18th century that the importation of goldfish into France from the Orient for aesthetic enjoyment created the demand for small aquariums; ceramic bowls, occasionally fitted with transparent sections, were produced. The large public aquariums in many European cities, built between 1850 and 1880, were influenced by current romanticism, and conscious efforts were made to create the illusion that the spectator was entering into the underwater world. More recently, the trend has been to emphasize the natural beauty of the specimens and to make a sharp distinction between the water and the viewing space.

Basic materials. Regardless of size—whether a small jar with a capacity of less than one gallon or a huge tank with a capacity of more than 1,000,000 gallons—aquariums must be constructed with care; many substances, especially plastics and adhesives, nontoxic to humans, are toxic to water-breathing animals.

Role of toxicity

Glass is probably the safest basic material, although polyethylene, polypropylene, acrylic plastics (plexiglass), and fluorocarbon plastics are normally nontoxic. Fibre glass has been widely used and is nontoxic if properly prepared. Adhesives for sealing include epoxy resins, polyvinyl chloride, silicone rubber (except for certain coloured preparations), and neoprene. Metals are not usually used, especially in seawater, which is highly corrosive. Stainless steel, which has a low toxicity, is often used, especially in freshwater systems.

The tank. A small aquarium can be constructed entirely of glass and without supporting frames by using silicone rubber as an adhesive. Fibre glass is probably the most practical supporting material for all but the largest tanks since it is lightweight, strong, does not deteriorate, and is easily fabricated into any shape. Wood, though widely used, is subject to rot and boring organisms and thus must be protected. Reinforced concrete, including special mixes for seawater, is the principal supporting material used in the construction of large aquariums.

In modern aquariums tanks of a variety of sizes and shapes are often grouped into discrete units in order to avoid the "boxes of fish" look that characterizes some of the older, formal aquariums. Dry dioramas at the rear of the tank create the illusion of distance; the tank habitat can be a natural one or one in which fibre glass has been impregnated or painted to duplicate almost any environment. Modern aquariums attempt to illustrate the natural environment of the specimens displayed.

Polished plate glass, fully tempered polished plate glass, and plexiglass are the most commonly used glazing materials. Polished plate glass is usually used only in small aquariums because it breaks into large pieces when it fails. One generally accepted practice is to glaze large tanks with two or three layers of tempered glass so that if breakage occurs it is confined to one layer. Although plexiglass is easily scratched, it can be repolished.

Accessories. Accessories for individual tanks normally include filters, air pumps, lights, and electric thermostatically-controlled immersion heaters, or perhaps alternately, some means of chilling the water. In aquarium buildings the tanks are usually grouped so that they have a common filter and method of temperature control. Water sterilizers may be included. Plumbing in large aquariums with multiple systems is sometimes complex, involving a variety of automatic controls and water-quality monitoring systems. Because of its cost and fragility, glass plumbing (*e.g.*, for aeration or circulation of water within an aquarium) is used only in cases in which low toxicity is essential. Unplasticized polyvinyl chloride pipe is widely used. Fibre glass pipe and epoxy-lined asbestos pipe are sometimes used, but lead and hard rubber pipe are obsolete. In seawater systems the growth of fouling organisms such as mussels and barnacles is avoided by providing the system with duplicate pipes and alternating their use on a weekly basis. When a line is dry the few organisms present die and are flushed out when the line is again put into service.

Nonmetallic or plastic-lined pumps are better than metal ones in terms of toxicity, but stainless steel is often satisfactory. Airlift pumps (such as those used in home aquarium subsand filters) move large volumes of water when the lift pipes are of sufficient diameter.

Generally, the most effective illumination is by incandescent lamps placed above the front glass. Fluorescent lights provide even illumination but may overilluminate the tank walls; coloured lights accentuate natural colours; and mercury-vapour lamps encourage maximum growth of marine plants.

Plants. The introduction of some form of aquatic plant life is of practical value in an aquarium, although the presence of plants can cause complications. Aquatic plants consume dissolved oxygen and give off carbon dioxide; under the influence of bright light, plants also consume carbon dioxide and give off oxygen while engaged in photosynthesis. This operates very well so long as light of a certain intensity falls on the plants—the animals thus give off what the plants can use and vice versa. Aquariums in which the plants and animals are believed to balance each other in a respiratory sense are generally referred to as balanced aquariums. This condition of balance is rarely attained, however, because under certain conditions (*i.e.*, inadequate light intensity) the plants are in direct respiratory competition with the animals. Moreover, the atmosphere constantly enters the picture, affecting the exchange of gases through the water surface. The quantities of animal and plant life should be so related to the surface area of the tank that they can survive indefinitely by means of gaseous exchange through the surface film. There is an additional relationship between plants and animals in such an aquarium: the waste products of the fishes form fertilizer or food for the plants and are consumed by them. It is, strictly speaking, such water that aquarists consider conditioned and most suitable for the specimens of interest. Furthermore, dense, slow-growing plants that consume much of the waste products of relatively few fishes as fast as they are formed usually furnish the most stable and attractive aquariums. A variety of such plants is available.

Maintenance problems. The design of a large aquarium must take into account the requirements of the specimens, especially since exhibits at modern aquariums include all types of aquatic organisms: mammals, birds, reptiles, amphibians, and invertebrates as well as fishes. Among the many factors that must be considered are

traffic flow patterns of visitors, reflections off glass, and acoustics; and tank-maintenance problems such as water clarity, dissolved wastes, temperature, tank decor, disease treatment, and nutrition.

Water quality. The primary requirement for maintaining aquatic organisms is water quality. The water supply must be free of pollutants, including sewage and industrial wastes, and it should be in gaseous equilibrium with the atmosphere to ensure adequate oxygen and to avoid supersaturation with nitrogen. In recirculating systems, water treatment must not only ensure clarity of the water but also purification of metabolic wastes. The source of freshwater is usually water supplies from which chlorine and other additives have been removed, either by carbon filtration or by the addition of a chemical. Marine organisms can be maintained in either natural or artificial seawater; the latter has the advantage of being initially free from disease organisms and pollutants but may not be as suitable for some organisms.

Types of water systems

There are three basic types of water systems: open, closed, and semiclosed. In open systems the water flows through the aquarium once and is discarded. This provides water quality comparable to that of the natural environment and there is no buildup of toxic metabolic wastes; however, temperature control and pumping are usually costly, and filtration often is necessary.

Water is continuously recirculated in closed systems and is only renewed periodically. Metabolic wastes must be treated since they are not continuously flushed from the system. An important problem is that ammonia must be rapidly removed or transformed because it is harmful even at very low concentrations. In the aquarium the bacteria that convert ammonia to nitrite reside primarily in the filter material, and a slow sand filter with a large surface area is usually provided to ensure their abundance. Plant growth in the aquarium, especially in marine systems, is not usually sufficient to utilize all the nitrate produced by bacteria from nitrite. Although some aquariums have operated many years with a minimum of water renewal, it is normally necessary to replace from 1 to 10 percent of the water per month to maintain a low level of nitrates. The use of charcoal in both freshwater and seawater systems helps to slow the accumulation of nitrogenous wastes. Metabolic wastes also cause an increase in the acidity of the water. Carbonate compounds are commonly used to maintain an optimal level of acidity, particularly when water renewal is infrequent.

Semiclosed systems are essentially the same as closed except that there is a constant connection to the water supply, and the problem of dissolved wastes is controlled by the regular addition of new water; this system is less costly than the open one with regard to temperature control and pumping.

Filters vary from simple flow-through systems to completely automated recirculating systems, with special provisions for monitoring and controlling the physical and chemical characteristics of the water.

The turnover rate, or rate of water replacement, of individual aquariums is important and should be no more than two hours. In addition, aeration by means of air stones (diffusers) should be provided to guard against asphyxia in the event of an unexpected water-supply failure.

Fishes and invertebrates can also be maintained without filtration or aeration in aquariums that are "balanced" with plants; however, the balance between plants and animals is very difficult to attain on a large scale or even in a normally stocked aquarium, especially a seawater aquarium.

Freshwater pools for mammals and birds present special problems. They generally require a higher filtration rate and greater filter capacity because they accumulate large amounts of fecal wastes. Air-breathing animals, however, are not highly sensitive to water quality; thus, chemical treatments, such as chlorination, which would kill fishes, can be used to control bacteria and to improve water clarity. Seawater formulas are simpler; for example, a 2 percent sodium chloride solution will satisfactorily maintain whales and dolphins. Seals and sea lions have been kept in freshwater, but this may increase their eye problems because of the osmotic effect of the freshwater on the eye tissues.

Feeding. The diet provided for aquarium specimens should approximate the natural diet as closely as possible. In addition to products available from the fish market, several commercially prepared diets and various live foods are available to the professional and home aquarist. As a general rule fishes do better if provided with a varied diet. Brine shrimp (*Artemia*) are a convenient source of food and are available both as newly hatched larvae or adults, either alive, frozen, or freeze dried. It should be noted, however, that they do not fill all of the nutritional requirements of many plankton-feeding fishes.

Diseases. Disease prevention is a constantly strived for goal, since diseases, once introduced into the aquarium, are often difficult to eliminate. Specimens therefore must be provided with water of suitable quality; the proper habitat, including protective cover if necessary; compatible tankmates; and an adequate diet. New specimens are usually quarantined for two or three weeks prior to placement in the display tanks.

Categories of infectious diseases

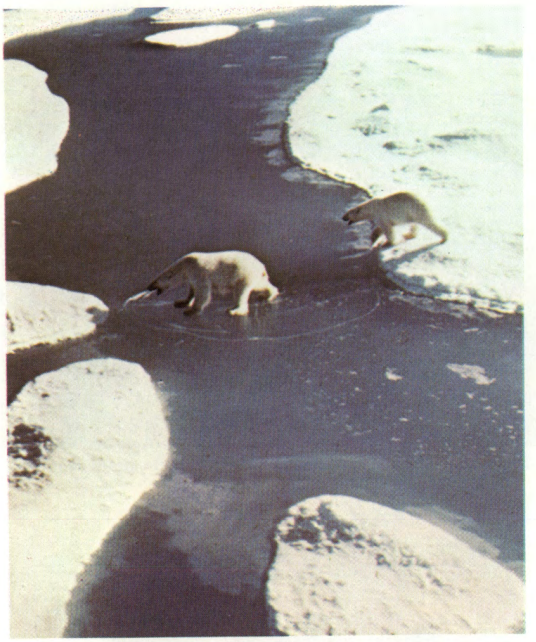
Several categories of infectious diseases can occur in aquariums. Protozoans are a common problem, especially those causing freshwater and marine white-spot diseases. Few treatments have thus far been completely successful. The problem in marine aquariums is so severe that many facilities maintain in the water a very low dose of copper sulfate as a preventive measure, despite its undesirability in terms of toxicity. Formalin and insecticides, although both toxic, are common treatments for the larger parasites that attach themselves to fish. Treatments that necessitate handling the specimens are avoided if possible. Large fishes that must be handled are usually anesthetized in order to reduce stress.

Ultraviolet sterilizer units can prevent the spread of disease in aquarium water systems if strategically placed and of sufficient capacity. Antibiotics should be used only if essential for the cure of a specific disease; indiscriminate use can not only result in resistant strains of the disease-causing organisms but also in destruction of nitrifying bacteria.

Staff. A large aquarium requires many kinds of specialists, including engineers, accountants, animal trainers, educators, and biologists. The people in the field cannot obtain formal training; as a result, aquarium experience is gained largely on the job.

BIBLIOGRAPHY. ERWIN AMLACHER, *Taschenbuch der Fischkrankheiten* . . . (1961; Eng. trans., *Textbook of Fish Diseases*, 1970), a textbook designed for those who wish to know how to identify and study fish diseases; H.R. AXELROD and WILLIAM VORDERWINKLER, *Encyclopedia of Tropical Fishes*, 11th ed. (1968), with emphasis on the breeding of fish; *Communications Ier Congrès International d'Aquariologie*, Bulletin de l'Institut Océanographique, Monaco, special no. 1, vol. A, B, C, D, the proceedings of the only international conference of professional aquarists, with information on water-treatment systems, display technique, etc., in various languages, but with English summaries; C.W. EMMENS, *How to Keep and Breed Tropical Fish* (1962), one of the best general books on keeping fish; JIM KELLY, *Pet Library's Aquarium Guide* (1969), on keeping of fish as a hobby, for the beginner; W.J. MCNEIL (ed.), *Marine Aquiculture* (1970), proceedings of a symposium on raising fish for food, with important chapters on nutrition, diseases, and cultivation methods; CRAIG PHILLIPS, *The Captive Sea* (1964), an account of the origin and organization of the Miami Seaquarium, one of the successful oceanariums; S.H. SPOTTE, *Fish and Invertebrate Culture* (1970), on the theory and management of water-quality control in aquaria, and culture systems (designed for the professional or student, but of great value to any serious aquarist); C. VAN DUIJN, *Diseases of Fishes*, 2nd ed. (1967), the most comprehensive list of treatments published in English; BARBARA and JOHN WATERS, *Salt-Water Aquariums* (1967), techniques and problems involved in establishing and maintaining saltwater aquariums; W.D. WICKLER, *Das Züchten von Aquarienfischen* (1963; Eng. trans., *Breeding Aquarium Fish*, 1966), a description of the theory and special conditions necessary to breed tropical fish, with information on the behaviour of individual species.

(D.W.W.)



Polar bears crossing ice in Norwegian Bay, Northwest Territories, Canada.



Lake near Bergen, Norway, fed by a glacier.

Vegetation clothing banks of hot springs in Iceland.



Extremes in aquatic environments

Mineral-rich thermal area, Yellowstone National Park, Wyoming.



Algae-stained thermal pool, Yellowstone National Park, Wyoming.

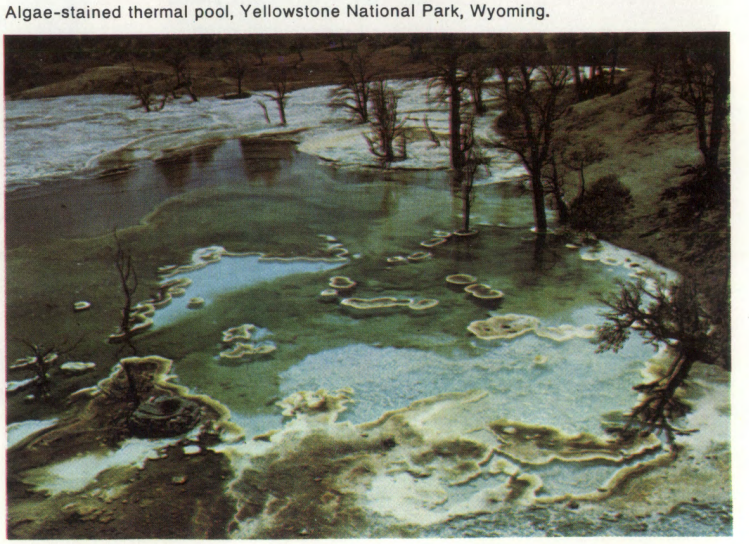


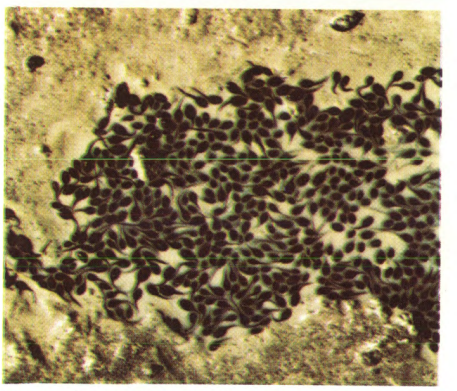
Plate 1: (Top left) Ted Grant—Information Canada Phototheque, (top right) Sven Samelius, (centre right) Peppe Traversa—FPG, (bottom left) Charles Steinhacker, (bottom right) Anthony Mercieca from Root Resources—EB Inc.



Salt-encrusted shore of Great Salt Lake, Utah.



Lake Baikal, Soviet Union, the world's deepest lake, rich in unique species.



Toad tadpoles massed in the remains of a drying puddle, Kenya.

Lake Manyara, Tanzania, a soda lake, containing only alkali-tolerant species.



Permanent pond, Loughrigg Tarn, Westmorland, England.

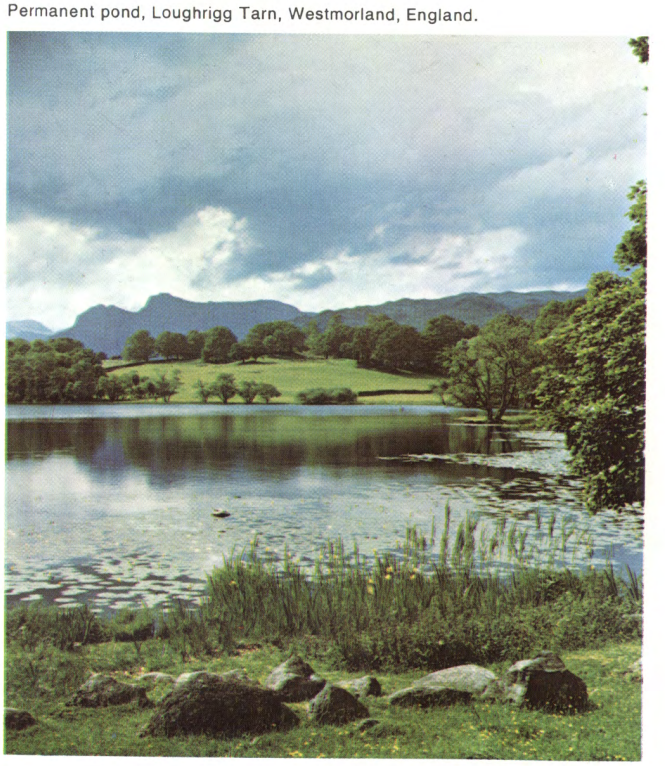


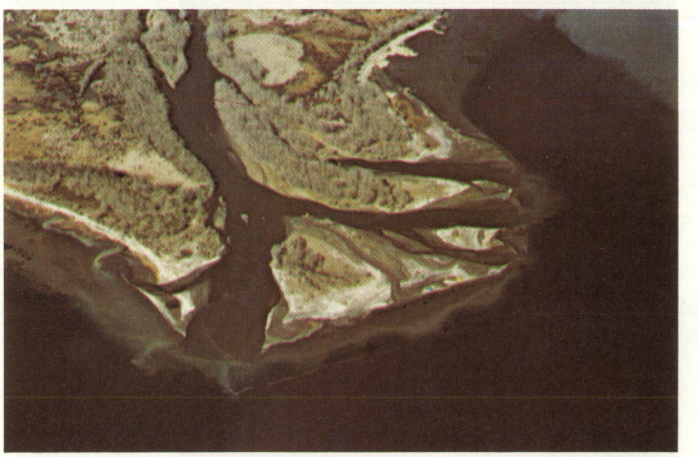
Plate 2: (Top left) Barbara Van Cleve, (top right) Harrison Forman, (centre right) Jane Burton—Bruce Coleman Ltd., (bottom left) from the book *The African Experience and The Tree Where Man Was Born*, text by Peter Matthiessen, pictures by Eliot Porter, published in 1972 by E.P. Dutton & Co., Inc., and used with their permission, (bottom right) Tourist Photo Library

Plate 3: (Top) Laurence R. Lowry—Rapho Guillumette, (centre right) G.R. Roberts, Nelson, New Zealand, (bottom left, bottom right) Bruce Coleman Inc., (bottom left) Mira Atkeson, (bottom right) James R. Simon



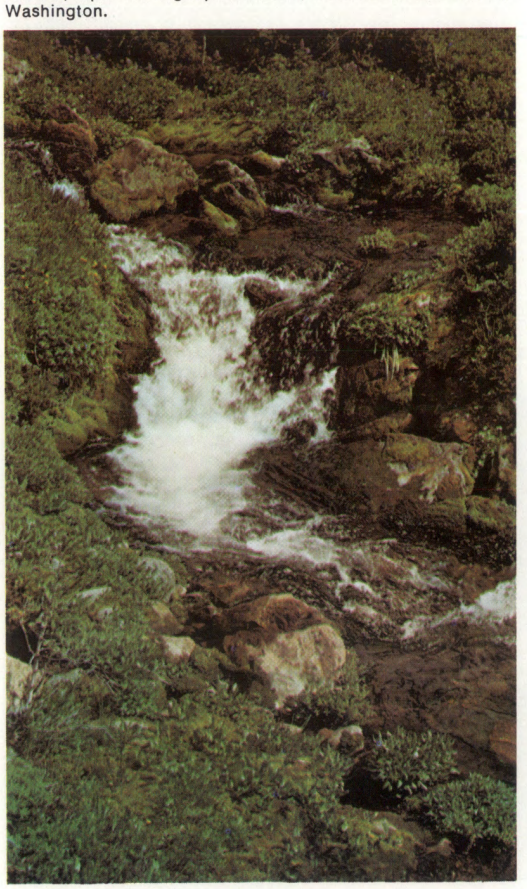
The Salinas River, California, emptying muddy water into Monterey Bay.

Delta of the Tongariro River, New Zealand, composed of accumulated silt washed downstream.

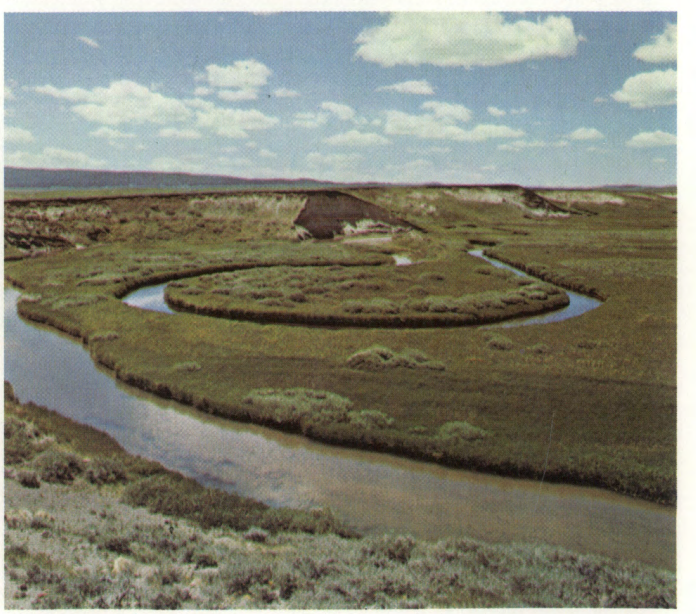


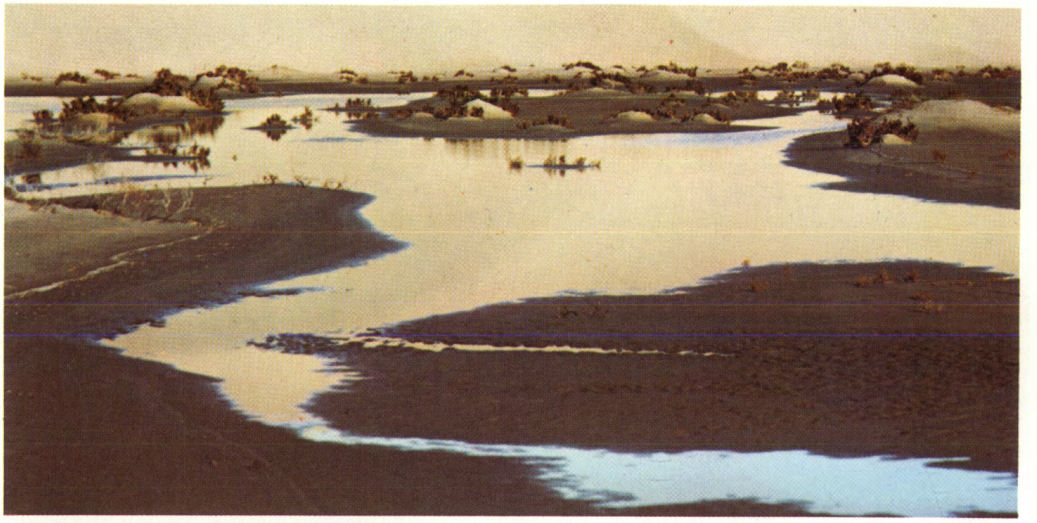
Inland water systems

A clear, rapid-moving alpine stream, Goat Rocks Wilderness, Washington.



Trout Creek, a slow-moving, mature river, Yellowstone National Park, Wyoming.





Sand dunes encroaching on salt marshes bordering the entrance of Scammon's Lagoon, Baja, California.



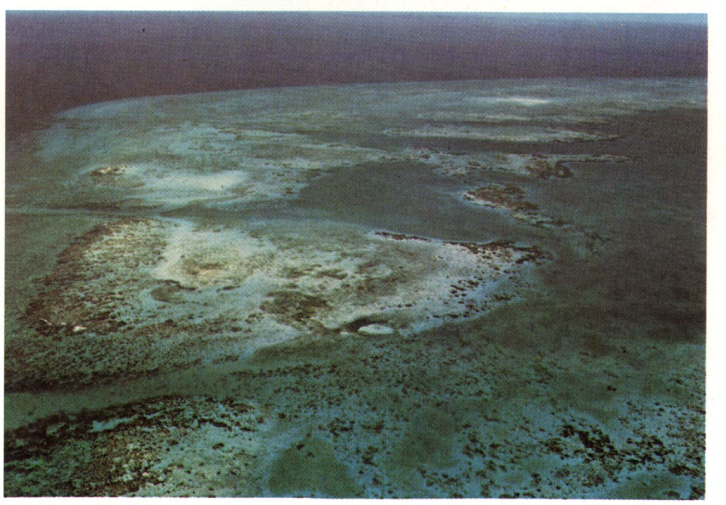
Marine life typical of that found in cold water, on a piling in Hodgkins Cove, Cape Ann, Massachusetts.

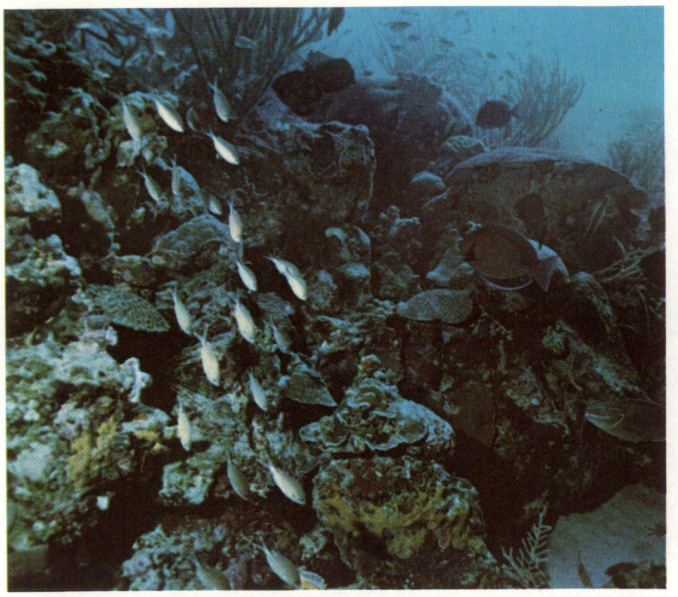


Kelp seen at low tide, Point Lobos Reserve State Park, California.



(Left) Marine life typical of that found in tropical waters, Palau Islands, Trust Territory of the Pacific Islands. (Below) Coral reef, Belize, British Honduras.





Sea bottom near Bonaire Island, Lesser Antilles.



Land spit at Hog Island, Virginia, forms the tranquil lagoon (bottom right) by separating it from the sea (top left).

Marine environments



Algal community, Sea of Japan, near Shimane.

Leaf barnacles and colony of tube worms exposed at low tide, Santa Cruz, California.

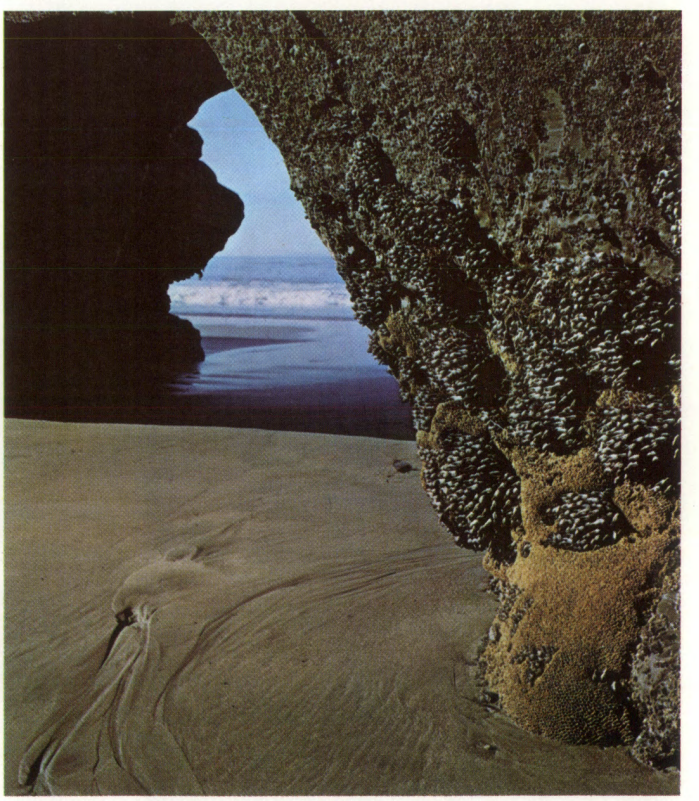


Plate 4: (Top) David Cavagnaro, (centre left, bottom left, bottom right) Douglas Faulkner, (centre right) Tom Myers

Plate 5: (Top left) Jeff Meyer—Animals Animals, (top right) © Costa Manos—Magnum, (bottom left) W.H. Hodge, (bottom right) Al Lowry—Photo Researchers



Detergent foam on Escondido Creek, California.



Duck feeding in refuse, Lake Mendota, Madison, Wisconsin.

Polluted water systems



Paper mill pollution, St. Catharines, Ontario, Canada.

Oil in surf at Santa Barbara, California.



Straw-covered tide pools soak up oil at Agate Beach, near San Francisco, California.



Plate 6: (Top) Barbara Adams from Root Resources—EB Inc., (centre left, centre right) National Audubon Society, (centre left) Raymond L. Nelson, (centre right) Lincoln Nutting, (bottom left, bottom right) Tom Myers

Aquatic Ecosystem

An aquatic ecosystem consists of plants and animals interacting with the chemical and physical features of a watery environment from the smallest puddle to the world ocean. And, while conditions in a short-lived puddle differ considerably from those in the ocean deeps, there are many points of similarity imposed upon the communities that inhabit such environments by the unique properties of water. Living forms have adapted to these environmental extremes and to the myriad intermediate aquatic environments available on Earth.

Types of aquatic ecosystems

A common distinction is usually made between seawater and freshwater or between oceans and inland waters. The oceans of the world constitute a deep, interconnecting system of saline water (salty with dissolved minerals) surrounding the continents. This world ocean is characterized by complex circulation and tidal movements. Inland waters, on the other hand, comprise a widely disparate series of bodies of water that tend, generally, to be fresh (*i.e.*, low in dissolved salts) and to drain into the oceans.

Many variations in the aquatic environment are products of organisms themselves; for example, the diatom (one-celled organisms with skeletons composed of silica) population of a lake in May is determined by chemical nutrients, but the chemical composition of the lake, particularly the aspect of it most important to diatoms, is very much the product of diatom activities in the same lake a month earlier. It is impossible to understand the life of waters without taking into account a reasonably large part of the web of circular causality that binds the system together.

THE AQUATIC ENVIRONMENT

Water as a medium for life. Because organisms are themselves aquatic systems, water provides a more kindly environment than does land (and the air above it), the other principal medium of the living world. Yet problems exist that must be overcome. Maintaining the proper internal concentration of water may cost organisms some energy in salty environments. An organism living in a concentrated desert lake or coastal lagoon may be faced with problems of water balance similar to those of its terrestrial neighbours. At the other end of the salinity scale, in freshwater environments, organisms must work to keep an excess of water from flooding in and disrupting their metabolism. In general, though, the problems of gaining and holding an adequate supply of water are very much less in aquatic ecosystems than they are in terrestrial systems.

Water has a high specific heat, which, simply stated, means that water gains or loses a large amount of heat before its temperature changes appreciably. This property of water moderates seasonal, daily, and local extremes of temperature. In even the coldest polar environments, aquatic ecosystems commonly provide refuges beneath the winter ice where the temperature is above the freezing point and far above that of the overlying air. In hot regions, evaporative heat loss rises with increasing temperature under most circumstances and keeps the water relatively cool. As a further consequence of the high specific heat of water, aquatic ecosystems do not show the pronounced microclimatic variability of many terrestrial systems. The contrast between the higher temperatures close to the ground in a sheltered hollow and the lower temperatures a few feet above an exposed ridge—a characteristic of high mountains in summer—is not found in mountain ponds, at least the surface water of which is nearly the same temperature throughout. High specific heat is also an important factor in maintaining the temperature of deep ocean water, even at the Equator, always a few degrees above freezing.

The supporting ability of water

Water is much denser than air, so the structural difficulties of maintaining a large organism in it are much less. In water, whales and sharks, much larger than any terrestrial organism—including the dinosaurs—are supported with ease. It is largely this property of water that has permitted the evolution of a style of life—that of the plankton—that has no analogue in air. Plankton is a collective term for all the aquatic free-floating organisms that are buoyant or nearly so. They sink so slowly that normal turbulence keeps most of them in suspension. Birds, bats, insects, and other organisms use the air as a medium of transport, to get from place to place, but typical plankters, by contrast, are completely at home in the water—breeding, feeding, and dying there. Their movements, if any, simply keep them in suspension and do not afford transport.

Water exhibits some important differences from related substances—such as ammonia, hydrogen fluoride, and hydrogen chloride—that theoretically could have provided a fluid medium for the maintenance of life. The hydrogen atoms in the water molecule are on one side, the oxygen atom on the other. The hydrogens being positively charged and the oxygen negatively charged creates a strong polarity and is responsible for the unusual physical and chemical behaviour of water as a solvent for many solids. Some idea of the importance of this capacity in organic evolution can be gained by comparing the relative abundance of the elements in the Earth's crust and in organisms. The elements that accumulate in living things are those that form soluble ions (charged particles) under the conditions of oxidation and acidity normally found in aquatic ecosystems.

The asymmetric arrangement of the hydrogen atoms also creates strong forces of attraction between the molecules of water, so that they tend to clump together. With, on the average, six molecules clumped together, water has much higher values for properties such as viscosity (resistance to flow), surface tension, specific heat (heat required to raise the temperature of one gram of water one degree Celsius), melting point, and boiling point than it would have if its molecules remained separate.

One other important consequence of the asymmetry of the water molecule is worth particular note. Most substances contract (*i.e.*, become denser) as they become colder. Water is unique in contracting and increasing in density down to about 4° C, after which it starts to expand again, undergoing a great expansion as it freezes into ice. If it were not for this anomalous thermal expansion of water, ice would form on the bottoms of polar seas and temperate lakes and rivers. Instead of winter water being insulated from further freezing by a surface layer of ice, summer ice would be insulated from melting by a surface layer of water. The ice would gradually increase in volume until all the aquatic ecosystems that are commonly ice-covered in winter became solid ice, and life in them would be impossible.

Physico-chemical control of life in water. Water, however, does place limitations on its denizens that air in terrestrial ecosystems does not. Since water is denser and more viscous than air, it is more difficult for an organism to move through water, a matter of consequence to an organism attempting to travel quickly or to move large quantities of water over its gills. Mixing of ingredients occurs slowly in water by comparison with air. Although a productive forest on a clear day or a downtown avenue at rush hour may show departures from the normal levels of carbon dioxide or monoxide, generally the atmosphere is well mixed. In aquatic ecosystems, mixing is so slow that local depletions and excesses of biologically important substances are common.

Mixing of ingredients in water

The mass and metabolic activity of organisms in the water are governed by many physical and chemical factors. For the earth as a whole, the most important of these is certainly light. If the light intensity throughout the deep ocean water were raised to the level in the upper 10 metres, the effects would be to increase the amount of photosynthesis and the activities of plants and animals dependent on it more than any other single change in physical and chemical factors. The fundamental control imposed by the supply of solar energy is so all-pervading that it is usually taken for granted, and investigations are limited to the experimental manipulation of other factors, particularly dissolved chemical substances known to be essential to plant growth. For particular places at particular times, it is possible to enhance photosynthesis by ex-

perimental additions of quite a wide variety of substances. Phosphorus, nitrogen, sulfur, silicon, iron, and other elements may all produce important increases in productivity when used singly, and quite dramatic increases when used in combination; commonly, however, the biggest increase for a single element is obtained with phosphorus. Potassium, a common constituent of agricultural fertilizers, is seldom or never in short supply in aquatic systems.

It is reasonable to expect an enhancement of photosynthesis if the supply of carbon dioxide is increased in a system with a very high light intensity. Certain elements required in very small amounts and vitamins, known to be essential to the nutrition of many aquatic plants, are occasionally scarce in aquatic ecosystems, but for the most part these substances seem to be available in adequate supply. This is undoubtedly because nutrients are more readily circulated through the water than they are on land.

Aquatic environments that restrict life

Animals are affected by anything that influences the plants on which they depend for food. In many aquatic situations, they also must contend with a shortage of oxygen. Many plants and animals are excluded by the extremely acid or extremely alkaline waters of a few lakes in volcanic regions or by the extreme salinity of coastal lagoons or lakes in arid lands saturated with sodium chloride. Whereas the entire range from salt-saturated lakes to fresh snow meltwater provides living space for communities of organisms, the constituent species are seldom able to tolerate more than a very small part of that range.

Major divisions of the aquatic environment. The world ocean constitutes the largest ecological subdivision on earth. Inland waters, which occupy far less area, are generally grouped into two categories on the basis of movement: standing water, or lentic environments; and flowing water, or lotic environments.

The marine division. The world ocean is enormously bigger than any lake, is much older, and has much greater local variation in surface illumination and temperature. Its biota is richer. Its chemical composition is relatively uniform, with about 3½ percent minerals, mostly salts, and a smaller percentage of organic matter in solution. Unlike the separate lakes that constitute the rest of the standing water of the hydrosphere, the ocean may be thought of as a single system, with no part completely isolated from any other.

The inland divisions. Inland waters include the standing waters of puddles, ponds, and lakes; the critical element is size. Ponds and lakes present many of the same limiting factors and stratification. Smaller bodies of standing water, such as temporary ponds and puddles, are severely limited by drying out and thus maintain transient and quite variable communities.

In common usage many lakes are referred to as seas—*e.g.*, the Caspian Sea, the Dead Sea, and the Aral Sea. The important distinction between a lake and a true sea, such as the Baltic Sea or the North Sea, is that a sea is in open communication with the world ocean; however, a lake, if connected to an ocean at all, is connected through a river that flows in only one direction, from the lake to the ocean. The common criterion of saltiness, which separates most lakes so clearly from the ocean, is not universal. Some lakes are about ten times as concentrated as open ocean water, and there are parts of the ocean, such as the upper Baltic, that would hardly be considered oceanic on the basis of salinity alone.

Flowing waters

The other inland waters are the flowing waters, which include springs, brooks, creeks, streams, and rivers. Although springs present a wide spectrum of temperatures and chemical properties, they remain relatively stable—in some cases very stable—with time. The march of the seasons does not affect them to nearly the same extent that it does other ecosystems, so the observations on one aspect of their function may legitimately be compared with observations on other aspects taken at different times. Although the inorganic chemical composition of springs is diverse, springs share the feature of having a very small part of their total content of nutrient salts combined into organic matter, either living or dead. During its long sojourn underground, spring water has lost by oxidation most of the dissolved and particulate organic matter it once contained but tends to be richer in available nutrients than most river water and therefore capable of supporting active plant growth as it issues forth into the sunlight. Rivers and streams are much more uniform in chemical composition than are springs, being integrations of the discharge of many individual springs and shallow seepage and surface runoff water. They are highly responsive to the vagaries of weather. From week to week, season to season, and year to year, the discharge and chemical composition of a river may show great variation. During a flood, when the water is laden with silt and spilling over the bank onto the floodplain, a river is a very different ecological system than it is during the clear, quiet times of low discharge.

THE OCEAN AND ITS COMMUNITIES

The oceanic environment. The ocean is by far the largest of all aquatic ecosystems. Once the ocean basins seemed stable and permanent features of the earth, but it is now known that they are continuously changing in size and configuration at rates that are almost directly observable. Sea-floor spreading from the midoceanic ridges carries the entire mass of suboceanic sediments, together with everything living on them, on a giant "conveyor belt" that spreads out from the middle ridge, apparently to plunge into the relatively small and localized oceanic deeps.

There is good reason to believe that the ocean basins have been growing steadily since their beginning. The ecosystem consequences of the concept of young and dynamic ocean basins have not been fully examined, but it is very likely that they will influence knowledge about the geochemistry of the oceans, and possibly the age, origin, and distribution of the ocean biota as well.

Oceanic circulation

The water of the ocean circulates on a global scale with a pattern that is vital to the organisms (particularly the deep-sea organisms) dwelling in it. The circulation pattern of the ocean is determined both directly, by observation, and indirectly, by inference from the distribution of water density. The basic data have been obtained from oceanographic vessels and consist of vertical profiles of temperature and salinity, the two main variables influencing the density of seawater. The driving force of oceanic circulation appears to be a combination of wind stress on the water surface, so that the sea is carried along by the atmospheric circulation, and differential heating and evaporation of the surface water at different places in the ocean.

One of the most important features of the entire circulation, with biological consequences for the deep waters of the entire world, is the cooling of surface water in the Arctic North Atlantic to the temperature of maximum density. This water, fully charged with oxygen, then sinks to the ocean floor and spreads out to all the oceans of the world. As it moves south and away from the source area there is some mixing with warmer surface water, so that a perennial thermocline (a middle layer of water with a temperature gradient) is set up at a depth of many hundreds of metres. In the upper part of the mixed layer over this permanent thermocline in temperate climates, a seasonal thermocline develops, with a thin layer of light, warm, surface water riding over the much deeper layer of seasonally mixed water.

Much of the particulate organic matter carried down in the cold, descending water masses of the North Atlantic decomposes as the water sinks. In addition, there is a steady rain of dead organisms and feces from the overlying surface water in all parts of the ocean into the cold, slow-flowing, deep water of North Atlantic origin. The decomposition of organic matter consumes oxygen, but the rate of renewal of the deep water and its oxygen content are so great that the oxygen depletion of the deep water, though easily measurable, is not sufficient to prevent oxygen-requiring animals and bacteria from living in most of the deep sea. In only a few localities, such as the Black Sea, the Cariaco Trench, off Venezuela, and

many fjords of heavily glaciated coastlines, as along Scandinavia and Alaska, is the renewal of water restricted enough to produce severe oxygen depletion comparable to that in the hypolimnion (stagnant water below the thermocline) or monimolimnion (water rich in dissolved salts, below the hypolimnion) of lakes.

Decomposition and mineralization

As in lakes, plankton and detritus sink to the deep water, releasing the nutrient salts needed to sustain plant growth. When this nutrient-rich water wells up to the lighted surface, as it does in a ring around Antarctica, off the coast of West Africa, off California, off Chile, and in various other places, it supports rapid growth of planktonic algae, which, in turn, supports large populations of animals. Many of the major fisheries of the world, such as the Antarctic whale fishery and the Peruvian anchovy fishery, are sustained by the productivity of these localized areas of nutrient upwelling.

The mineralization of plankton and detritus in the sea is never complete. A residue of undecomposed organic matter remains in extremely dilute solution and extends to the greatest depths. Since living creatures are abundant only in the illuminated surface layer of water, the mass of dissolved organic material in the sea greatly exceeds the total mass of particulate matter of all kinds —from phytoplankton (*e.g.*, algae) to whales, both living and dead.

The upper atmosphere mixes with ocean surface water quite rapidly—in about 10 years—but exchange of gases at deep levels is much slower—of the order of 1,000 years.

In other words, an average molecule of deep oceanic oxygen or carbon dioxide left the atmosphere about 1,000 years ago.

The ocean is the great sump, or drain, of the hydrosphere. Its chief contributors, the dissolved and suspended loads of rivers, are relatively easy to measure. Other gains and losses are less easily measured, but the ocean appears to have been in a chemical steady state before the advent of industrialized society. The main buffer in the system consists of reactions between seawater and solid compounds (silicates). This, for example, is the mechanism that buffers the ocean in the face of a continual input of acidic gases from volcanoes, although short-term and local buffering is dominated by reactions with carbonates and bicarbonates.

Pollution problems

The ocean receives much of man's industrial, municipal, and agricultural wastes. Some persistent synthetic chemicals, such as the chlorinated hydrocarbon insecticide DDT, are carried to the sea, where they accumulate in the fatty tissues of organisms and become concentrated as they are transferred up the food chain. The lead content of surface seawater appears to be increasing rapidly. Oil spills and leaks from oil wells, which formerly had disastrous effects only locally, now appear to be polluting the surface water on a worldwide scale. The ecological effects of these chemical changes, though they are not likely to have such devastating and immediate results as some alarmists have claimed, are nevertheless cause for serious concern. If they are reinforced by further expansion of the present regrettable practice of using the deep sea as a dumping ground for noxious substances, such as radioactive wastes and nerve gases, it is quite possible that the deep parts of the entire ocean will be profoundly modified. Because of the slowness of the deep circulation, serious and irreversible harm to the ocean ecosystem could be inflicted before the effects became noticeable.

Oceanic communities. The fossil record shows that a rich and diverse marine fauna has persisted on Earth since the beginning of the Cambrian Period (at least 570,000,000 years ago). The present oceanic fauna—clearly and smoothly traceable from that period—is incomparably richer than the fauna of any lake or river. The echinoderms, brachiopods, and several smaller animal phyla are restricted to the ocean, as are numerous groups of lower taxonomic rank.

The benthic community. Most of the world's ocean shoreline is inhabited by a flourishing benthic (bottom-dwelling) community. On shifting sandy beaches the benthos is replaced by psammon (organisms living in the water filling the spaces between sand grains). Although marine seed plants are relatively few in species, they are of considerable local importance. In general, the community of seacoasts is dominated by large algae or seaweeds: green and brown algae in the shallow waters and red algae at greater depths.

Species of the seacoasts

On wave-beaten or ice-scoured coasts the benthos must resist heavy mechanical forces, and species there show parallel evolution of body forms for attaching tightly to rocks or for burrowing in them. The rise and fall of the tides impose additional environmental stresses in terms of alternating moisture and temperature changes. Despite these difficulties, rocky shores and coral reefs are among the biological wonders of the world.

On softer shorelines, where the bottom is silt or mud, the richness of the biota is less immediately apparent because many species live in burrows in the bottom ooze. This soft-bottom fauna of the shallow seas grades into the benthos of the continental shelves and slopes and the ocean basins and trenches. In the deep-sea bottom the population of individuals is sparse, but the diversity of species is great. The ocean deeps, with their cold, chemically rich water, probably provide the most stable environment on Earth. It is not surprising, therefore, to find that their sparse food resources are partitioned among a great many species.

The benthos has considerable structure, with subcommunities of restricted depth distribution and enormous linear extent festooned along the contours of the ocean basins.

The planktonic community. The phytoplankton of the oceans is dominated by single-celled algae. In a few places, such as the Sargasso Sea, where surface currents are weak or convergent, there is some development of larger floating seaweed that contributes to the open ocean food chain, but for the most part the phytoplankton nourishes the consuming organisms. Among these consumers the most important are the copepods, which are small crustaceans that are also well represented in the lake plankton, but many other groups of animals have also adapted successfully to the planktonic niche in the oceans.

It is important that planktonic organisms remain in suspension in the water to take advantage of the light and nutrients. The consumer organisms also need to remain in the algae-rich illuminated surface layers to survive. Several mechanisms are utilized for suspension. Certain species use flotation devices such as oil droplets or gas bubbles. Many organisms are so small or so shaped that they can be maintained in suspension by ordinary turbulence. Since seawater is generally denser than freshwater, larger organisms have been able to maintain a planktonic existence in the denser medium. Even such an unlikely group as the mollusks, with their large calcareous shells, have been able to adapt successfully to a partly planktonic life in the open sea.

Flotation mechanisms

Many planktonic species, both in the sea and in lakes, go through an extensive vertical migration each day. In one common pattern, the population spends most of the daylight hours at low levels in the water, rising toward the surface at dusk. Toward midnight there may be a slight tendency to sink, followed by an early morning rise, and, finally, descent to the day depth as the sun comes up. Apparently, the plankton tends to follow a preferred illumination intensity in the water, complicated by a tendency to stop swimming and to sink passively during the darkness. Experiments have shown that the actual behaviour is more complex and involves biological rhythms set from time to time by reference to changing light intensity.

In one respect marine and freshwater plankton differ markedly. Like many other groups of marine organisms, some members of the oceanic plankton are strongly bioluminescent—*i.e.*, capable of producing light. The biochemical basis of bioluminescence has been understood for many years, but its function in the lives of most of the organisms showing it is still obscure. The light of a firefly is a sexual signal, and that of a deep-sea angler fish conceivably might be a lure, but it is not known

what role light plays in the lives of so many marine plankters—particularly since their close freshwater relatives who live similar lives do not bioluminesce. Perhaps the light-producers adjust the frequency of their spontaneous flashing to transmit solar time signals into the depths.

Despite the size of the ocean and its differences from place to place, marine biologists have achieved an enviable level of insight into the dynamics of the plankton community. It is not possible, however, to extend this approach to include the nekton (free-swimming organisms), largely because the larger animals have more varied tastes. Their position in the food chain covers several trophic (nutritional) levels and cannot be specified with the precision of the phytoplankton or the herbivorous zooplankton.

The nektonic community. The nektonic (free-swimming) community of the sea is diverse and specialized.

Dominance of vertebrates in the nekton

Vertebrates predominate, but there are also highly evolved mollusks in the nekton, particularly squid and cuttlefish, which are admirably adapted to a predatory existence.

The nekton includes sea snakes, turtles, crocodiles, whales, carnivorous mammals, cephalopod mollusks, and fishes. Its most characteristic mode of life is peaceful filtering of plankton from seawater, in the fashion of herring, whale sharks, and baleen whales, but the fast, effective predators, such as hammerhead sharks and killer whales, with their greater potential for human injury, exert a firmer hold on the human imagination. For no other ecological community has the mythological fauna survived so tenaciously into modern times—nourished by the sufficiently awesome reality of giant squid and great white shark, by the unexpected discovery of living lobe-finned fishes (*i.e.*, the coelacanths), and by the undeniable possibility that the enormous expanses of the open sea may harbour some large animal radically different from those already known.

INLAND WATERS AND THEIR COMMUNITIES

For purposes of convenience, ponds and lakes are described as lacustrine ecosystems and springs and rivers as riverine ecosystems.

Lacustrine ecosystems. *The lacustrine environment.* Most natural processes tend to obliterate pond and lake basins, either by filling them with sediment or by draining them, so lakes are, geologically speaking, rather ephemeral features of the landscape. Their distribution is spotty, but most are to be found in areas that have been glaciated recently (within the last 10,000 to 15,000 years), that are presently underlain by permafrost or soluble rocks, or that are subject to volcanic or earth-shaping activities. Glacial lakes are most abundant and vary in size from very small ponds to Lake Superior; few of them are older than 15,000 years. Most of the more ancient lakes of the world occupy basins formed when parts of the earth's crust fell below the level of the countryside around them. These lakes are ancient for several reasons: the initial basin may be very deep, so that it takes a long time to fill with sediment; the sill is very broad and not readily breached by an outflowing stream; and, perhaps most important, renewed faulting activity continues to deepen the basin.

Seasonal cycles of lakes

Most lakes experience seasonal cycles of heating and cooling. Even in the tropics, where annual changes in air temperature are muted, the change from dry to rainy season greatly decreases the solar energy delivered to the surface of a lake. In the temperate zone and in the Arctic, the difference in heat income between summer and winter is one of the dominant features of a lake ecosystem.

If a lake is reasonably deep relative to its surface extent, the wind is unable to mix into the deep water all the heat that is delivered to the surface during the warm season, so that a layer of light, warm, surface water rests upon a layer of heavy, cool, deep water. With the approach of winter the surface temperature may fall below 4° C (39° F) and, especially when ice forms on the lake, inversion of water occurs, and a winter stratification becomes established. Such a stratified lake consists of an upper layer (epilimnion), stirred by the wind to an almost uniform seasonal temperature, a middle layer (thermocline), in which the temperature changes rapidly with depth, and a lower layer (hypolimnion), in which the temperature change is slight. In a few lakes, an additional layer (monimolimnion) of water made dense by dissolved salts lies beneath the hypolimnion, from which it is separated by a constant chemocline layer.

In very transparent water there may be sufficient light in the hypolimnion to support some photosynthesis, but food is commonly produced only in the epilimnion, although it is consumed throughout the lake. The epilimnion is always fully charged with oxygen, because both photosynthesis and free exchange with the atmosphere occur there; the hypolimnion loses oxygen gradually through the season of stagnation because of its consumption by animals and bacteria. At the same time, the supply of nutrient salts in the epilimnion is depleted and that of the hypolimnion enriched. If the hypolimnion becomes sufficiently anoxic (devoid of oxygen), the release of nutrient materials from the mud is enhanced and fish may be unable to live in it.

Importance of knowledge about lake stratification

Recognition of the features described above is critical to the understanding and rational management of any stratified lake. The changes that have occurred in the Swiss lakes, which are stratified, are typical of what is happening to similar lakes round the world. Many of the Swiss lakes formerly supported large populations of whitefish, which could not tolerate high surface temperatures and retreated to the hypolimnion each summer. As cities grew by the lakes and began enriching the surface water with nitrogen and phosphorus from their sewage-treatment plants, the further depletion of hypolimnetic oxygen became so severe that all the whitefish died.

Lacustrine communities. Most young lakes are inhabited by organisms with a wide distribution and with the ability to adapt from one transitory lacustrine environment to another. Many of the smaller species have resting stages resistant to drying and are readily transported by the wind or by wading birds. Others are normally riverine inhabitants that moved in to exploit the living space afforded by a new lake on their home river system. Ancient lakes, however, have an endemic (native) fauna of fishes, mollusks, and crustaceans. Lake Baikal, in Siberia; lakes Victoria, Tanganyika, and Nyasa, in Africa; Lake Ohrid, in Yugoslavia; Lake Titicaca, in the Andes; and Lake Lanao, in the Philippines, are all famous for their endemic species, each of which is restricted to one lake. The mechanism of evolution of these species has been the subject of much conjecture and debate but is still obscure (see also EVOLUTION).

Large lakes more than a few metres (a few yards) deep are dominated by the plankton community, in which the producing organisms are mostly minute algae; in smaller or shallower lakes and in ponds a variety of rooted aquatic plants may be more important. In clear water, the aquatic plants may consist mostly of a basal rosette of leaves, from which a flowering stalk rises to the surface for pollination. A more common life form, however, and one that is more effective in turbid or highly coloured water, consists of a rooted stem that carries both its leaves and flowers above the water. The leaves are commonly long and slender, as in grasses, sedges, rushes, and cattails, or shaped like an arrowhead.

In deeper water, in which it is not feasible to form and maintain a stalk of adequate length in the face of winds and waves, the common life-form of rooted aquatic plants consists of a long flexible stem attached at the base to embedded roots or rhizomes and a flattened leaf floating on the water surface; water lilies are a common example.

The neuston community

Another lacustrine community of importance is the neuston, consisting of unattached organisms associated with the surface film of water. Most neustonic organisms are microscopic and are noticed only when they reach such numbers as to impart to the water surface a brilliant green or vermilion colour; others, such as duckweed, Nile cabbage, and various water ferns, are large

enough for individuals to be seen easily. They sometimes form a mat dense enough to destroy communities in the water beneath them and to hinder navigation. The growth of ferns on the surface water of sheltered bays around Lake Kariba, a body of water on the Zambezi River, in Africa, is solid enough to support the weight of an elephant, and the water beneath it is dark and devoid of oxygen.

Organisms living in and on the lake bottom away from the shore zone are commonly referred to as the benthos. In deep water they are mostly invertebrate animals and bacteria. Worms, nematodes, mollusks, midge larvae, and crustaceans are common benthic organisms. Some authors follow the German usage of separating the organisms attached to a surface, or *Aufwuchs*, from the ooze dwellers. Still others recognize the community of organisms attached to submerged stems and leaves of plants as the periphyton.

A smaller community in most lakes, but an ecologically interesting one, is the psammon, the assemblage of plants and animals inhabiting the tiny spaces between the grains of sand on a beach. Considerable light penetrates a sand beach because the sand is translucent rather than opaque. A well-developed community of photosynthetic organisms as well as animals and bacteria lives in this unlikely niche.

Probably the most familiar of the communities inhabiting lakes is the nekton, which is composed of the large active swimmers, many of them predatory, such as fishes, seals, turtles, otters, hippopotamuses, and crocodiles.

All of these various communities are composed of producers, the photosynthetic organisms, and consumers, the animals that subsist on the producers; a class of organisms called reducers include fungi and bacteria, which decompose organic debris. Some of the animals are herbivores, feeding directly on plants; others are carnivores, preying on the herbivores and other carnivores. The specification and quantitative description of these feeding relations is seldom easy, but such information is more central to understanding how an aquatic ecosystem functions than is a description or map of the various communities living in it.

Ecological significance of flowing waters

Riverine ecosystems. *The riverine environment.* By volume, the rivers of the world constitute a very small fraction (1 percent) of the total amount of water in aquatic systems, but their high rate of flow gives them great ecological significance. As active agents of mechanical and chemical erosion, rivers exert a major control upon the terrestrial ecosystems of their watersheds, as well as any lakes or oceans into which they discharge.

There are significant differences among the rivers of the major continents regarding chemical composition, concentration, and rate of erosion. Australia is so dry and Antarctica so cold that they make little contribution to the global budget for river transport. The streams of Asia carry almost three times as much suspended material per unit area as do those of any other continent. The river waters of South America are the most dilute; those of Africa, being similar in composition, are about twice as concentrated. North America, Asia, and especially Europe differ from the two more southern continents in carrying a much larger quantity of calcium and bicarbonate. The low salt concentration of South American rivers is offset by their very great discharge.

Rivers, particularly in arid lands, show great annual fluctuations in size and other properties. Variations of four times the normal discharge and six times the quantity of suspended particles are not unusual and are accompanied by great changes in current velocity, concentration of dissolved substances, and shape of the channel. The formulation of the general properties of a riverine ecosystem is accordingly much more difficult than it is in lakes and the ocean, and, although considerable attention has been devoted to the biology of streams, the results have tended to be descriptive and individual, with little applicability to streams in general.

Stream zonation

Differences between the headwaters and the downstream reaches of many streams make them very different habitats for plants and animals. In general, the headwaters tend to have steep beds, to be rocky, and to be shaded by trees. Near the mouth the gradient of the stream bed usually is less, the bottom composed of finer material, and the river too broad for its central part to be shaded by trees. Contrary to expectation, there is no marked difference in current velocity between the headwaters and the downstream stretches of most rivers. A headwater stream cascading down a rocky mountainside gives a subjective impression of fast current, but its channel is rough and has a small cross-section, both of which more than compensate for the steep gradient. Downstream the bed is much smoother, and the cross-sectional area much greater, so a more gentle gradient commonly results in a current velocity that is greater than the velocity of the turbulent headwaters.

Riverine communities. Rocky streams tend to be broken up into alternating stretches of riffles and pools. In riffles the flow is fast and turbulent, the water shallow, with light reaching the bottom. The rocks of riffles provide a good surface upon which plants can attach, and the animals that live among the rocks and plants frequently display adaptations of form befitting their swift-flowing environment. Microscopic organisms, such as protozoans, are able to live in a thin layer close to the rock surfaces; thus, they are not likely to be swept downstream. Larger animals, such as mayfly nymphs, however, may be flattened and streamlined to minimize the risk of being torn from their living place and carried away. Such evolution of body form has occurred also in fishes, the most extreme adaptations of which are to be seen in some Himalayan species that have suction-cup-like fins that hold them close to the rocks.

Obtaining an adequate oxygen supply is no problem in the riffles, and so the development of gills and pumping mechanisms is not so striking as in the inhabitants of pools.

Plankton, especially phytoplankton, is poorly developed in most rivers, except in the broad, open, downstream waters. In most river systems, attached algae contribute more to photosynthesis than do planktonic algae, and higher plants may also be important. Some of the latter, such as the Podostemaceae, are so specialized for living in rivers that they have lost most of the obvious features that would identify them as flowering plants.

The importation of food into riverine ecosystems

Most of the food energy to support the food chain of a riverine ecosystem, though, is not the product of stream-dwelling plants; rather, it is from terrestrial ecosystems. Under natural conditions leaf litter is probably the most important element in the basic nutrition of a stream; in thickly populated regions, leaf litter is commonly supplemented by massive quantities of municipal sewage that support, directly or indirectly, much denser populations of animals, including fish, than would be maintained by any ecosystem dependent on its own photosynthesis.

It is clear from the size and specialization of riverine fish faunas that most large rivers have had a separate and continuous existence for a very long time. Temperate rivers that were covered by continental glaciers 10,000 to 15,000 years ago, such as those of western Europe, have a poor fauna by comparison with old rivers, such as the Mississippi, which maintained ice-free downstream waters while its upper tributaries were covered by the Pleistocene ice sheets. Large tropical rivers, older still, have an even richer fauna.

BOUNDARY ECOSYSTEMS

Between the water and the land, and between one kind of water and another, a diversity of ecosystems exists. Small in extent, they differ from each other more, in some respects, than do standing and flowing waters; for convenience, they are here treated as a group.

Between waters. Boundary systems in which there is an appreciable current are most like oceans, lakes, and rivers. Estuaries, in which seawater is diluted by freshwater inflow, share many features with rivers and the ocean. An estuary has a net seaward flow equal to the discharge of its tributary rivers, but that flow is affected by the surge of tidal currents. It also has other distinctive

Estuaries

features of circulation. As river water in an estuary reaches seawater, the river water, being lighter, rides on top and, having a considerable velocity, tends to drag the upper part of the seawater underlayer along with it. Under steady-state conditions, this salt water carried seaward by the fresh surface flow is continuously replaced by a landward flow of seawater at greater depths. Variations of this basic pattern are produced either by differences in size and shape of the estuary or by differences in the relative importance of river discharge and tidal mixing. The typical estuary may, however, be thought of as a two-way flow: the seaward-flowing water is lighter, at the surface, and of low salinity, and it courses along the right bank of the estuary (facing downstream) in the Northern Hemisphere and along the left bank in the Southern Hemisphere, because of the inertial effect of the Earth's rotation; the landward flow is heavy, deep, and saline and courses along the bank opposite from the seaward, surface flow.

A pollutant introduced into an estuary, therefore, is partitioned between the two flows and moves landward as well as seaward (see ESTUARIES).

Coastal lagoons in arid regions have a current system opposite to that of estuaries. Seawater flows inward on the surface at the inlet to replace the water lost from the lagoon by evaporation, and very dense water flows outward at depths and maintains the salt balance of the lagoon. The same sort of circulation operates in the Mediterranean under modern climatic conditions, but in the past, with a different climate and with a different sea level, some parts of the Mediterranean were anoxic, suggesting that it has changed from a gigantic estuary to a gigantic lagoon (see LAGOONS).

Similarity of coral atolls and salt marshes

Between water and land. Coral atolls and salt marshes are both systems in which a solid surface is bathed by a continuously renewed supply of water. In the case of the atoll, it is full-strength seawater, swept over the reef by the trade winds. In the case of the salt marsh, it is estuarine water flushed in and out by the tides.

The coral community is diverse, with many species of organisms. The salt-marsh community, however, is simpler, with many individuals of a relatively small number of species. Both systems provide a surface to which producers can attach as they extract nutrient salts from the flow of water past them (see CORAL ISLANDS, CORAL REEFS, AND ATOLLS; LAGOONS).

Groundwater is not without life. A sparse fauna lives in subterranean streams, feeding on the detritus washed down from the surface or carried into caves by other animals. Blanched and eyeless groundwater fishes and crustaceans are found chiefly in caves or where wells pierce the water-bearing layer in which they live (see SPRINGS AND WELLS; CAVES AND CAVE SYSTEMS).

In addition to salt marshes, with their abundant tidal flow, there are many kinds of wetlands that have so little flow that they are best considered as semiterrestrial analogues of standing waters. The principal distinction of importance is between swamps and marshes, on the one hand, which have access to mineral-rich groundwater; and bogs, on the other, which are insulated from groundwater by a thick layer of peat and receive most of their water and mineral salts through the atmosphere. The terminology of wetland types is confused by the use of many names, such as fen, mire, and carr, all of which, like swamp, marsh, and bog, are, by origin, simply local names meaning "wet land." For ecological purposes it seems sufficient to distinguish wetlands with trees as swamps; wetlands without them as marshes; and wetlands receiving their water and salts from rain as bogs.

Because of reactions between peat and bog water, the supply of dissolved salts is decreased below that in rain. As a result, the nutrients needed by plants for growth are in very short supply, and many bog plants—such as sundews, pitcher plants, and bladderworts—have evolved the habit of trapping insects and other small animals to augment the meagre supply of mineral food in the water.

The great accumulation of acidic peat under a bog—often raised some metres above the level of the surrounding countryside—is an extreme case of one of the general puzzles of aquatic ecosystems: in few of them is all the organic matter produced in photosynthesis broken down completely after the death of organisms. A certain residue of organic matter remains and is preserved in a basin of sediment. Evidently, the sedimentary bacterial community is incapable of anything approaching complete decomposition of the organic matrix in which it lives. It is a curious fact that no micro-organisms make use of this undecomposed residue—a fact to which man owes his supply of fossil fuels, such as coal, gas, and oil.

Incomplete decomposition in bogs

PRODUCTIVITY IN AQUATIC ECOSYSTEMS

The problem of determining productivity. The best basis for comparison among ecosystems is probably photosynthesis by green plants. There are aquatic systems without photosynthesizers—a cave stream or the depths of the ocean, for example—but they have no primary production. The food that sustains them is imported from outside the system. They may have a productivity of blind crayfish or brittle stars, but their net production of organic material is negative. They are thus consuming, not producing, systems.

Unfortunately, there is no generally applicable method for measuring primary productivity in all aquatic systems. In places with a strong seasonal pulse of growth—as in a temperate cattail marsh—the maximal amount of plant material present after a growing season can be harvested in order to estimate net primary production, which is somewhat smaller than the gross rate because it takes no account of food consumed by the cattails to maintain themselves over the season of growth or of cattail substance that may have been consumed by animals or bacteria during the same span of time. Other methods depend on measuring changes in the quantities of various substances, commonly oxygen or carbon dioxide, in the medium in which the plants are growing.

Reasons for low productivity in open oceans

Comparisons of productivity. The open oceans are among the least productive aquatic ecosystems. Conditions of light and temperature are very favourable for plant growth over the tropical oceans all the time and over temperate and some polar seas at least seasonally, but nutrients are in short supply. For a brief period in spring, photosynthesis is rapid in temperate waters, but the nutrient supply is quickly exhausted. Thermal stratification reduces the rate of delivery of new nutrients from deep waters, and there is a continual loss, during the summer season, of the nitrogen, phosphorus, and silicon that were incorporated into organic substance during the flush of spring growth. Not all is lost, however, because local recycling of nutrients in the surface water permits some photosynthesis to take place, but at a diminished rate.

Places in the open ocean that maintain a high level of production are those in which upwelling brings a continuous supply of nutrient-rich water to the surface. Most oceanic plankters are effectively trapped in a parcel of water, from which they lack the swimming power to escape. After exhausting the local supply of nutrients, they are obliged to curtail their activities. The plankters in a region of upwelling, by contrast, are slowly sinking through a rising water column and thus are exposed to a continuously renewed medium from which they may extract nutrients indefinitely. It is no accident that the most productive open-sea fishery in the world, the Peruvian anchovy fishery, is based on plankton-eating fish living in areas of deep-sea upwelling.

The attached plants of a coral reef, seashore, or salt marsh are able to maintain a much higher rate of production than the plankton, because water is continuously washing over them. Near a continental landmass there is a good chance that the water will be reasonably high in nutrients that have just been brought to sea by rivers, or very high if much sewage is present. These systems, although they may be as productive as any in the world, are small relative to the great expanse of open ocean and contribute little to the total for oceanic primary production. The very extensive blue-water areas of the ocean sea compare in productivity with deserts on land,

with the result that the sea, although it comprises about 70 percent of the area of the Earth, accounts for only some 30 percent of its photosynthesis.

The limited resources of the sea

This low productivity of the ocean generally is a fact of social as well as scientific significance. Early estimates of marine production were misleadingly high and generated a premature hope that the sea represented a vast untapped food resource that could be used to feed an expanding human population for a long time to come. Actually, the food resources of the sea are quite limited. Marine organisms are small and have a very active metabolism; thus, a large part of marine production is consumed by the producing organisms themselves. Because many generations per year contribute to the modest production of the sea, there is never a large harvestable crop available for convenient gathering, as there is in a forest or cornfield. Even the zooplankton, which consists of larger organisms with a longer life cycle, is not a feasible food source, because the energy needed to collect the zooplankton is equal to the food energy it yields.

The best hope for a sustained food supply from the sea seems to be the whales and clupeid fishes, such as herring, anchovies, and sardines. Whales are currently being harvested faster than they can replenish their populations, and it is suspected that most of the clupeid and other pelagic fish stocks of the world have become overfished as well. At the same time, pollution and general disturbance of estuarine breeding grounds threaten shellfish and game fish of the inshore waters; these make a negligible contribution to man's caloric intake but a very important one to the quality of life.

The primary production of rivers is even less thoroughly known than that of the ocean. In the food budget of most rivers, photosynthesis, the gauge of primary production, is less important than imports from the land. Headwater springs, however, can be very productive, because most of their nutrients are in a mineralized condition and are immediately available for plant growth as they issue from the ground. Hot springs are interesting because of their extremely high productivity and the insight they provide into the evolutionary adaptability of some organisms, particularly blue-green algae, to the demands and opportunities of a sustained high-temperature environment.

Where nutrients are scarce, as in areas of thoroughly leached soil over old crystalline Precambrian rocks, lakes are extremely unproductive. The same is true of polar lakes, which are cold, in the dark much of the year, and ice-covered during most of the illuminated season. Lakes underlain by strata that weather readily to yield a good supply of nutrients, such as some sedimentary and some volcanic rocks, may be very productive. In the tropics, where the water is warm and well lighted at all seasons, lakes surrounded by fresh volcanic ash—like coral reefs and hot springs—approach the upper limit of productivity.

Wetlands with a good supply of water rich in nutrient salts are rather productive. African papyrus marshes appear to compare favourably with fertilized and irrigated fields of sugarcane and corn (maize) and to be considerably more productive than drier terrestrial systems. The surface of acid bogs, however, is unable to support a high rate of organic production.

BIBLIOGRAPHY. N.J. BERRILL, *The Life of the Ocean* (1966), a popular account of marine biology by a professional marine biologist; RACHEL CARSON, *The Sea Around Us*, rev. ed. (1961), an elementary account of the oceans and one of the most successful popular science books of all time; D.G. FREY (ed.), *Limnology in North America* (1963), a geographically organized account of the lakes, and to some extent of the rivers, springs, estuaries, and wetlands, of North America, with a useful annotated bibliography; R.M. GARRELS and F.T. MACKENZIE, *Evolution of Sedimentary Rocks* (1971), includes a lucid and informative synopsis of the geochemistry of aquatic systems; E. GORHAM, "The Development of Peat Lands," *Q. Rev. Biol.*, 32:145–166 (1957), a short account of the main features of wetland ecology; JAMES GREEN, *The Biology of Estuarine Animals* (1968), a store of information about estuarine animals and how they fit into the ecosystem; G.E. HUTCHINSON, *A Treatise on Limnology*, vol. 1, *Geography, Physics, and Chemistry* (1957) and vol. 2, *Introduction to Lake Biology and the Limnoplankton* (1967), the first two parts of a planned 3-volume encyclopaedic work on lakes, probably the most comprehensive ever attempted for any type of aquatic ecosystem; H.B.N. HYNES, *The Ecology of Running Waters* (1970), a comprehensive book on the biology of rivers; L.B. LEOPOLD, M.G. WOLMAN, and J.P. MILLER, *Fluvial Processes in Geomorphology* (1964), a moderately technical introduction to the variability of rivers; D.A. LIVINGSTONE, "The Sodium Cycle and the Age of the Ocean," *Geochim. Cosmochim. Acta*, 27: 1055–1069 (1963), a balance sheet for gains and losses of sea salt over geologic time; G.K. REID, *Ecology of Inland Waters and Estuaries* (1961), a unique attempt to cover the ecology of lakes, rivers, and estuaries in relatively simple fashion; F. RUTTNER, *Grundriss der Limnologie*, 3rd ed. (1962; Eng. trans., *Fundamentals of Limnology*, 3rd ed., 1963), a short introduction to lakes, with emphasis on their biology; H.U. SVERDRUP, M.W. JOHNSON, and R.H. FLEMING, *The Oceans: Their Physics, Chemistry and General Biology* (1942), a classic text on oceanography, with a clarity of exposition that makes this still a useful introduction to the sea despite many advances since it was published; J.M. and M. TEAL, *Life and Death of the Salt Marsh* (1969), a lyrical but authoritative account of the salt marsh ecosystem for nonspecialists; P.K. WEYL, *Oceanography: An Introduction to the Marine Environment* (1970), a spirited attempt at a nontechnical exposition of general oceanography.

(D.A.L.)

Aqueduct

The word aqueduct comes from the Latin *aqua*, "water," and *ductus*, "to lead," and in the broad sense refers to a main man-made conduit for carrying water. In a more restricted sense it is a structure used to conduct a water stream across a hollow or valley. As generally used in modern engineering, the term aqueduct applies to a system of pipes, ditches, canals, tunnels, and supporting structures used to convey water from its source to its main distribution point. Subsidiary distribution lines are not normally classed as part of the aqueduct. For a comprehensive discussion of modern water supply, see WATER-SUPPLY SYSTEMS.

Early history of aqueducts

IN PRE-ROMAN TIMES

The *qanaat*

Although the Romans are considered the greatest aqueduct builders of the ancient world, they were by no means alone nor were they the first. Examples from other civilizations are many. From time immemorial *qanaats*, or tunnels, driven horizontally into hillsides transported water by gravity to irrigate the plains of Persia. This construction was particularly important in areas where small amounts of water were transported great distances and evaporation was a serious problem. Before 500 BC the *qanaat* was known in India, Armenia, Egypt, and many Near Eastern countries. In Mesopotamia, clay tablets dating possibly to the 25th century BC make frequent reference to canals or reservoirs supplying cities between the Tigris and Euphrates rivers. Letters of Hammurabi (*c.* 1760 BC) refer to cleaning of canals. The Minoans, who flourished on Crete (*c.* 2000–1400 BC), had highly developed sanitary facilities, flushed with water, presumably tied to an aqueduct system.

One of the best known early tunnel aqueducts was built about 700 BC under King Hezekiah to supply Jerusalem with water. Cut through solid rock, the conduit (1,750 feet, or 530 metres, in length and six feet, or two metres, in height) brought water from the spring of Gihon to the reservoir or pool called Siloam. Built about the same time (691 BC) under Sennacherib was the great Assyrian aqueduct of Jerwan, which joined earlier systems to bring water from a tributary of the Greater Zab to Nineveh, some 50 miles (80 kilometres) away. One of the most striking features of this system was a limestone bridge over 900 feet (300 metres) long that carried the waters over a stream and valley. Standing 30 feet (10 metres) high in some places, this portion of the aqueduct had five corbelled arches in the centre and carried a channel exceeding 50 feet (15 metres) in width. It is estimated that over 2,000,000 blocks of heavy limestone were used in its construction. Employed for both irrigation and domestic

purposes, this appears to be the earliest of significant public water supplies extant.

Another notable aqueduct, named by Herodotus as one of the three greatest of Greek engineering works, was that on Samos, probably dating from the 6th century BC. Built under the direction of Eupalinus of Megara, the first hydraulic engineer whose name has been preserved, the aqueduct consisted of a tunnel and a pipeline supported on stone pillars. This tunnel, some 3,300 feet (1,000 metres) in length and nearly six feet (two metres) square, was bored from both ends (the headings missed connections, resulting in a jog in the middle) through a mountain of limestone rock nearly 1,000 feet (300 metres) high. The water was carried through clay pipes laid in a sloped recess in the bottom of the tunnel.

Most ancient aqueducts simply conducted the water by gravity, free of pressure, in pipes, channels, or tunnels along the hydraulic gradient, gradually descending from a higher to a lower elevation. Thus, water channels were normally carried across valleys on piles of stones or arches. Introduced by the Greeks in Asia Minor and elsewhere, the siphon (a pressure pipeline or tunnel) was a new hydraulic device that could be used as a short cut to cross steep valleys. Since it usually followed the contour of the ground down one side of the valley and up the other, it was, of course, subjected to greater pressures in its lower portions. The scarcity of good materials for piping and jointing obviously limited its use. One example of a sizable Greek siphon was that built at Pergamon *c.* 180 BC. The water was carried by gravity some 35 miles (55 kilometres) from mountain springs to two settling tanks at Áyios Yeóryios two miles (three kilometres) east of and about 130 feet (40 metres) above the city. From this point the water passed through pressure pipes that crossed two valleys—one of which was more than 600 feet (200 metres) below the reservoir—and an intervening ridge, ending at the citadel. Still standing are stone collars about four feet (1.2 metres) apart with holes approximately 12 inches (300 millimetres) in diameter that indicate the course of the line. The nature of the pipe material is unknown, but the fact that pressures up to 300 pounds per square inch (20 kilograms per square centimetre) had to be accommodated in the lower section seems to rule out clay or lead pipe. Speculation indicates bronze or possibly reinforced wooden pipes. In Roman times the siphon was replaced by an elevated system, less subject to leakage and more reliable.

Introduction of the siphon

IN ROME

Writing in AD 97, Sextus Julius Frontinus, water commissioner of Rome, exclaimed, "With such an array of indispensable structures carrying so many waters compare, if you will, the idle pyramids or the useless, though famous, works of the Greeks!" Today, nearly 2,000 years later, Roman aqueducts are still impressive largely because of their magnificent arches, which stand in many parts of Greece, Italy, France, Spain, North Africa, and Asia Minor. Not widely known is the fact that the course of these aqueducts was mainly underground. For instance, in the time of Frontinus the combined length of all aqueducts entering Rome added up to about 260 miles, of which only 30 miles, or about one-ninth, were on arches.

The Roman water system. In general Rome's water system, originating in the hills several miles from the city, was gravity fed, low-pressure, and designed primarily to serve urban needs rather than those of the farm. The water was lowered gradually some 800 feet in elevation through a series of free-flowing conduits, in large part underground, to elevated distribution tanks in the city. These tanks (247 in 1st century AD) had no great storage capacity, for the system worked on the principle of constant takeoff. Overflow was regularly used for sewer flushing. At the tanks the system was converted from one of gravitational flow in conduits to that of a low-pressure supply in pipes. In building the conduits, trenching was employed where possible, but tunnels, the more expensive channels on arches, and even siphons on rare occasions were used. Tunnels, which might be 50 feet or more underground, had shafts at intervals of 240 feet or less, to prevent air locks and to permit inspection and cleaning. The conduits and pipes, depending on time and place of construction, were made of a variety of materials. There is evidence of stone-built ducts, open ducts of masonry, fittings of bronze, and pipes of stone, terra-cotta, wood, leather, and lead. In the older aqueducts, such as Appia, the channel was lined with cut stone walls made of a friable gray tufa called *capellacio*. Tufa is porous rock formed as a deposit from springs or streams. A harder reddish brown tufa, cut stone, and varieties of local limestone appeared later. Still later, in place of expensive cut stone, a rough concrete was introduced, and various combinations of block and brick, mixed rubble, and reticulate (diamond-shaped stones) came into being for both channels and arches.

The channels varied in width from about a foot and a half to five feet and in height from three to ten feet. On the average, they measured three and one-half by six feet and were generally lined with two or three layers of fine cement. To keep out rain and debris and to hold down evaporation they were covered over with flat, vaulted, or gabled roofs. Very often the floors and the roofs of the channels were made of the more highly prized peperino or travertine stones.

Since the system operated by gravity, determination of the gradient was highly important. The Roman surveyors relied principally on a number of simple instruments: a plumb-bob level; the *groma*, which traced right angles; the *dioptra*, which measured horizontal angles and sighted on a levelling staff; the *chorobates*, a water level some 20 feet long; and the *libra*, a simple water level. Their surveying competence is indicated by the gradients they achieved; for example, the duct of La Brevenne in France had a minimum slope of 1 to 1,400. Despite this capability and the 1 to 200 minimum recommended by Vitruvius, there seems to be no continuity or consistency in the fall from one section to another in the aqueducts entering Rome. Aqua Marcia, for instance, had grades varying from over 1 to 1 to those of 1 to 300.

The major aqueducts of Rome. During a span of over 500 years, from 312 BC to AD 226, 11 major aqueducts were built across the plains of the Campagna around Rome to supply the city with water. In order of construction they were: Appia (312 BC), Anio Vetus (272 BC), Marcia (144 BC), Tepula (127 BC), Julia (33 BC), Virgo (19 BC), Alsietina (2 BC), Claudia (AD 52), Anio Novus (AD 52), Trajana (AD 109), and Alexandrina (*c.* AD 226). Here are brief notes on the 11:

Aqua Appia. Built under the direction of the censor Appius Claudius Caecus, this aqueduct in the time of Frontinus measured 11,190 Roman paces (52,320 feet) from its spring source on the Lucullan estates to its distribution point in the city. Over 99½ percent of the line ran underground.

Aqua Anio Vetus. Financed by booty from the war with Pyrrhus, this was the first aqueduct to draw water from the upper valley of the Anio some 43 miles away. The contract for the work was let in 272 BC by the censor Manius Curius Dentatus and completed three years later. Its turbid and unwholesome waters, which flowed almost entirely underground, were directed to gardens and meaner uses.

Aqua Marcia. Marcius Rex, the Praetor, was commissioned by the Senate in 144 BC to restore the old aqueduct and to investigate new water sources. Four years later he completed Marcia, the longest and the first high-level aqueduct in Rome. Its waters were the best, deriving from three spring sources in the region of the upper Anio, about 57 miles distant. Though much of the channel was underground, cut through virgin rock, nearly 7,463 Roman paces (about seven miles) were above ground. This work is considered among Rome's first notable experiments in arch construction.

Aqua Tepula. Named for the tepid character of its waters, which originated a few miles from the city at the foot of the Alban Hills, this aqueduct was brought in by

the censors Gnaeus Servilius Caepio and Lucius Cassius Longinus. No remains of the original channel have been identified, but a new structure built by Agrippa in 33 BC continued to draw from the original springs but brought their water into the city on the arches of Marcia.

Aqua Julia. Built by the eminent engineer, Marcus Vipsanius Agrippa, this aqueduct tapped a group of rich springs some two miles above the headwaters of Tepula. It may have been built as early as 40 BC or as late as 33 BC, the year Agrippa extensively repaired and restored all preceding aqueducts. Julia also came into the city on the arches of Marcia.

Aqua Virgo. Largely to supply the new great public baths in the Campus Martius, Agrippa brought in this supply from springs on the Lucullan estates a little over eight miles from the city. The circuitous route, however, extended the length of the aqueduct to nearly 14 miles. Frontinus claims this aqueduct was named in honour of a young girl who showed the springs to soldiers who were looking for a water source. Still in operation today, Aqua Virgo feeds the famed 18th-century fountain, the Fontana di Trevi, whose sculpture bears a representation of the legendary maiden.

Aqua Alsietina. In 2 BC Augustus built a huge basin across the Tiber measuring 1,200 by 1,800 feet. Named the Naumachia, it was designed to stage mock sea battles. To supply this basin, Augustus tapped the Alsietinian Lake and channelled the unwholesome water over 20 miles, almost entirely below ground.

Aqua Claudia. Caligula razed part of Aqua Virgo to make a site available for an amphitheatre. To compensate for this loss he directed the start of two new aqueducts in AD 38. Both were completed 14 years later, during the reign of Claudius. The first, Claudia, tapped a source close to Marcia but, with a length slightly over 40 miles, was considerably shorter than Marcia. Improved engineering practices involving long tunnels and high bridges accounted for this advance. In the Claudian period the aqueducts reached their full development in height and magnificence. One of Claudia's bridges in the hills reached a height of 130 feet. Originally over 1,000 stately arches—some rising to a height of 90 feet—crossed the Campagna, and even today some 350 of these *opus arcuatum* are still in evidence.

Aqua Anio Novus. The companion to Claudia, this aqueduct originally drew its waters from the Anio River and the Rivus Herculaneus. Initially the quality of its water was poor, but after having its intakes shifted during the time of Trajan, it ranked just after Marcia in quality and quantity. Its original length is unknown, but during the time of Trajan it extended 54 miles. In general it was the highest of the aqueducts, achieving a maximum height above the ground of 110 feet.

Aqua Trajana. To serve the industrial quarter Trajan had this aqueduct built on the right bank of the Tiber, following nearly the same route as Alsietina. Five springs near the lake of Bracciano supplied this aqueduct, which ran no less than 35 miles, mostly below the surface, though in some places its height above ground exceeded 108 feet.

Aqua Alexandrina. This aqueduct was built to supply the luxurious bath in the Campus Martius built by the Emperor Alexander Severus. Its water came from springs on the eastern edge of the Patano some 14 miles from Rome. Much of the channel was carried on arches that reached a maximum height of 68 feet.

Seven of these aqueducts took their waters from springs, two from lakes, and two from rivers. They entered the city at various heights from about 238 feet above sea level down to 55 feet in descending order: Trajana, Anio Novus, Claudia, Julia, Tepula, Marcia, Anio Vetus, Alexandrina, Virgo, Appia, and Alsietina. Their lengths and quantities of water delivered varied, of course, with their state of repair and restoration, but the usual statistics given on them date to the end of the 1st century AD (Trajana and Alexandrina excluded).

Augmenting substantial archaeological evidence, much of the information comes from the treatise *The Two Books on the Water Supply of the City of Rome* by Sextus Frontinus (*c.* AD 35–104). After serving in many important posts, (including governor of Britain) he was appointed curator aquarum, water commissioner, by Emperor Nerva in AD 97. His treatise was a brief history of public water supplies in Rome, a detailed account of the system then in existence (nine aqueducts) and a commentary on his stewardship. It was begun shortly after he took office and was completed under Nerva's successor, Trajan. According to Frontinus, slightly less than one-third of all water delivered by the nine aqueducts was distributed outside the city, about 58% for private parties, and 42% for the emperor's disposal. Within the city the distribution approximated 44% to public uses (75 public buildings, 39 ornamental fountains, 591 water basins, and the legions' barracks), 37% to private parties, and 17% to the Emperor. To measure water flow, Frontinus used the *quinaria*, whose introduction he attributed to either Agrippa (63–12 BC), the first permanent water commissioner of Rome, or to Vitruvius (1st century BC), engineer and author of the classic *Ten Books on Architecture*. *Quinaria* was the name given to a standard pipe, five quarter digits in diameter (0.728 inch). It was also the term for a volume of water flowing through this standard orifice under "normal velocity." Since the velocity obviously varies, there is no way to determine how much water flowed, despite the large number of statistics generated by Frontinus. His American translator, hydraulic engineer Clemens Herschel, calculates that with all aqueducts running a daily total of 84,000,000 gallons was carried but that with diversions, leakages, outages, and theft perhaps only 38,000,000 gallons reached the inhabitants in the city. Based on an estimated population of 1,000,000, this indicates that the Roman engineers supplied 38 gallons (144 litres) per capita per day, a figure not uncommon in European cities of the 20th century.

Sources of information about Roman aqueducts

Elements in the construction of the Roman system. Selecting a good source of water was, of course, of prime importance. Vitruvius observed that water was probably good if (1) the general health of the people living near the source was good; (2) the water could be sprinkled into a Corinthian vase or other good bronze without leaving a spot; (3) the water could be boiled and then poured off without leaving sand or mud; (4) green vegetables could cook in it quickly; (5) the water was clear and moss and reeds were absent when it was flowing. He did not favour lead pipe because of the dangers of lead poisoning. Waters flowing into Rome were kept separated when there was a great difference in purity, so that Marcia was used primarily for drinking water, while Anio Vetus was used for such purposes as washing clothes.

Selecting water sources

There is little evidence that Rome's water was treated, though settling tanks were usually placed near the middle and at the ends of aqueducts, and *piscina* performed a kind of filtering operation by catching pebbles to keep them from clogging the pipes. Private treatment was no doubt used on occasion, for Pliny wrote that, "It was the Emperor Nero's invention to boil water and then enclose it in glass vessels and cool it in snow. . . . Indeed, it is generally admitted that all water is more wholesome when it has been boiled."

In the time of Frontinus five of the aqueducts had sufficient head to supply any part of the city, but illegal tapping sometimes disrupted the system. In the words of the water commissioner himself, "The cause of this is the fraud of the water men, whom we have detected diverting water from the public conduits for private use; but a large number of proprietors of land also, whose fields border on the aqueducts, tap the conduits; whence it comes that the public water courses are actually brought to a standstill by private citizens, yea for the watering of their gardens."

Frontinus indicates that the organization for maintaining the aqueducts shifted at the time of Agrippa. In the Republican period, censors and aediles were in general charge. They let the maintenance work out to contractors and then inspected the work. As the first water commissioner, Agrippa employed his own workmen—about

240 slaves. Later, when the Emperor Claudius brought his aqueduct into the city, he employed a gang of 460. Both gangs had foremen, masons, linewalkers, reservoir keepers, and other classes of workers.

Repairs, said Frontinus, were necessitated by age, weather, lawlessness of the owners of property through which the aqueducts passed, and poor workmanship in original construction—particularly in the later aqueducts. Tree roots were also a source of trouble, bursting top coverings and sides. Of particular concern were the arches, the hillside sectors, and the concrete lineage of the channels, for injury to the latter resulted in leaks that damaged the sidewalls of the channel and the substructure.

Throughout the Roman period the aqueducts tended to be under almost continual repair or restoration. There are records of restorations taking place under Diocletian and Constantine early in the 4th century and under Honorius and Arcadius very late in the same century. Much of the system was destroyed when the Goths besieged Rome in AD 537, but Belisarius, Justinian's general who reconquered much of Italy, effected extensive repairs shortly thereafter. Two centuries later Pope Adrian I and after him successive popes into the 9th century and possibly beyond attempted partial restoration, but the whole system continued to decay and fell into disuse during the Middle Ages, when Rome declined to a fraction of its ancient population. In 1870 an English company tapped the sources of the Marcian aqueduct above Tivoli and brought in a supply in modern pipe. Three others of the old aqueducts have been restored for modern use: Acqua Vergine (Virgo) restored in 1570; Acqua Felice (part of Alexandrina), restored by Sixtus V in 1585, and Acqua Paola (Trajana) restored by Paul V in 1611.

Aqueduct restoration

Aqueducts elsewhere in the empire. Throughout what was once the Roman Empire there remain evidences of the same engineering genius that created the magnificent water supply for the mother city. It is likely that the remains of more than 200 of these old Roman aqueducts—many with arches more striking than those around Rome—are extant. Prime among them is Agrippa's Pont du Gard (*c.* 19 BC) near Remoulins, France, which was damaged by the barbarians in the 5th century AD but repaired in 1743. The structure itself, which rises to 160 feet above the valley floor, is only a small section of 25½ miles of conduit, mostly underground, which was used to supply the city Nîmes from the Eure and Airon Rivers. Three

Archivo Mas, Barcelona

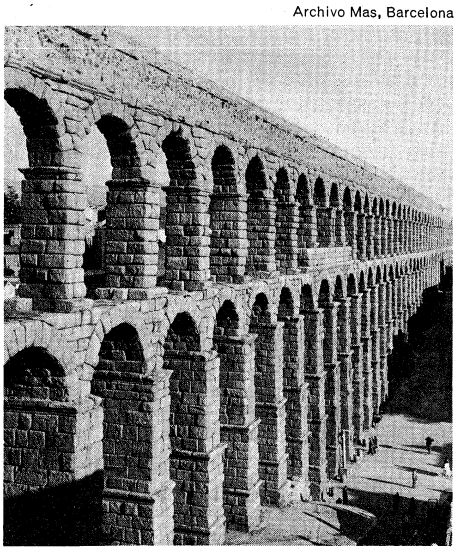


Figure 1: Roman aqueduct at Segovia, Spain, probably built under Trajan (*c.* AD 100–110). Still in use, the 2,700-foot structure carries water from the Río Frio to the city of Segovia.

tiers of arches support the water channel. Another spectacular example is the section of the Spanish aqueduct at Segovia (see Figure 1), built *c.* 100–110 during the reign of Trajan. Standing on slender piers of cut stone without lime or cement, the 2,700-foot structure dominates the city, rising some 119 feet (36 metres) above the streets on two tiers of arches. Near Carthage the aqueduct of Zaghouan, erected in Hadrian's time, has had a long and exciting existence, being wrecked by the Vandals, restored by the Byzantines, demolished by the Spaniards, and used in part again by the Bey of Algiers in the 19th century. Significant use of lead-pipe siphons is evident in Lyons, another site of some first-class aqueducts. These include Agrippa's Mont d' Or; the Craponne, probably built about the same time; the Brévenne, 1st century AD in the time of Claudius; and the Gier, later in the 1st century under Hadrian. Examples seem endless. Suffice it to say that long after the Roman Empire had died, her aqueducts remained to be used and admired and to serve as an inspiration for builders of subsequent ages. For the most part, the engineering works of the Romans were not surpassed until the 19th century.

IN THE MEDIEVAL WORLD

As the Roman Empire gradually fell apart, the strong central authority it represented, with its taxing and public works funding capacity, came to a halt. There were exceptions in a number of areas, of course. In 6th-century Byzantium, the surviving eastern portion of the Roman Empire, Justinian is credited with restoring and improving upon the water system built under Constantine. Outstanding features of the system included an imposing arched conduit and two great underground reservoirs. One of them, known as the Hall of a Thousand and One Columns (AD 528), was two stories high and measured 60 by 70 yards. The other, which still holds the city's water supply, has 365 marble columns in 28 rows, each column standing over 40 feet high.

In the Arab world there are a number of substantial examples, among them the 9th-century, arched conduit built by Ibn Kātib al-Faighanī to bring water from the southern desert to the new city of Cairo.

Arab aqueducts

Though remnants of Roman works remained in some of the larger urban centres in western Europe to be patched and used for centuries, most people were forced to return to wells and other local sources.

Clerics took the lead in matters material as well as spiritual. The earliest plans for a monastery water supply extant are those of the Benedictine Priory of Christ Church, Canterbury, carried out by Prior Wibert about 1153. Wells already existing in the cloister infirmary and outer cemetery were used for reserve, and the main supply was brought from springs three-quarters of a mile from the monastery, to a circular conduit house, thence through a perforated plate that trapped large impurities, and by lead pipe through five settling tanks. It then crossed a moat on a bridge, penetrated the city wall, and entered the monastery, where it was conducted underground to various points such as lavatory basins. This type of work by the churchmen was widespread. Cistercian abbeys had aqueducts in England, France, Germany, and Italy. Black Friars, Carmelites, Carthusians, Dominicans, Knights of St. John, and many other orders all developed respectable water systems.

Throughout medieval Europe, water supplies developed by the churchmen for their religious communities were expanded to serve the townspeople and ultimately passed to the control of the municipal authorities. One such example was that at Southampton, where in 1290 the Franciscan friars received a license to enclose the fountain of Colwell. Twenty years later they gave the use of their surplus waters to the town. By the 15th century the pipes were in disrepair, but the friars were too poor to maintain them, so the town took over the system. The dissolution of the monasteries in England in the 16th century and the decline of their influence elsewhere accelerated the transition to lay control, but the burgeoning of trade and the expansion of towns stimulated secular systems even before that time. Dublin had a water supply brought in at the expense of the town in the middle of the 13th century; in this case the town gave water to the monastery. Bruges in the 14th century was a flourishing city

Dublin's water supply

of some 40,000, served by a complete system of underground conduits to public cisterns and fountains. Built in the latter part of the preceding century, the system originated with a collecting reservoir near St. Bavon. From there the water was conveyed to the Bruges *waterhuis* where it was raised by an ancient pumping device, a man or horsedriven wheel with buckets on a chain, and distributed by gravity. Bruges was a wonder in its day, but its water system could hardly be compared to that of ancient Rome. Despite the proliferation and growth of towns throughout Europe, the art of supplying water made no significant advance until the forcer pump was introduced in the late 15th and early 16th centuries.

Originating in southern Germany and installed all over Europe by travelling specialists, this pump apparently was first developed to remove water from mines. Driven by river currents, large waterwheels—installed under bridge arches—were connected to wood forcers, covered with leather, which worked in barrels of wood or bronze. Before the 16th century, wheels over 20 feet in diameter were in operation. The resulting system was usually not very reliable, for it provided an intermittent supply with varying heads, but it marked great progress in municipal engineering. Heretofore, main systems originated in an area elevated above the city, relying on gravity for their flow. Pumping devices such as the Archimedean screw and variations of the waterwheel were known in Greek and Roman times, but they were used only to lift water to cisterns to supply individual dwellings. Now, pumping machinery became a significant adjunct to municipal water systems. The shift from the gravity-flow to pressure systems was underway.

Development of the modern aqueduct

ACHIEVEMENTS IN LONDON

In London, the first forcer pump appeared in 1582 when Peter Morice, a Dutchman, installed an undershot waterwheel under one of the arches of London Bridge. He amazed the population by pumping Thames water over the steeple of St. Magnus the Martyr. He installed a second wheel under another arch, and soon five arches were being used to give London its first pumped supply. Lifted to a high square tower at the bridgehead, the water was distributed by gravity through lead pipes to the eastern end of the city. Morice's franchise, drawn up to last 500 years, remained in his family until 1701. The works existed until finally demolished along with the old bridge itself in 1822.

Piecemeal attempts to supplement Morice's system were made as London, toward the close of Elizabeth's reign, neared a population of 150,000. The first large, successful, communal effort was made by Sir Hugh Myddelton, citizen and goldsmith of London, who formed the New River Company early in the 17th century. Completed in 1613 after a construction period of five years, the New River was an open ditch that drew its waters principally from the Chadwell near the town of Ware and flowed some 38¾ miles to a reservoir in London. Yet it was more than a long ditch. With over 40 sluices, crossed by over 200 small bridges, the stream traversed several roads by elevated timber troughs lined with lead. One such, near Edmonton, was an overhead aqueduct 660 feet long and five feet deep. Another, near Islington, known as Myddelton's Boarded River, was 460 feet long and 17 feet high. These timber structures were replaced with clay embankments toward the end of the 18th century. At Islington the water is reported to have been raised by a "great engine," powered originally by sails and then by horses. Thirteen wells along the way augmented the supply which originally totalled 13,000,000 gallons (50,000,000 litres) per day and was subsequently doubled. In the city the company's distribution system ultimately included over 400 miles of wooden pipe. The New River Company remained in existence until 1902 when the Metropolitan Water Board took over, ending the era of the great private companies.

The New River Company

Following the lead of the New River Company, more private water companies came into existence in London, mainly in the 18th and 19th centuries. In the early 18th century a private supply to middle class houses was common, and some tenements were also served. By the middle of the 19th century practically every house had a cistern filled at stated times; but a constant 24-hour supply was not introduced until 1873, and 35 percent of the total supply was on an intermittent basis as late as 1891. In 1867 nine companies, drawing their water from wells in the Chalk, Hertfordshire, the Lea, and the Thames, supplied London with 108,000,000 gallons daily.

ACHIEVEMENTS IN PARIS

The first significant step toward a Paris municipal water supply came sometime before 1200 when the monks of St. Laurent and Saint-Martin-des-Champs built aqueducts to carry water in conduits of lead and terra-cotta to the populace and the monasteries. These sufficed as the city's water source, along with private wells and the river, until the 17th century. Henry IV engaged Jean Lintlaer, a Flemish engineer, to install at Pont Neuf a pumping system similar to that at London Bridge. Lintlaer's machine, which continued in operation for two centuries, included a large waterwheel and four pumps capable of delivering 120 gallons per minute, most of which was used to supply the gardens of the Louvre and Tuileries. About 1670 a similar machine with three pumps was erected at the Pont Notre-Dame. Largely in order to serve his gardens at Versailles, Louis XIV began two water supply projects near Paris, which were of great moment in their day. The first, a pumping plant at Marly, raised water over a ridge and into the aqueduct 525 feet above the Seine, with an ingenious combination of 14 undershot waterwheels, a series of reservoirs, and 253 pumps arranged in three lifts. The second, a gravity-fed aqueduct to carry water from the River Eure, was started with great ceremony by army troops, but the project was never finished. In the 19th century, however, long aqueducts were built to bring distant waters to Paris. Particularly outstanding were the Dhuys Aqueduct, transporting spring waters 81½ miles, and M. Belgrand's Vanne Aqueduct, carrying spring water from Villeneuve l'Archevêque, 106 miles southeast of Paris.

Projects of Louis XIV

OTHER NOTABLE AQUEDUCTS

Many examples of Roman influence persist, particularly in southern Europe. At Coutances, France, ruins of a Gothic aqueduct built *c.* 1277 may still be seen; at Limoges, France, a subterranean aqueduct dates from an earlier time. On Malta, the Order of St. John had a long Roman-arched aqueduct built from Notabile to Valletta in the period 1610–15. Lisbon is served by an aqueduct carried on pointed arches with a Roman design.

The Hispano-Roman influence spread to the New World in such structures as the Mexican aqueduct of Zempoala (1553–70), built under the direction of a Franciscan monk, and the arched structure at Rio de Janeiro (1750), later converted to carry streetcars.

In North America early waterworks developed similarly at about the same time as in Europe. The first public supply in Boston in 1652 was spring-fed water conveyed by gravity through wooden pipes to a reservoir. In 1796 the Aqueduct Corporation brought water into the city from Jamaica Pond in Roxbury 4½ miles away, and in 1848 a supply was brought by gravity through an oval brick conduit 18 miles from Lake Cochituate, developed by J.B. Jervis and E.S. Chesbrough to a reservoir in Brookline, Massachusetts.

Montreal pioneered waterworks in Canada. In 1800 a private company furnished spring water through wooden pipes from Mount Royal, and by 1815 water was pumped from the St. Lawrence to elevated reservoirs. Water from outside the city limits was finally brought in with the construction of the five-mile Canal Aqueduct in 1853–56. Built under the direction of Thomas C. Keefer, this aqueduct carried water from the river above the Lachine Rapids. This reaching out from the city limits to greater and greater distances became particularly pronounced in the 19th century when new materials and

The Canal Aqueduct in Montreal

technical advances made such extensions feasible and economical. The main progress came in pumping systems capable of generating greater pressures, in the development of materials for handling higher heads, and in construction techniques, particularly tunelling.

THE INTRODUCTION OF PUMPS

When the York Buildings Waterworks of London were established in 1691, a horse wheel was first employed to pump water from the Thames. Waterwheels, however, continued to provide the power for most of the large waterworks up to, and in some cases well into, the 19th century. From the 16th to the 18th century, small pumps were often made of lead with handles of iron; brass or wooden pumps were also seen. Practically all the larger pumps were wooden; a transition to iron in all sizes came about gradually during the 18th century. This was reflected in the first U.S. attempt to pump a public supply. In 1754 at Bethlehem, Pennsylvania, a Danish millwright, Hans C. Christiansen, forced spring water through log pipes to a wooden reservoir 400 feet away.

The use of steam followed closely behind the development of the cast-iron pump. Early workers were more concerned with removing water from mines than with pumping a public supply. Though intermittent attempts were made to apply the steam engine to water supply in the 18th century, no substantial results were achieved until early in the 19th century. At the York Buildings Waterworks the horse wheel was replaced by a Savery engine that, in turn, gave way about 1720 to a Newcomen engine, which was improved by John Smeaton. In 1787 an early Boulton and Watt engine was installed; such progressive London waterworks installations as Hull (1795), Marly (1803), and Chelsea (1810) followed.

From England steam pumping operations spread rapidly to other countries. In the U.S., the practice was introduced by Benjamin H. Latrobe, who installed pumps built by Nicholas J. Roose (1800) to raise the waters of the Schuylkill at Philadelphia. Thus in the 19th century the steam engine and cast-iron pipe ushered in the era of modern waterworks.

USE OF THE PIPE AND OTHER CONDUIT MATERIALS

Nearly all ancient aqueducts were free-flowing channels; modern engineering has supplemented these with large-diameter pressure conduits, in which the water flow is completely enclosed and entirely fills the conduit or tunnel. Thus, conduits eliminate much of the large-scale construction the Romans needed to carry their aqueducts across valleys and depressions. Instead, the water flows under pressure (either by siphon or pumping) and follows the depressions, elevations and other topographic features of the aqueduct route. Because water in such an enclosed conduit exerts pressure in all directions, much of the history of the modern aqueduct concerns the search for suitable materials.

Wood. Though wooden water mains were used occasionally by the Romans, their main period of development came in the 17th and 18th centuries. Typical were the mains of London's New River Company that, for the most part, were bored elm logs with the bark left on, varying in bore diameter from two to ten inches, in outer diameter from six to 13 inches, and in length from ten to 22 feet. The joints were of spigot-and-socket type, the smaller tapered end being coated with white lead and driven into the larger end that was reinforced by an iron band.

Wyckoff's wooden pipe

In 1855 A. Wyckoff of Elmira, New York, patented a wooden pipe banded with iron, steel, or bronze spiralled around the pipe. The outside was protected by an asphalt coating; the pipe could withstand pressures up to 172 pounds per square inch. At about the same time, continuous wood stave pipes appeared. By 1851 B.H. Hall made pipes in 12-foot lengths of pine planks. A fine 20th-century example was the mile long, 13.5-foot diameter continuous pipe built in 1913 by the Pacific Coast Pipe Company for the N.W. Electric Company of Portland, Oregon. Diameters up to 17 feet have been installed.

Cast iron. In the 16th century cast-iron cannons up to 3 tons were made; it is possible some water pipe of the same material was cast at about the same time. There are no records of its extensive use, however, until the next century, when over 15 miles of mains were laid to supply Versailles. In 1901 some of these pipes were taken up and found to be of good grey iron, clean on the inside, and only slightly rusted on the outside. All about 3 feet long, they were joined by bolted flanges.

Despite its early use by the French, cast-iron water pipe remained largely in the experimental stage throughout the 18th century. In 1745 the London Bridge Waterworks Company had over 54,000 yards of wooden pipe, 3,860 yards of lead, and only 1,800 yards of cast iron. In the following year the Chelsea Water Company of London laid a cast-iron flanged pipe but was forced to re-lay it in 1791 because of defective joints. Joining trouble prompted Thomas Simpson of the company to design the first ball-and-spigot, and lead joints in 1785. The ball-and-socket, flexible joint was devised some years later (1810) by James Watt, while he was laying a main on the bed of the Clyde for the Glasgow Water Company. During the second half of the 18th century other sporadic attempts were made to change over to cast iron, such as that in Edinburgh where a lead line was replaced with a cast-iron main in 1755. In Dublin iron pipes were contracted for in 1797, more wooden pipes were laid in 1806, and, finally, a complete changeover to cast iron was started in 1809. Thomas Telford used cast iron in his aqueduct, which carried the Ellesmere Canal over the Dee at Pontcysyllte, some 20 miles from Chester, England. One thousand feet long, Telford's structure was a cast-iron trough 20 feet wide and six feet deep on 20 spans resting on tall masonry pipes.

Telford's Ellesmere Canal

During the early 19th century the change-over to cast iron was particularly rapid in England and in France. The practice spread to the United States when, in 1817, the Watering Committee of Philadelphia imported from England some cast-iron pipes to replace the old bored logs. The first waterworks in Cincinnati, established in 1820, utilized the new material but hung on to the old as well. A force pump, operated by a horse or ox treadmill, pumped water through wooden pipes to a reservoir of oak timber, 160 feet above the river. From here the water was distributed in cast-iron mains, wood branches, and lead services. Nine years later Lynchburg, Virginia, installed cast-iron pipes in what was claimed to be the first modern high-pressure main. The use of cast iron was further extended in 1848 when Dr. Robert Angus Smith, an Englishman, patented a cheap, effective exterior coating for protection against corrosion. His mixture of gas-tar, pitch, linseed oil, and resin is basically unchanged. The coating with a hydraulic cement mortar, both inside and out, on sheet-iron water pipes was invented about the same time by Jonathan Ball and was used for domestic supplies in 1845 at Saratoga, New York. Cast iron continues in favour today, though a number of other materials are proving competitive as technology advances.

Wrought iron and steel. Most development of wrought-iron and steel pipes for water carrying has been American. The first large-scale use made of wrought-iron pipes was in 1856 in certain hydraulic mining operations in California. Between 1862 and 1892 some 72 miles of wrought-iron pipe were laid as part of the San Francisco waterworks. In addition, a wrought-iron submarine pipe 13,800 feet in length was laid across San Francisco Bay. About 1890 soft steel began to supersede wrought iron in aqueduct conduits, particularly inverted siphons, a notable example being New York's Catskill Aqueduct, constructed early in the 20th century, which included 14 siphons totalling six miles. These were rivetted steel pipes, encased in concrete and lined with portland cement mortar, up to 11 feet 3 inches in internal diameter. Larger steel pipes have since become commonplace, notably the 30-foot diameter pressure pipes at Hoover Dam. Continued advancement in welding techniques has contributed to the replacement of rivetted by welded

pipe and, in general, has greatly expanded the use of steel pipe. Among steel's advantages are its larger size limits than cast iron, its resistance to rupture, its lack of joints (which increases watertightness), its lightness and ease of handling. It is particularly adapted to long span, self-supporting structures.

Stone and concrete. Though terra-cotta and stone pipes were used in ancient waterworks, their inability to withstand high pressure has made them impractical for modern use. While steam pumping was still in its infancy, the Weymouth Water Company of Weymouth, England, experimented with stone pipes in 1797. The West Middlesex Water Company of London also tried them early in the 19th century, but results were unsatisfactory. A coarse glass, covered on the outside, was tried briefly in France, but was unsuccessful.

Reinforced-concrete construction came into favour for large inverted siphons early in the 20th century, notably two siphons in the province of Huesca, Spain, the Sosa Siphon and the Albelda Siphon (1909). The latter, larger one, was built to accommodate a head of 97 feet measuring 13.12 feet in diameter, 2,363 feet in length, and 7.87 inches in thickness. A part of the Los Angeles Aqueduct (1913), ten feet in internal diameter, was constructed for heads from 40–75 feet, and on the Catskill Aqueduct (1915) heads less than 50 feet were provided with reinforced concrete siphons, 7¾–17 feet in diameter.

Other materials. About 1915 cement-asbestos pipe began to appear commercially. Ease of handling, strength in tension and compression, and chemical inertness made it popular in water supply systems. Other pipe materials in use, less important in the large pipe field, include lead, copper, fibrous compounds, and plastics.

These materials have been largely associated with pressure piping; many have also been used in varying degrees in nonpressure sections of aqueducts—covered or open conduits, tunnels, and canals. In long unlined canals loss by seepage can run as high as 30 to 40 percent. The search for more effective canal linings has involved use of masonry, hand-placed concrete, machine-placed concrete, mortar and plaster, asphaltic concrete, membranes, rubber and plastic sheets, compacted soils, soil and chemical sealants. Portland cement concrete is favoured for canal linings, but its relatively high cost encourages the search for effective substitutes.

The search for canal linings

TUNNELLING TECHNIQUES

The long modern aqueduct has been made possible not only through the improvement in materials and pumping operations but also through new techniques in tunnelling. The first significant progress in rock tunnelling was the introduction of black powder on the Canal du Midi, in France, in the 17th century. Dynamite replaced black powder and mechanical drills replaced hand drills in the latter half of the 19th century, while early in the 20th century hammer drills and mechanical loaders appeared, followed by detachable and tungsten-carbide bits. Later full-face boring machines proved increasingly successful, even in hard rock. Where the estimated rate of linear advance in tunnelling through rock (diameters 15–30 feet) in the 17th century was only two feet per week, it increased to 100 feet by the end of the 19th century, and up to 800 feet after the advent of full-face boring.

Where tunnels are low pressure and run in hard, sound, unfissured rock they are often left unlined, but a smooth lining of concrete in conjunction with the supporting of unstable ground and sealing of fissured places by injection of cement grout is generally necessary. A notable pressure tunnel is the Eucumbene-Tumut Tunnel of Australia, part of the Snowy Mountains irrigation-water–hydroelectric project; 14 miles long, with a 21-foot finished inside diameter, the tunnel has a concrete lining 18 inches thick reinforced and supplemented, as ground conditions require, by heavy reinforcing steel rods, steel support beams and roof support anchor bolts, the surrounding rock being pressure grouted with cement.

Conduit shapes. Hydraulically, the most efficient cross-sectional shape is the semicircle running full. In a closed conduit, the most efficient shape is the circle running full, but it is less efficient than any reasonably shaped open channel. Most pipes are circular, but canals —due to construction restrictions—are generally trapezoidal. Concrete-lined tunnels are generally circular or horseshoe-shaped to permit equipment access. In modern work the rectangular shape, favoured by the Romans, appears rarely under specialized conditions.

Long modern aqueducts

Swelling populations, exhaustion of local supplies, improvements in materials and machines, and advances in construction techniques all caused municipalities in the 19th and 20th centuries to reach out for water with longer and longer aqueducts. Glasgow became the first city in Britain to bring water from a distant upland source. Built during the period 1855 to 1860, the aqueduct extended a distance of 25¾ miles from Loch Katrine to the reservoir at Mugdock, eight miles from Glasgow. Tunnels accounted for 13 miles, cut and cover for nine miles and iron pipes across valleys for 3¾ miles. In 1881 Liverpool began work on a 68-mile line from the Vyrnwy Valley; shortly thereafter Manchester developed an even greater system by going to Lake Thirlmere 95⅞ miles away. A significant early 20th-century work was Vienna's Second Emperor Franz Joseph High Springs Aqueduct planned by Francis Berger. Of its 113⅓ miles almost 44 miles were made by tunnelling. Even larger was the 30-inch Coolgardie pipeline (1903) that stretched 350 miles across Western Australia to supply the gold fields. Also notable is the Apulian Aqueduct in Italy, designed to bring 110,000,000 gallons of water a day from the Caposele Springs through a ridge in the Apenines by means of a 9½ mile tunnel, thence west and south to Taranto 152 miles away. Numerous gigantic projects have appeared in the Western United States. Outside of California one of the most significant is the Colorado-Big Thompson Project (1938 to 1959), which includes 13 reservoirs and regulating basins, 27 earth and rockfill dams and dikes, six power plants, three major pumping plants, 20 tunnels (including the 13.1-mile Alva B. Adams Tunnel through the continental divide), 14 canals, 16 major siphons, and eight penstocks—traversing terrain that ranged from less than one mile to over 2.7 miles above sea level.

Water for Australia's gold fields

IN NEW YORK CITY

New York City draws nearly all of its water from three main systems, the Croton, the Catskill, and the Delaware. Its first successful municipal supply dates back to the Manhattan Company formed in 1799. By 1836 the company had 25 miles of mains and supplied 2,000 houses; it was replaced about 1842 by the Croton system.

The Old Croton Aqueduct, ultimately extending 45½ miles from the dam to the Murray Hill Reservoir, was begun in 1837, began delivering water in 1842, and was completed in 1848 under the direction of J.B. Jervis. A distinctive feature was its 1,450-foot (440-metre) Roman-type masonry bridge that carried the water across in three cast-iron pipes. Additional water requirements were met by the New Croton Aqueduct (1885–93). Nearly the entire line of 31 miles was tunnelled. In combination with other city water supplies, Croton's safe yield is about 325,000,000 gallons per day.

Construction on the first phase of the Catskill watershed project by Jonas Waldo Smith was carried out from 1907 to 1917; the first water was delivered in 1915. The aqueduct extended 92 miles from the Ashokan Reservoir to the city, including 55 miles of cut and cover, 14 miles of concrete grade tunnels (24), 17 miles of pressure tunnels (seven), six miles of steel pipe siphons (14), and nearly two miles of flexible cast-iron pipe across the Narrows of New York Harbor. To keep the tunnel in solid rock, some remarkable depths were achieved, such as the rock tunnel siphon 14 feet in diameter, 1,114 feet below sea level, crossing under the Hudson near West Point. Under Manhattan itself City Tunnel No. 1 runs for 18 miles at depths from 200–752 feet below ground. The system

was extended to about 120 miles from the city with the completion of the Gilboa Dam and the placing in service of the Schoharie Reservoir in 1924. Its safe yield is considered to be 555,000,000 gallons per day.

The Delaware system

In common with the Catskill system, which draws its waters from the Schoharie and Esopus creeks, the Delaware system also draws from Catskill mountain streams —from Rondout Creek, the Neversink River, and the East Branch of the Delaware River. Both systems start at an elevation so high that they feed by gravity to the Kensico Reservoir at 355 feet above sea level, thence without pumping to the city where they arrive at a head some 160 feet above that of the Croton supply. Construction on the Delaware System began in 1937, was suspended during World War II (although placed in limited use in 1944), and resumed shortly thereafter. The first stage, the Delaware Aqueduct, is a circular pressure tunnel running deep in bedrock for its entire length of 85 miles from the Rondout Reservoir to the Hill View Reservoir. With its extension in City Tunnel No. 2 its total length is 105 miles (169 kilometres). In some places it runs to a depth of about 2,500 feet below the surface and reaches a maximum diameter of 19 feet six inches in its last 13.6 miles. Subsequent extensions included the six-mile Neversink Tunnel in 1953, the 25-mile East Delaware Tunnel in 1955, and the 44-mile West Delaware Tunnel from the Cannonsville Reservoir to the Rondout Reservoir in 1965. With this addition the city could count on a total water supply from all sources amounting to 1,810,000,000 gallons per day.

IN CALIFORNIA

Of all political areas in the world the state of California must rank as a leader in the conveyance of water. Fundamentally, its need to move water stems from 70 percent of its surface supply originating in the northern third of the state, and 77 percent of its demand lying in the southern two-thirds. Agriculture uses between 80 and 90 percent of the state's imported and controlled water; approximately 95 percent of the crops harvested received some water supplemental to rainfall.

The first project of significance in the state was the Los Angeles Aqueduct (1908–13), which brought waters from the Owens River Valley in the Sierras some 250 miles away. Designed to deliver over 280,000,000 gallons per day, the aqueduct proper consisted of a series of storage reservoirs and 23.7 miles of unlined canals, 37 miles of lined canals, 97.6 miles of covered conduit, 42.9 miles of tunnels (plus 8.8 miles of power tunnels), 12 miles of siphons, and two miles of penstocks. In the recent past it has averaged 330,000 acre-feet annually (one acre-foot equals a depth of one foot [300 millimetres] of water over an acre which equals 325,851 U.S. gallons or 1,233,446 litres). Following closely was San Francisco's aqueduct, which originates in the Hetch Hetchy Valley in Yosemite Park some 160 miles distant. The system was designed to carry 400,000,000 gallons per day. Of particular interest are two welded steel, inverted siphons across the San Joaquin Valley, 47½ miles long, ranging to 66 inches in diameter and operating under maximum heads of 500 feet.

To the south, Los Angeles, in company with other municipalities, formed the Metropolitan Water District of Southern California in 1928 to build the Colorado River Aqueduct. Construction was begun in 1932; statistics released by the district in 1963 included the following: total length of aqueduct system, 672 miles (1,081 kilometres), including 242 miles of main aqueduct. Capacity 1,212,000 acre-feet annually, or 1,000,000,000 gallons of water per day. The aqueduct system includes: five pumping stations which lift the water 1,617 feet over mountain barriers; 92 miles of tunnels; 63 miles of concrete-lined canals; 55 miles of concrete conduits; 29 miles of inverted siphons, totalling 144 in number; three miles taken up by reservoirs and pump delivery lines; 430 miles of distribution lines to service areas in the district, including 395 miles of large diameter pipe lines, 15 miles of concrete-lined canals and 20 miles of tunnels; nine reservoirs; 321 miles of high-voltage power lines from Hoover Dam to aqueduct pump lifts; softening and filtration plant.

In 1947 the state legislature authorized an $8,000,000 water resources study which took ten years to complete. The final phase of the investigation, published in 1957, was a comprehensive master plan for the control, conservation, distribution, and protection of water in the state entitled *The California Water Plan*. This plan addressed itself to two categories of development: local works in various parts of the state, and long-distance system to transport waters via an aqueduct of record length from the north coastal region and Sacramento River basin to areas of deficiency—largely in the south.

The California Water Plan

Implementation of the second portion of the plan got underway in 1960. The program (California State Water Project), scheduled to take 25 to 30 years to complete, is designed to yield 4,230,000 acre-feet of water annually. Its cost was estimated at $2,800,000,000. It extends some 600 miles from the Oroville Dam on the Feather River in the northern part of the state to Perris Dam south of San Bernardino. Interconnections at Perris with the Metropolitan Water District of Southern California make possible the transfer of northern waters all the way to the Mexican border. The major physical works of the project include: 20 dams, seven powerplants, 21 pumping plants, 483 miles of lined canal, 166 miles of pipelines, 21 miles of conveyance tunnels, and 28 miles of unlined channels and reservoirs.

Among the outstanding facilities that form part of the project is the Oroville Dam (1962–67), at 770 feet (235 metres) the highest in the United States. Another is the A.D. Edmonston Pumping Plant, the largest plant of its kind in the United States. Located northeast of Ventura on the California Aqueduct, it raises the waters to a point where they can cross the Tehachapi Mountain Range and flow into southern California. It is designed to raise the water in a single lift operation, 1,926 feet from the pump outlet to the penstock. The principal conveyance structure of the whole system, the California Aqueduct, runs from the delta east of San Francisco, south through the San Joaquin Valley, and over the summit of the Tehachapi Mountains, a distance of 293 miles. At this point it divides into an East and West Branch, the former terminating some 444 miles from the delta. Sizes of channels vary along the aqueduct—a typical section being that of the initial reach, the North San Joaquin Division. This is a concrete-lined canal, 40 feet wide at the base, with sides sloping outward. The average depth of the flow is 30 feet.

In summation, a description of the aqueduct empire of California is one of superlatives. Never in the history of the world has man moved such volumes of water such distances, and the end is not in sight.

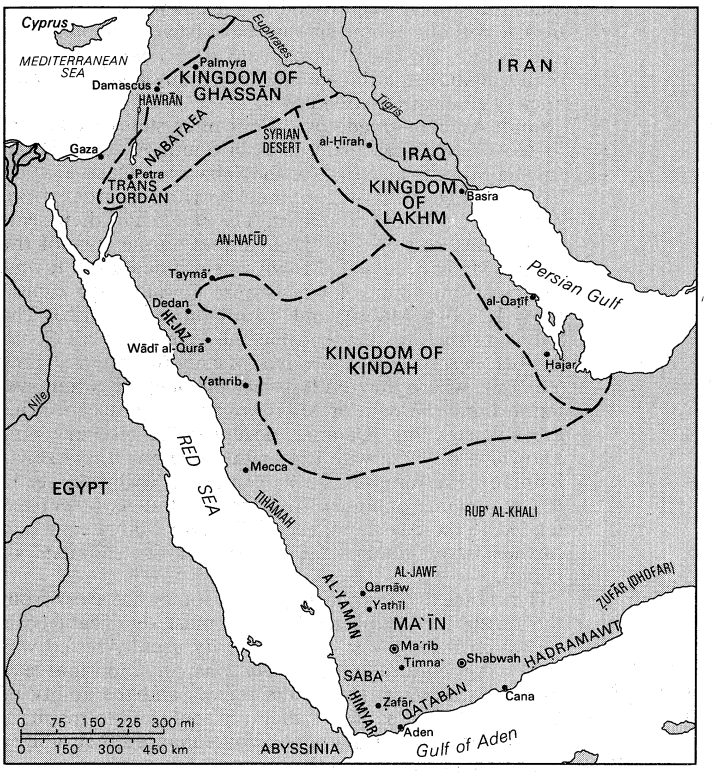
BIBLIOGRAPHY. T. ASHBY, *The Aqueducts of Ancient Rome* (1935), an exhaustive, detailed account of the Roman aqueducts based on modern archaeological findings, with many illustrations showing dimensions and an extensive bibliography; E. COOPER, *Aqueduct Empire* (1968), a history of the development of all major aqueduct systems in California; L.S. DECAMP, *The Ancient Engineers* (1963), a history of engineering works of major civilizations from earliest times to the Renaissance; J.K. FINCH, *The Story of Engineering* (1960), a history of civil engineering by eras from earliest times to the present; C. HERSCHEL, *Frontinus on the Water Supply of the City of Rome* (1899), a commentary on the translation of Frontinus, water commissioner of Rome in 98 AD, with a description of the aqueducts and their operation; C.J. MERDINGER, *Civil Engineering Through the Ages* (1963), a history of civil engineering by disciplines from earliest times to the present; F.W. ROBINS, *The Story of Water Supply* (1946), a history of water supply from earliest times to the present with emphasis on practices in England; E.B. VAN DEMAN, *The Building of the Roman Aqueducts* (1934), the most authoritative and extensive historical work in English on the ancient Roman aqueducts; E. WEGMANN, *The Water Supply of the City of New York, 1658–1895* (1896), descriptions of the planning and building of New York City aqueducts through the 19th century.

(C.J.M.)

Arabia, History of

Some time after the rise of Islām in the first quarter of the 7th century AD and the emergence of the Arabian Muslims as the founders of one of the great empires of history, the name "'Arab" came to be used by these Muslims themselves and by the nations with whom they came in contact to indicate all people of Arabian origin. The very name Arabia, or its Arabic name Jazīrat al-'Arab, has come to be used for the whole peninsula. But it is much less certain that all the people of that peninsula called it Jazīrat al-'Arab before the rise of Islām. Even in Islāmic sources the definition of Jazīrat al-'Arab is far from unanimously agreed upon; in its narrowest applications it indicates much less than the whole peninsula. The same can be said, perhaps with more certainty, about the use of the name Arab (*'arabī*) for every native of that same peninsula before and after Islām. It is, therefore, often more realistic to call them Arabians.

Arabia has been inhabited by innumerable tribal units, forever splitting or confederating; its history is a kaleidoscope of shifting allegiances, although certain broad patterns may be distinguished. A native system has evolved of moving from tribal anarchy to centralized government, frequently theocratic in type, and relapsing again into anarchy. The tribes have dominated the peninsula even when controlled by secular or theocratic arbiter-rulers. The inhabitants of Arabia, tribal or otherwise, have lived mostly in settled villages, dependent on agriculture; but nomadic or seminomadic camel- and sheep-owning pastoralists have occupied the dry steppe or desert, and fishing communities flourish along the coasts. The cities have the merchant and religious classes.



Principal sites in pre-Islāmic Arabia.
Adapted from C. Brockelmann, *History of the Islamic Peoples*

Character and achievements of Arabian culture

Arabian culture is a branch of Semitic civilization; because of this and because of the influences of sister Semitic cultures to which it has been subjected at certain epochs, it is sometimes difficult to determine what is specifically Arabian. Because a great trade route passed along its flanks, Arabia had contact along its borders with Egyptian, Greco-Roman, and Indo-Persian civilizations. The Turkish overlords of the Arabic-speaking countries affected Arabia relatively little, however, and the dominant culture of western Europe arrived late in the colonial era. Isolation preserved many parts of Arabia's interior from the changes of the centuries.

Arabia was the cradle of Islām and, through this faith, it influenced every Muslim people. Islām, essentially Arabian in nature, whatever superficial external influences may have affected it, is Arabia's outstanding contribution to world civilization. The Arabian ideal of religion inclines to strict adhesion to law. Islāmic law (Sharī'ah) is one of innate conservatism, and Islām itself does not readily accept change unless labelled a return to pristine purity.

The flexible and expressive Arabic language—the vehicle of Islām—is, along with it, the important cultural legacy of the peninsula and a language of which Arabs are inordinately proud, regarding Arabic verse as its loftiest secular achievement. Arabic not only developed a vast and important literature outside Arabia but it has affected every language it has contacted—Persian, Turkish, Urdu, Malay, Swahili, Hausa—just as Arabian Islām and Sharī'ah law have affected the intellectual and material cultures of Muslim countries. The holy book of Islām, the Qur'ān or Koran, is still widely studied, and the pilgrimage to Mecca remains a constant means for the renewal and diffusion of Islām. (See also HAMITO-SEMITIC LANGUAGES.)

On the material side, Arabia may boast of the magnificent ruins of pre-Islāmic architecture in the Yemen and also of many splendid medieval buildings there and in the Hejaz. Fine building still continues in western Arabia, which can point also to its great irrigation system, as can Oman. Arabians have also made significant advances in navigation, seafaring, and other activities.

PRE-ISLAMIC ARABIA, TO THE 7TH CENTURY AD

Prehistory and archaeology. At one time Arabia may have had a greater rainfall and richer vegetation than it does today; and the evidence of dried riverbeds, or wadis, running toward the sea in different directions, is an indication of a probable richer past. However, climatic conditions seem to have changed little since human habitation began, and life—settled or nomad—has had to cope with the harsh realities of this vast subcontinent.

Origins of the Arabian peoples

A well-known theory considers Arabia as the birthplace and homeland of the Semitic nations; but many scholars now prefer to see the movement the other way round. Ethnically, the people of Arabia have been related to different groups, whose homelands lie in almost all directions outside Arabia proper. The view that sought to see in all Arabians a single race has never been valid. Besides, non-Arabian ethnic elements in Arabia have included African races in the Tihāmah (the western coastal plain along the Red Sea), Aramaean stock in the north and east, Persians in the east, and Jewish communities in Arabian towns.

Arabia's known history shows that for most, if not all, of the time it could be divided into three main zones, which, in order of importance, were South Arabia, Northern and Central Arabia, and Eastern Arabia.

The western part of South Arabia, roughly corresponding to modern Yemen, is the most fertile part of Arabia. Irrigation works were skillfully and extensively constructed by the ancient South Arabians to utilize with a great degree of efficiency the water of the adequate rainfall brought by the monsoons of the Indian Ocean. There the highest degree of Arabian civilization and the most recognizable pattern of state and community organization developed. Knowledge of South Arabia derives mainly from the inscriptions found on slabs of stone, on the walls of public buildings and temples, on rock faces, on a variety of cult objects, and also on plaques of bronze. The vast majority of these inscriptions, however, has been found or located without proper archaeological excavations or surveys. The few and inconclusive excavations in South Arabia made after 1950 have yielded useful information and therefore serve as an indication of the rich archaeological possibilities that the land holds. The situation in other parts of Arabia is roughly the same. Inscriptions in a variety of languages, dialects, and scripts—Liḥyānī, Thamūdī, Safaitic, Aramaic, Greek,

and Latin—have been found in various parts of the peninsula. Although some surveys have been made in North and Eastern Arabia, the archaeology of these areas is still really unknown, although sites are many and links with ancient civilizations are known.

South Arabian kingdoms. Four main states are known to have established themselves in South Arabia—Ma'īn (of the Minaeans), Saba' (of the Sabaeans), Qatabān (Qitbān) and Ḥaḍramawt, in their territorial order from north to south and southeast. The states had their seats on the eastern slopes of the highland on the edge of the desert. Ma'īn occupied al-Jawf, where it had many towns, the most important of which were Qarnāw, the capital (now Khirbah Ma'īn), and Yathīl (now Barāqish). The Sabaeans, before they became the masters of all South Arabia, were centred around Ma'rib (in what is now Yemen), where they built a very efficient irrigation system around the dam of Ma'rib, whose magnificent ruins are still standing. Qatabān occupied the territory now known as Bayḥān, with its capital at Timna' (now Hajar Kuḥlān). Ḥaḍramawt had its capital at Shabwah, set in the mountains of upper Ḥaḍramawt; but its power extended to the coastland, especially over Mahra land and farther into Dhofar (Ẓufār), now in the sultanate of Oman.

The choice by these states of seats somewhat away from the fertile land indicates that the most significant basis on which their power and prosperity usually rested was trade. Agriculture undoubtedly was an important element of the prosperity of the region, and the ability of these states to provide for their residents and to maintain a sedentary core of the population enhanced their ability to control and carry out the trade.

Importance of trade

A main trading commodity was frankincense, a commodity then in great demand for ritual and other purposes in the lands of high civilization—in Egypt and Mesopotamia north of Arabia, and later across the Mediterranean in Greece and the Roman Empire. Most of the frankincense was grown or collected in South Arabia itself, particularly in the Mahra region and in Dhofar. This trade also carried the commodities brought by sea from India and East Africa to the ports of Cana (Ḥiṣn al-Ghurāb) and Aden.

Ma'īn. Ma'īn civilization seems to have been based entirely on trade activities. Autonomous Minaean trading colonies, whose task must have been to attend to trading affairs, are known to have existed in Dedan in northern Hejaz, in Timna', and in Shabwah. A group of Minaeans, indicated in inscriptions as *m'n m ṣrn*—taken by many scholars to refer to the Minaeans of the colony at Dedan—should most probably be identified as "Ma'īn of the caravan" or "Ma'īn of the expeditionary force," indicating the section of Ma'īn whose task was to lead or escort the caravans.

Everywhere the Minaeans identified themselves as belonging to "their tribe of Ma'īn," the real meaning of which must have indicated a group of people of equal citizen rights rather than a "tribe" based on kindred. Because of this, and also because it was based on a trading community, Ma'īn exhibited many features of government and public life typical of small democracies of the city-state pattern. The kings of Ma'īn were heads of state who ruled and promulgated edicts and orders in collaboration with a number of well-defined groups, most precisely enumerated in an inscription (obviously in order of standing and power after the kings themselves), as: "all the councillors of Ma'īn and the magistrates and the [assembly of] patricians and the class of [public] administrators and the whole tribe Ma'īn." The councillors formed a kind of cabinet, or council, normally mentioned immediately after the king or kings in the formula "the councillors of Ma'īn in the council of sovereignty [or, probably, government]."

The inscriptions make no mention of wars or military expeditions undertaken by the Minaean kings or state (although a Sabaean text mentions Sabaean attacks made within Ma'īn territory at a time when Ma'īn still stood as a state). This relatively peaceful existence of Ma'īn, coupled with its trading activities, suggests that it enjoyed a kind of covenant of safe-conduct with its neighbours and along the trade routes.

Saba'. The history of Saba' was rich and varied enough to explain to the outside world why Saba' often seemed synonymous with the whole of South Arabia. It is commonly held that the earlier form of government in Saba' was under *mukarrib*s, high priest-princes who combined religious and temporal power in the state and who peacefully gave way to kings. According to a less commonly held view, kings and *mukarrib*s could have existed and functioned simultaneously in the same state. It seems that *mukarrib*s functioned when the state was still based on a type of confederacy between a number of tribes; but when the rulers became autocratic kings, the office of *mukarrib* dropped. (*Mukarrib*s were also known in Qatabān and Ḥaḍramawt, but not in Ma'īn where everyone was always considered as belonging to the one and same tribe.)

For a long time Saba' was based in its capital, Ma'rib, on the trade route, and closely hemmed in, on both sides to the north and south, by Ma'īn and Qatabān, respectively. These two neighbouring states had a common interest in checking Saba', perhaps because of a common ancestry, attested by distinctive similarities in their dialects, or because of a close common stake in the caravan trade. After a long and occasionally militant history, Saba' first benefitted, and later suffered, from the emergence of the state of Ḥimyar in part of the territory under the authority of Qatabān. Ḥimyar made its capital at Ẓafār in the highlands.

Emergence of Ḥimyar

The beginning of the Ḥimyarī state has been tentatively linked with the beginning of the Sabaean era, by which a number of inscriptions are dated. The beginning of this era has been placed sometime between 115 and 109 BC, and a chain of events seems to have followed upon the birth of the new state. One suggestion has been that the overthrow of Ma'īn took place about this time, because Saba' probably found the weakening of Qatabān an opportune time to attack Ma'īn unopposed by Qatabān.

Around this time also, whether by cause and effect or by coincidence, the straits of Bab el-Mandeb at the southern tip of Arabia were opened for direct traffic from Egypt to India. This weakened the position of the overland incense traffic, which almost vanished within a century. The South Arabian states must have suffered a great blow; but another group which must have suffered were the Bedouin—the *'arab* and *a'rāb* of the inscriptions—who took part in the overland traffic and sold or hired camels to the South Arabians for this purpose. These Bedouin, whose homelands must have been outside the traditional territory of South Arabia, then started to become restless. They became involved in the conflicts in South Arabia that raged in the first two or three centuries AD. They served as fighters, normally attached to the king of a given state, most probably in the capacity of mercenaries. They fought on camel and horseback, and it seems that their presence in the South Arabian fighting forces gave the horse importance in South Arabia.

The Romans sent an expedition against South Arabia in 25–24 BC under the command of Aelius Gallus, the prefect of Egypt; it ended in failure. Epigraphic evidence suggests that the South Arabians thought the expedition was part of the expansionist plans of the Nabataeans of northern Arabia, who tried to use the Romans for their own ends.

Somewhat later the Ḥimyarites occupied Ma'rib, and the conquering kings styled themselves "Kings of Saba' and Dhū Raydān." From that time on South Arabia was in constant turmoil, with wars between rival claimants and factions even involving the kings of Aksum and Abyssinia in northeast Africa. With the reign of Yāsir Yuhan'im and his sons, beginning late in the 3rd century AD, order re-emerged in South Arabia. The whole region, including the kingdoms of Qatabān and Ḥaḍramawt, fell to these kings, who styled themselves "Kings of Saba' and Dhū Raydān and Ḥaḍramawt and Yamanāt."

Qatabān. Next to Saba', Qatabān seems to have been the strongest state in South Arabia. Qatabān had both

*mukarrib*s and kings. In Qatabān the title *mukarrib* may have been used by the kings themselves when they ruled a vast territory in which they controlled many tribes; the title was abandoned when Qatabānī domains shrank because of the emergence of the state of Ḥimyar in a substantial part of that territory. Government in Qatabān was democratic in form and institutions, and the landed classes played an important role.

The economy of Qatabān rested on its share of the South Arabian trade and on an efficient system of irrigation and agriculture. Qatabān seems to have also drawn revenue from pastoral activities of other groups within its territory. The temple administration also owned land and seems to have controlled a sizable part of the economy, including the trade.

Ḥaḍramawt. Knowledge of Ḥaḍramawt is limited. The royal family there may have been related to that of Maʿīn, a relationship possibly reflected in the close interest both states had in the frankincense trade—Ḥaḍramawt as prime source and producer, Maʿīn as prime carrier.

Kings alone seem to have ruled in Ḥaḍramawt, although reference to a *mukarrib* has been found. The kings, at least in the later period, were invested in public ceremonies, which were attended occasionally by ambassadors from other states (in one case from India, Palmyra, and others).

Unified South Arabia. By the end of the 3rd century AD South Arabia found itself involved in the affairs of other parts of Arabia as a result of the fall of Palmyra in the north and of the attempt of the different powers of the area to replace the latter's authority and trading activities in Northern and Central Arabia. The South Arabians tried to extend their political and probably their military presence into those regions. Eastern Arabians invaded by sea substantial parts of the Persian domains and occupied them for about a decade, possibly with the backing of South Arabia. About 325 the Persians retaliated, throwing out the Arabian invaders and inflicting a devastating punitive action on eastern, Central, and northeastern Arabia; they destroyed settled life, agriculture, and irrigation and filled up most of the wells and water sources, giving the nomad elements the upper hand and removing more cultured elements to the outlying parts of Arabia and even to Persia. Tribes loyal to South Arabia fled there, and South Arabia suffered a setback. But by the early 5th century South Arabia had recovered and expanded to Central Arabia; and the title of their kings became further augmented by the addition of "and of their Arabs in the Coastal Plain [Tihāmah] and the Highland [Ṭawd]."

Attempts at expansion

Another significant change was that a monotheistic religion, with "the Lord of Heaven and Earth" as the sole god, had become the faith of the state. Formerly the religion of South Arabia had been basically stellar, with ʿAthtar holding the senior place as the patron of land and agriculture and, in Maʿīn, even of trade. Each state or tribe had had its own moon god under a national or local name. The temples had been centres of religious life, and the priests of the moon gods had normally provided oracle services. Pilgrimage had been performed to certain temples of the moon gods, with rituals similar in many details to those of the pre-Islāmic and Islāmic pilgrimage to Mecca. Minor deities of agriculture and irrigation had been known, especially in Saba', as well as house gods and genies. (See also ARABIAN RELIGIONS.)

The tribe Amir had practiced a semimonotheistic religion—that of Dhū-Samāwī (Lord of Heavens). They worshipped no other god, although they paid homage to the gods of other communities in which they occasionally found themselves. The god's name resembles that of the "Lord of Heaven and Earth" of the monotheistic religion that emerged in the 5th century, but it is not known whether the latter developed from it.

Early Arabian monotheism

Jewish or Christian identification with the monotheistic god has been suggested. Christianity was introduced formally into South Arabia in the second half of the 4th century through a mission sent from Byzantium; and Judaism could have been introduced there about the same time, if not earlier. Later, in the 6th century, the struggle between local Jews and Christians was to play a decisive role in the fate of South Arabia.

Central and Northern Arabia. Central and Northern Arabia were the homelands of the Arabs proper. The region bordered on all sides lands of ancient civilizations. Reference to places or groups of people in that part of Arabia can be gleaned from Egyptian sources, from the Old Testament, and particularly from Assyrian records of the 9th–7th centuries BC. The Assyrians were concerned with keeping open the trade routes through the lands of these Arabs as well as ensuring the control of the nomads living on the edges of Mesopotamian cultivation. The last king of Babylon, Nabonidus (Nabu-na'id; ruled 556–539 BC), spent ten years of his reign in Taymā', in northeastern Arabia, making it the seat of his power, and he subdued Northern Arabia as far south as Yathrib (Medina).

Thamūd. The Thamūd tribe, known from the Assyrian records, from classical authors, and above all from the Qur'ān, lived in Northern Arabia and particularly northern Hejaz, near the Red Sea. Thamūdic inscriptions and graffiti, in a script derived from that of South Arabia, have been found all over Arabia, with the greatest concentration in Northern and Central Arabia. The inscriptions show a similarity in language and a unity in religion that suggest the existence of a Thamūdī form of authority over Northern and Central Arabia. It is doubtful, however, whether Thamūd enjoyed full sovereignty there; stronger external empires or caravan cities—particularly those of the Nabataeans and the Palmyrenes, who had a big stake in the safety of the trade routes through these lands—would have kept them either under direct control or closely bound.

The Liḥyānites. Another Northern Arabian group was the Liḥyānites, whose main seat was in northern Hejaz near the coastland and whose main town was Dedan, where a Minaean colony existed for some time. Liḥyānite inscriptions, in the South Arabian script, are of fuller content than the Thamūdic graffiti and mention names of some of their kings. The Liḥyānites may have been the ruling branch of Thamūd, for some time at least, or they may have been Thamūdīs who, through long contact with the better civilized Minaeans at Dedan, had a more developed system of government. Liḥyān seems to have survived as a state into the early part of the 4th century AD, and after that some of its people seem to have seized power in Ḥīra (al-Ḥīrah, what is now Iraq) from 380 to 385. A tribe of Liḥyān survived into Islāmic times, with a homeland not far from that of the Liḥyānites of antiquity.

The Lakhmids. The Nabataeans and Palmyrenes, on the outskirts of Arabia, were Arabian powers whose authority extended to the regions of Northern and Central Arabia, through which their trade routes passed. The fall of Palmyra in AD 273 was followed by far-reaching developments and changes in Arabia. The immediate task of policing the Syrian Desert for both the Romans and the Persians fell to the Lakhmids of Ḥawrān and Transjordan, perhaps in collaboration with, or possibly as successors to, the Tanūkh, the tribe in charge of the Arabian outskirts of Mesopotamia, to whose king, according to tradition, the new Lakhmid prince was related. When the Persians under Shāpūr II retaliated about AD 325 against the Arabs of Eastern and Central Arabia the second Lakhmid prince was entrusted by the Persians with the control of all the Arabian tribes conquered and lands devastated; he was styled "the King of all Arabs" and was recognized by both Romans and Persians. His domains touched on the territory of South Arabia proper.

The adoption of Christianity by the Roman emperor Constantine I (early 4th century AD) and the subsequent involvement of the Arabs on the side of the pagan emperor Julian the Apostate against Persia (363) must have left the pagan Lakhmids in a very awkward position vis-à-vis the Christian emperors of Byzantium and the Persians alike. But after the death of Shāpūr II in 380, the Lakhmids moved their seat to al-Ḥīrah and thus came completely under Persian suzerainty.

Adoption of the Arabic language

The Lakhmids were the first to adopt Arabic as an official language, writing it in the Nabataean Aramaic script. They thus departed from the practice of the Nabataeans and the Palmyrenes, whose official language had been Aramaic, and from that of the Arab tribes of the Syrian and North Arabian deserts, whose script was derived from the South Arabian. The fully developed Arabic language of the 6th-century pre-Islāmic Arabic poetry was mainly the work of tribes under Lakhmid authority, and the Lakhmid court was the focal point of most of it.

The Romans removed the Lakhmids from control of the Syrian Desert and replaced them with a dynasty from Quḍāʿah—a great confederation of tribes of the Syrian Desert and northeastern Arabia. Power in western Arabia was in the hands of a group of the Azd tribe, noticeably in Mecca and Yathrib (later Medina), perhaps under a loose alliance with South Arabia and with the tacit support of the Byzantine Romans.

Late in the 5th century Quṣayy ibn Kilāb, probably a defecting chieftain from among the Arabs under the Lakhmids, was installed as prince of Mecca with the active help of Quḍāʿah and the Byzantines. He thus established the rule of Quraysh in Mecca and gave a fresh significance to its sanctuary. About the same time the Kindah tribe of west central Arabia, supported by South Arabia, wrested control of much of Central and Northern Arabia from the Lakhmids and was able for a short time to replace them in al-Ḥīrah itself. The Kindah received occasional support from the Byzantines; but the Lakhmids fought back, and the kingdom of Kindah fell in 528.

Abyssinians in South Arabia. The fall of the Kindah tribe was precipitated by the fall of their supporter, South Arabia, to the Abyssinians (Ethiopians) about 525. A Ḥimyarī king, converted to Judaism, had acceded to the throne of South Arabia about a decade earlier and had drastically punished and persecuted the Christian population of Najrān. The Abyssinians, then already Christians, attacked South Arabia with the physical and moral support of the Byzantines, overthrew the Jewish king, and installed a local vassal Christian king. When the Kindah fell, the Byzantines sought the help of this vassal king in reinstating the Kindah in power; but he could not help. Later Abrahah, the commander of the Abyssinian forces in South Arabia, seized the throne and styled himself with the full title of the kings of South Arabia; he later was granted recognition by all the outside and local powers, including Abyssinia. Abrahah led an active Central Arabian policy against the tribes in and around Mecca through his political and religious activities. Muslim traditions associate the birth of the Prophet with the year Abrahah marched against Mecca but was probably thwarted by a breakout of disease among his men.

Rule of Abrahah

Abrahah was followed in quick succession by the brief reigns of his two sons. The Ḥimyarites drove out the Abyssinians with the active help of the Persians, who, shortly after that, assumed direct rule in the Yemen and controlled the country until it fell to Islām just before Muḥammad's death.

The Ghassānids. In Northern Arabia the Salīḥ dynasty of Quḍāʿah, which controlled the Arabs under Byzantine authority, was replaced about the end of the 5th century by Ghassān, dominion of a tribe of Azd that had moved northward from western Arabia. Ghassān with the Kindah checked the influence of the Lakhmids and their Persian overlords. When the Kindah fell in 528, the Ghassānid king al-Ḥārith ibn Jabalah defeated the Lakhmid al-Mundhir, received recognition from the Byzantines and lived a long active life.

Ghassān ruled and effectively controlled the Syrian Desert and northwestern Arabia to the outskirts of Yathrib, and it occasionally interfered in the politics of both Yathrib and Mecca. The Ghassānids were not as great patrons of Arabic poetry as were the Lakhmids; and Syriac, rather than Arabic, was their official language. They were Christians and played an important part in the religious conflicts of the Byzantine Church.

About the end of the 6th century Ghassān broke up into small states, some of them falling under Persian control. The decline of Ghassānid power in northwestern Arabia gave the Jews there a chance to wrest power from the Azd tribes in Yathrib and in Wādī al-Qurā. The Lakhmids fell out of favour with the Persians at about the same time, probably because of their conversion to Christianity.

Wars between Byzantium and Persia early in the 7th century further weakened the importance of both Ghassān and the Lakhmids, whose political existence ended after Islāmic conquest of Syria and Iraq in the time of the first two caliphs; and the picture of the region changed.

The influence of Northern and Central Arabia's close relations with other nations and cultures is best shown in the considerable number of loanwords from Persian, Aramaic, Syriac, Greek, and Latin found in classical Arabic, many of which are well attested in the most authentic text of classical Arabic—the Qur'ān.

ARABIA SINCE THE 7TH CENTURY

Arabian and Islāmic expansion. In the 6th century, Quraysh—the noble and holy house of the Kinānah confederation of the Hejaz controlling the sacred enclave (*ḥaram*) of Mecca—contrived a chain of agreements with the northern and southern tribes that opened the highways of Arabia to commerce. Under Quraysh aegis, caravans moved freely from the Aden and Ḥaḍramawt coasts in the south to Mecca and thence northward to Byzantium or eastward to Iraq. Another agreement made trade with Aksum (in what is now Ethiopia) and the African coast secure, as was also the Arabian coastal sea route. Furthermore, members of the Quraysh house of ʿAbd Manāf concluded pacts with Byzantium, Persia, and rulers of the Yemen and Abyssinia, promoting commerce outside Arabia. The ʿAbd Manāf house could effect such agreements because of Quraysh's superior detached position with the tribes. Quraysh had some sanctity as lords of the Meccan temple (the Kaʿbah) and were themselves known as the People of Allāh or the Protected Neighbours of Allāh; the tribes on pilgrimage to Mecca were called the Guests of Allāh; Quraysh made a contribution toward entertaining these pilgrims.

Quraysh control of Mecca

In its *ḥaram*, Quraysh was secure from attack; it arbitrated in tribal disputes, attaining thereby at least a local pre-eminence and seemingly a kind of loose hegemony over many Arabian tribes. Temple privileges held by Quṣayy, who established the rule of Quraysh, passed to his posterity, the ʿAbd Manāf house of which exercised the important function of collecting the tax to feed the pilgrims. It appears that the Kaʿbah, through the additions of other cults, developed into a pantheon, the cult of other gods perhaps being linked with political agreements between Quraysh—worshippers of Allāh—and the tribes.

The life of Muḥammad. Muḥammad (*q.v.*) was born of the Hāshimite (Banū Hāshim) branch of the noble house of ʿAbd Manāf; though orphaned at an early age, and with little influence in consequence, he never lacked protection by his clan. Marriage to a wealthy widow improved his position, but he began to make his mark in Mecca through prophetic inspiration. He preached the oneness of Allāh—possibly reacting against the pantheon in favour of the Quraysh deity Allāh. Rejected by the Quraysh lords, he sought affiliation with other tribes, though unsuccessfully until he managed to negotiate a pact with the tribal chiefs, the *naqīb*s, of Medina, whereby he obtained their protection and became theocratic head and arbiter of the Medinan tribal confederation (*ummah*). Those Quraysh who joined him there were known as *muhājirūn* (refugees or emigrants), while his Medinan allies were called *anṣār* (supporters). The Muslim era dates from the *hijrah* (hegira)—Muḥammad's move to Medina, in AD 622.

Beginning of the Muslim era

Muḥammad's men attacked a Quraysh caravan, thus breaking the vital security system established by the ʿAbd Manāf house, and hostilities broke out against his Meccan kinsmen, who dubbed him "divider" (*mufarriq*). Within Medina two problems confronted him—the necessity to enforce his role as arbiter and to raise supplies

for his moves against Quraysh. He overcame internal opposition, removing in the process three Jewish tribes whose properties he distributed among his followers. Externally, his ascendant power was demonstrated following Quraysh's failure to overrun Medina when he declared it his own sacred enclave. He foiled Quraysh offensives and ultimately marched into Mecca. After taking Mecca he became lord of the two sacred enclaves (*al-ḥaramayn*), but although he broke the power of such Quraysh lords as Abū Sufyān of the Umayyad branch of ʿAbd Manāf, his policy thenceforth was to conciliate his Quraysh kinsmen.

The rise of Islām. After Muḥammad's entry into Mecca the tribes linked with Quraysh came to negotiate with him and to accept Islām; this meant little more than giving up their local deities, already proved ineffective, and worshipping Allāh alone. They had to pay the tax, but this was not so novel because the tribal chiefs had already been taxed to protect the Meccan *ḥaram* from tribes that respected neither it nor the holy months of truce. Tribesmen probably regarded this episode of Muḥammad's as a typical Arabian family power squabble, and they waited to join the winner. Doubtless they cared little for Islām—many tried to break away (the so-called apostasy)—on Muḥammad's death.

Islāmic expansionism

Islām, however, was destined for a world role. Under Muḥammad's successors the expansionist urge of the tribes, temporarily united around the nucleus of the two sacred enclaves, coincided with the weakness of Byzantium and Persia. Tribes summoned to the banners of Islām launched a career of conquest that promised to satisfy their greed for booty and lands. With families and flocks, they left the peninsula. Population movements of such magnitude affected Arabia; in Ḥaḍramawt they possibly caused neglect of irrigation works, resulting in erosion of fertile lands. In Oman, too, in eastern Arabia, when Arab tribes evicted the Persian ruling class, its complex irrigation system seems to have suffered severely. Many Omani Arabs about the mid-7th century left for Basra (in Iraq) and formed the influential Azd group there. Arabian Islām replaced Persian influence in the Bahrain district and al-Hasa in the northeast, and in the Yemen.

As the conquests far beyond Arabia poured loot into the Holy Cities (Mecca and Medina) they became wealthy centres of a sophisticated Arabian culture; Medina became a centre for Qur'ānic study, the evolution of Islāmic law, and historical record. Under the caliphs —Muḥammad's successors—Islām began to assume its characteristic shape; paradoxically, outside the cities it made for centuries little difference to Arabian life. Sharīʿah (Islāmic law), promoted often by the Prophet's own descendants, who were naturally interested in doing so, took hold in the urban centres; but outside them customary law persisted, sometimes diametrically opposed to Sharīʿah. In time the Hejaz and the Yemen came to make notable contributions to Islāmic culture, but Islām's basically Arabian nature can be seen in the early mosque, which was like the pre-Islāmic temple and in the pilgrimage rites little altered from paganism.

Struggle for leadership. In Arabia offices were generally hereditary and elective; but on Muḥammad's death, Abū Bakr, the first caliph, abetted by his own eventual successor, ʿUmar, seized the leadership that Quraysh might have lost to others. They were not of the house of Hāshim, which, from the outset, felt cheated of its rights. ʿAlī, Muḥammad's cousin and son-in-law, became the focus of legitimist claims to succeed the Prophet. ʿUthmān, however, the third caliph, was descended from both the Umayyah and Hāshim branches of ʿAbd Manāf. The latter half of ʿUthmān's reign coincided with a slackening in the tide of conquest and plunder. ʿUthmān was censured for diverting property, revenues, and booty in Iraq and Egypt to his Quraysh relatives, though Muḥammad had set precedents for so doing. Squabbles with the tribes resulted in ʿUthmān's murder at Medina by tribal malcontents from Egypt. ʿAlī was now proclaimed caliph by the *anṣār*, but he lost the political battle with ʿUthmān's powerful relative Muʿāwiyah, governor of Syria, who demanded retaliation against the murderers. ʿAlī was later murdered by a Khārijite, a member of a dissident group that had forced mediation on a reluctant ʿAlī and willing Muʿāwiyah. ʿAlī had quitted Medina for Iraq, and the political power centre of Islām left the peninsula, never to return. ʿAlī's posterity, however, played an important part in subsequent Arabian history. (See also ISLAM, HISTORY OF.)

Departure of the caliphate

The Umayyad and ʿAbbāsid periods. Once Muʿāwiyah and the Umayyads had seized overlordship of the far-flung Islāmic empire, which they ruled from Damascus, the Holy Cities remained only the spiritual capitals of Islām. The Umayyad caliphs appointed governors over the Hejaz, Yemen, and Oman, whose degree of authority was variable; but in Iraq occasional powerful governors, such as al-Ḥajjāj ibn Yūsuf and Yazīd ibn al-Muhallab, managed to control the Persian Gulf provinces, the Gulf being an important maritime trade route, especially under the ʿAbbāsids. The brief unity Islām had imposed on the Arabians, however, was now irrevocably broken; from primitive power struggles, the main Islāmic sects took shape—the "Orthodox" Sunnīs and the "Legitimist" Shīʿah (who were distinguished from the Sunnīs principally by their tenet that the *imām* of the Muslim community must be descended from ʿAlī by Muḥammad's daughter Fāṭimah). (See also FATIMIDS.)

Umayyad forces defeated a Quraysh pretender, ʿAbd Allāh ibn az-Zubayr, who had been proclaimed caliph in the Hejaz. Medina was captured; Mecca was besieged, the *ḥaram* bombarded with missiles, and the Kaʿbah set on fire (the sacred Black Stone was split in three places). The harsh Umayyad general Ḥajjāj captured the city, and the pretender perished. The violation of the sacred enclaves by troops, including Arab Christians, was an act of sacrilege, but it broke any power remaining with the tribal "supporters" in Medina.

The Prophet's original simple mosque in Medina, already enlarged by the early caliphs, was rebuilt by the Umayyad al-Walīd (it has been much altered and restored since). The Umayyads spent lavishly on the Holy Cities and developed Hejaz irrigation.

Preservation of the Holy Cities

The Umayyads collapsed before the ʿAbbāsids in 750, a fall to which rivalry between the tribes, nationally aligned as northern and southern Arabs, contributed materially. The ʿAbbāsids claimed adherence of the Legitimists, since their ancestor, the Prophet's uncle, was of the Hāshimite house. The ʿAbbāsids maintained a policy of strict adherence to religious observance, and they too devoted large sums to supporting and embellishing the Holy Cities, to which they sent annually a pilgrim caravan. Zubaydah, wife of the caliph Hārūn ar-Rashīd, celebrated for her public works, is said to have built the *qanāt*, a tunnelled conduit that still takes water to Mecca. The threat of insurrection by Legitimist pretenders of the ʿAlīd branch of the Hāshimite house—who denied ʿAbbāsid rights to the caliphate as they had with the Umayyads—was a constant danger to the ʿAbbāsid caliphs. The ʿAlīd family developed both Sunnī and Shīʿah branches; but the latter split into a multiplicity of sects, of which the most important were the "Twelvers" (Ithnā ʿAsharīyah, or Imāmīs), who recognized 12 *imāms*, and the "Seveners" (also called Ismāʿīlīs, after Imām Ismāʿīl ibn Jaʿfar), who acknowledged seven only.

The Yemen. To quell a rising in the Yemen, the ʿAbbāsid caliph Maʾmūn dispatched an Umayyad, Ibn Ziyād, who refounded in 820 the southern city of Zabīd and became overlord of Yemen, Najrān, and Ḥaḍramawt. About a century later, the Najāḥids—Abyssinian slaves or local Afro-Asians—supplanted the Ziyādids in Zabīd; but, though independent, neither dynasty renounced vague ʿAbbāsid suzerainty. The Banū Yaʿfur (or Yuʿfir; pre-Islāmic nobility), lords north of Ṣanʿā', expelled the Ziyādid governor and ruled independently from 861 to 997, recognized as governors by the ʿAbbāsids. Najāḥid rule ended when a certain ʿAlī ibn Mahdī captured Zabīd in 1159. (See also CALIPHATE, EMPIRE OF THE.)

The Qarmaṭians. In the late 9th century an ʿAlīd revolt in Yamāmah, in central Arabia, cut communications

Ismāʿīlī threat to the ʿAbbāsids

between Iraq and the Hejaz; but a more serious loss to ʿAbbāsid power in Arabia was occasioned by the appearance of Ismāʿīlī propaganda in the Yemen about 880, in eastern Arabia about 899, and even briefly in Oman. From the Yemen, Ismāʿīlī missionaries reached North Africa, where the Fāṭimid movement arose and conquered Egypt and for a time seriously threatened the ʿAbbāsids. The Qarmaṭians, an extremist offshoot of the Ismāʿīlīs, founded a state in al-Hasa, in northeast Arabia. Careful planners and plotters, the Qarmaṭians set out to subvert Orthodox Sunnī Islām. They were opposed to many of the teachings of Muḥammad the Prophet; they urged discarding of the prayer, the fast, and the pilgrimage, while encouraging wine drinking, sexual promiscuity, and music. In the Yemen the Qarmaṭian ʿAlī ibn al-Faḍl broke with the contemporary Fāṭimid missionary Manṣūr and in 912 entered Ṣanʿā'; but Qarmaṭianism was short-lived there.

In 1037 ʿAlī aṣ-Ṣulayḥī of the Yām tribe proclaimed the Fāṭimid caliph al-Mustanṣir but set up a dynasty in Ṣanʿā'. The Ṣulayḥid dynasty ruled most of upper Yemen and warred with the pro-ʿAbbāsid Najāḥids. The Banū Maʿn (probably the ʿAwlaqīs of today) had controlled Aden, apparently without submitting even to the Ziyādids; the Ṣulayḥid expelled and replaced them with a section of the Yām tribe. From the latter stemmed the Shīʿah Zurayʿids, who ruled until the Ayyūbid Tūrān Shāh, brother of Saladin (founder of the Ayyūbid dynasty), invaded the Yemen in 1173–74 and took Aden from them.

In 930 the Persian Gulf Qarmaṭians plundered Mecca, carrying off the Black Stone to al-Hasa; they later returned it under Fāṭimid pressure. The Qarmaṭians were overthrown in 1077–78.

Oman. In the last decades of the 7th century the Ibaḍīyah, regarded as a moderate Khārijite sect, conquered southern Arabia, established a Kindite *imām* in Ḥaḍramawt, occupied Ṣanʿā', and took Mecca and Medina, before the Umayyads drove them back to Ḥaḍramawt. Oman had early become Khārijite; the first Ibāḍī *imām*, al-Julanda ibn Masʿūd, was elected at about the beginning of the ʿAbbāsid caliphate. After the Ibāḍī invasion of southern Arabia in 893, Oman wavered between independence and subjection to the ʿAbbāsids and their Būyid or Seljuq supporters of western Asia. By the 12th century the Seljuq hold had become precarious and local *imām*s existed. During periods when the Indian trade used the Persian Gulf, Omani ports flourished; but revenues diminished wherever trade was switched to the Red Sea. From the mid-12th century until 1406 the Nabhani dynasty controlled the interior of Oman, but Turkic Oğuz (Ghuzz), Persians, and others variously possessed the coastal flank of the mountains.

The Zaydīs and ʿAlawīs. In the Yemen, lasting movements were being shaped by the close of the 9th century; the *imām* al-Hādī, a theocratic arbiter-ruler of traditional type, founded the ʿAlīd Zaydī (Zaidi) dynasty in Ṣaʿdah of northern Yemen. About the mid-12th century a Zaydī *imām* extended his rule northward to Khaybar and Yanbuʿ (Yenbo) and southward to Zabīd.

In the mid-10th century a refugee from disturbances in Iraq, Aḥmad ibn ʿĪsā (known as al-Muhājir), arrived in Ḥaḍramawt, then under Ibāḍī domination, and founded the ʿAlawī Sayyid house, which was instrumental in spreading the Shāfiʿī school of Islāmic law abroad into India, Indonesia, and East Africa.

The Ayyūbids and Rasūlids. The Ayyūbids, when they invaded the Yemen in 1173, found it parcelled out between the Banū Mahdī in Zabīd and in the Tihāmah, the Aden Zurayʿids, the Zaydīs of Ṣaʿdah, the Banū Ḥātim at Ṣanʿā', and the Sulaymānīs at Ḥaraḍ. Ayyūbid objectives were probably part political, to find themselves a haven and destroy the Ismāʿīlīs, and to control the India trade route. They stayed in power until about 1229, generally controlling Aden, Ḥaḍramawt, the Tihāmah, and the districts south of Ṣanʿā'. They introduced an administrative centralization apparently adapted from Syro-Egyptian organization to Yemenite conditions.

With the Ayyūbids arrived the *amīr* ʿAlī ibn Rasūl, probably of Oğuz origin, whose descendants, at first Ayyūbid governors, grasped independence (*c.* 1229) during the reign of al-Malik al-Kāmil. The Rasūlid period is the most brilliant era of Islāmic history in the Yemen. These monarchs embellished their capital, Taʿizz, and other cities with fine buildings; several kings had a literary bent and, beside belles lettres, wrote treatises of some originality on such subjects as agriculture, medicine, and farriery. A fiscal survey still surviving provides an account of the trade through ash-Shiḥr, Aden, and the Tihāmah ports, with budgets for maintaining castles, troops, and hostages kept as surety of good tribal conduct. Aden was an important trade centre in a flourishing period of Arab and Jewish commercial enterprise. The Rasūlids kept the southern coast under loose control up to Dhofar, even holding Ḥaḍramawt to some extent and maintaining an antipirate squadron.

The Rasulid period

The sharīfs of the Holy Cities. At Mecca in the mid-10th century commenced the 1,000-year ascendancy of the ʿAlīd Sharīfian families. Mecca now became capital of the Hejaz, replacing Medina, the centre from which it had been ruled since the Prophet's days. The *sharīf*s, though at times subject to such foreign overlords as the rulers of Egypt and of other parts of Arabia, exercised virtual independence. Throughout the ʿAbbāsid–Fāṭimid struggle, however, the *sharīf*s took the opportunist line of supporting the side in ascendancy. When Saladin, after deposing the Fāṭimids in 1171, brought back orthodoxy, the *sharīf*s again recognized the ʿAbbāsids and Ayyūbids, and from being Zaydīs turned Shāfiʿī.

In 1181 the French crusader knight Reynaud de Châtillon, raided Arabia, robbing pilgrim caravans. He intended to attack Medina but, switching his plan, raided in 1182 the Red Sea ports as far south as Bab el-Mandeb; Saladin destroyed Reynaud's vessels and so ended the threat to Mecca.

By the early 13th century the Mūsāwī *sharīf*s had conquered the Hejaz, extending their power southward to Ḥalī; but when they sought support from Egypt, Syria, or the Yemen, the Rasūlids managed temporarily to dispute the overlordship of Mecca with the Egyptians.

After Baghdad fell to the Mongols in 1258, the pilgrim caravan from Iraq lost all political significance for the Hejaz. As Iraq declined, Egyptian importance mounted and the *sharīf*s became steadily more dependent on the Mamlūks of Egypt.

Mamlūk and Ottoman influence. Although the Yemeni Rasūlids sometimes disputed with the Mamlūks the overlordship of the Holy Cities, the Mamlūks generally prevailed. Egyptians and Meccans attacked al-Mujāhid the Rasūlid on a pilgrimage in 1350, and he was held prisoner in Egypt though released later.

The Mamlūks. During the 14th and especially the 15th centuries the Mamlūks became the dominant power, maintaining a political agent in the Hejaz and a body of cavalry in Mecca. Eventually they made or unmade the Sharīfian rulers, though the local Egyptian commander's policy sometimes ran counter to that of Cairo. From the mid-15th century the Mamlūks took charge of the customs at Jidda, Mecca's port, 50 miles to the east, allotting a portion of the revenue to the pasha of that port. Sharīf Muḥammad ibn Barakāt (ruled 1425–53), however, received one-quarter of the value of all wrecked ships, one-quarter of all gifts arriving from abroad for the Meccans, and one-tenth of all imported goods, including the cargo of the Indian ships destined for Jidda. About half his income was distributed among the leading Sharīfian families.

Mamlūk domination

By the mid-15th century the foundering of the Rasūlid dynasty in the Yemen made way for the Ṭāhirids; about the same time the Kathīrī tribe of southeast Arabia controlled Ḥaḍramawt, nominally at least, on behalf of the new dynasty.

The beginning of the 16th century witnessed Portuguese penetration of the Indian Ocean and the Red Sea. Though they failed to capture Aden, the Portuguese blockaded the Indian trade routes to Europe via the Persian Gulf and Red Sea, causing severe, lasting damage to the economy of Muslim Middle Eastern countries.

The Ottomans. To counter the Portuguese, the Mamlūks now found it imperative to occupy (*c.* 1516) the Yemen Tihāmah, using cannon and muskets, then new to South Arabia. In 1517, however, the Ottoman sultan Selim I conquered Egypt and proclaimed the Hejaz part of Ottoman dominions. Sharīf Barakāt II of Mecca sent his son to negotiate at the Ottoman court and was confirmed as lord of the Holy Cities and Jidda, subject to recognizing the Sublime Porte as overlord. Selim's successor, Süleyman I the Magnificent, at the zenith of Ottoman power, munificently subsidized the Holy Cities, devoting large sums to new building.

In the Yemen the Mamlūks of Zabīd and Ta'izz acknowledged Ottoman authority, and Turks took over naval operations against the Portuguese in the Red Sea and Indian Ocean. They seized Aden and forced the Yemenis into the mountains, capturing San'ā' and Shahārah; ultimately, however, the Yemenis drove them back into the Tihāmah. The Turks adopted Mocha (al-Mukhā) in southern Yemen as their base, and Aden declined in importance. After conquering Iraq in 1563, the Ottomans could operate in the Persian Gulf against the Portuguese, who had taken Hormuz and Muscat in 1507, Bahrain in 1521, and freely harried the Arabian coasts.

Contact with Europe

Coastal Arabia was by then coming into direct contact with the Christian European maritime nations, which had begun their commercial penetration of the Orient. The Dutch and English followed the Portuguese; then came the French and other nations. The Portuguese held Bahrain until 1602, Muscat until 1649. The Western nations traded with the Yemen through Mocha, whose coffee trade began in the 17th century; later the Dutch, British, French, and others opened factories there.

By 1635 the Zaydīs, supported by the northern tribes, had expelled the Ottomans from the Yemen, and the Zaydīs had their first great, if short-lived, expansion when their tribes moved into much of southern Arabia. The country is physically so broken that they could not maintain their supremacy, and local tribes drove out Zaydī garrisons by about the second decade of the 18th century.

In 1631 Turkish troops, dispatched to reinforce the hard-pressed Ottoman governor at Mocha, deserted in a body and, arriving by land at Mecca, sacked it; their drunkenness and their rape of members of noble families horrified Muslim sentiment. Security vanished, and the nomadic Bedouins again looted caravans freely until, with aid from Constantinople, Zayd, the new *sharīf* of Mecca, was able to exterminate the rebels. Through his influence with the tribes, Zayd was also able to thwart the attempts of the Turks to control the Hejaz. In the late 17th and 18th centuries, however, there was confusion and civil war in Mecca, with disputes among the sharifian tribes and struggles at Jidda with Turkish officials, who, notwithstanding the virtual independence of the *sharīf*s, still dabbled in Hejaz politics. A new element was introduced in Najd (in central Arabia) with the rise of the puritan Wahhābīyah, who—regarded by the *sharīf*s as dangerous heretics—were refused permission to make the pilgrimage from the 18th century. Though the prohibition was withdrawn in 1785, Sharīf Ghālib raided them in 1790 and fought them for the next 15 years.

Omani expansion. In Oman events took an independent course. The Ya'rubid dynasty—founded about 1624 when a member of the Ya'rub tribe (one of the al-Azd group) was elected *imām*—expelled the Portuguese from Muscat and set to harrying Portuguese possessions on the Indian coast. Embarking on expansion overseas—to Mombasa in 1698, then to Pemba, Zanzibar, and Kilwa—the Omanis became the supreme power on all the coastal regions of the Indian Ocean, and marauding Omani fleets were feared by European merchants.

The Persians captured Muscat in 1743. The Ya'rubids dissolving into dynastic dispute, a certain Aḥmad ibn Sa'īd set to liberating Oman from the Persians. He became *imām* in 1749, founding the Āl Bū Sa'īd dynasty. This period in Oman is marked by the crystallization of the political alignment of the tribes of the Banū Ghāfir (Ghāfirī) against those of the Banū Hinā (Hināwī).

During the 18th century the growth of the East India Company, and British paramountcy in India, was beginning to affect Arabian politics in the Indian Ocean coastal region, though the interior was little concerned. A series of agreements, including the peace treaty of 1820, which suppressed the Qawāsim pirates, paved the way to British control of the Gulf.

The Wahhābīyah. The Turks, clinging to the Hejaz for religious prestige and claiming to be custodians of the Holy Cities, had little power outside their garrisons in those cities and along the pilgrim route. The bribes they gave the Arabs for allowing the caravans to pass, and supplies to be carried to Mecca and Medina, however, prevented their expulsion. (See also OTTOMAN EMPIRE AND TURKEY, HISTORY OF.)

Religious reform. The Wahhābī movement, which introduced a new factor into the pattern of Arabian politics, was founded by Muḥammad ibn 'Abd al-Wahhāb, a reformer influenced by the writings of the 13th–14th-century pietist theologian Ibn Taymīyah, of the strict Ḥanbalī school of Islāmic law. Paganism still flourished in Arabia, in various disguises, and it was al-Wahhāb's intention to purify Islām of what he saw as polytheism and to return it to an idealized primitive state. Expelled from his home town in Najd, he moved to Dir'īyah, a village that had never been ruled by the Turks, and obtained the protection and the adherence of its chief, Muḥammad ibn Sa'ūd.

Resistance to the Ottomans. Propagating the doctrines of al-Wahhāb, Ibn Sa'ūd and his son mastered all Najd, from al-Qaṣīm to al-Kharj. Late in the 18th century the Wahhābīyah began raiding Iraq and in 1801, despite Turkish attempts to suppress them, sacked the holy Shī'ite city of Karbalā'.

The Wahhābīyah then besieged Mecca, aided by Sharīf Ghālib's brother-in-law, who defected to them. On Ghālib's withdrawal, the Wahhābīyah entered the city, where, though enforcing Wahhābī practice, they behaved in an orderly fashion. They failed before Jidda, however, and Ghālib re-entered Mecca that same year. In 1804 the Wahhābīyah took Medina and two years later retook Mecca. By then the Ottomans were so alarmed at the Sa'ūdī–Wahhābī peril that the Sultan urged Muḥammad 'Alī, viceroy of Egypt, to act as his deputy and drive the Wahhābīyah out of the Holy Cities. Egyptian troops invaded Arabia, and after a bitter seven-year struggle, with serious Egyptian reverses, the Viceroy's forces recaptured Mecca and Medina. Forced to surrender his capital, the Wahhābī leader, 'Abd Allāh ibn Sa'ūd, was sent to Constantinople and there beheaded. Foreign occupation of northern Arabia continued for some 20 years.

Extension of Sa'ūdī–Wahhābī dominance

The second phase of Sa'ūdī–Wahhābī dominance began when Turkī, of a collateral Sa'ūdī branch, revolted and in 1821 captured Riyadh in Najd and made it his capital. He was succeeded by his son Fayṣal. By 1833 Wahhābī overlordship was generally recognized in the Persian Gulf, although the Turks and Egyptians remained in the Hejaz. A new Egyptian push, at first successful in reaching the Gulf and deposing Fayṣal, collapsed with the fall of Muḥammad 'Alī's empire, and Fayṣal reassumed his rule.

After Fayṣal's death the fratricidal ambitions of his two eldest sons allowed Ibn Rashīd, ruler of Ḥā'il in Jabal Shammar to the north, to take Riyadh. Ibn Rashīd ruled northern Arabia until he died in 1897. Meanwhile, the Sa'ūdīs had lost the fertile al-Hasa to the Ottoman Turks, and the family ultimately took refuge in nearby Kuwait.

Ibn Rashīd's son and successor became involved in a struggle with the *shaykh* of Kuwait, which enabled the greatest of the Sa'ūdīs, Ibn Sa'ūd ('Abd al-'Azīz II), to retake Riyadh in 1902. By 1904, raiding and skirmishing, Ibn Sa'ūd had recovered all the Sa'ūdī territory, though he had to campaign against Ibn Rashīd, the Ottomans, and disaffected tribes and relations. In 1912, to bring the nomads under control, he introduced agricultural settlements (*hijrah*) colonized by Wahhābī warrior groups called Ikhwān, organized to fight and cultivate the land. In 1913 he drove the Ottomans from al-Hasa.

When World War I broke out, Kuwait renounced allegiance to Ottoman Turkey. Ibn Sa'ūd fought the pro-

Turkish Ibn Rashīd but otherwise remained inactive, though in contact with the British. (See also IBN SAʿŪD.)

The Hejaz. After Muḥammad ʿAlī Pasha of Egypt deposed Ghālib in 1831, the Meccan *sharīf*s were merely his nominees until 1840, when the Egyptians evacuated Arabia. Thereafter the *sharīf*s were independent until the Turks began to reassert authority by force, usurping the powers of the *sharīf*s. Improved communications as a consequence of the opening of the Suez Canal in 1869 allowed the Turks to send troops by sea to Arabia and to establish direct administration in the Hejaz. Ḥusayn ibn ʿAlī, appointed grand *sharīf* in 1908, travelled by the new Hejaz Railway from Constantinople to Medina.

The Yemen. In 1839 the British took Aden, ruling it and the island of Socotra (at the entrance to the Gulf of Aden) from India and thus forestalling the Egyptians, who had aimed at capturing it for themselves. In 1849 the Ottoman Turks occupied the Tihāmah, compelling the Zaydī *imām* to sign a treaty whereby he would govern the Zaydī highlands as a vassal of the Porte and admit a small Turkish garrison to Ṣanʿā'. The Turks, however, could not enter Ṣanʿā' until 1872. They were never able to break the resistance of the Zaydī tribes completely and were forced to an accommodation with the *imām* a few years before World War I.

British control of Aden

Aden developed into a large town and prosperous port, especially after the Suez Canal opened. Protectorate treaties concluded with the independent tribes around Aden eventually were extended to include the Ḥaḍramawt. Shortly before World War I an Anglo-Turkish commission largely settled differences over the Yemeni boundary with the Aden Protectorate.

The Gulf states. By the early 19th century the Qawāsim pirate tribes of the Gulf, conquered and largely inspired by the Wahhābīyah, had extended their depredations to India. In 1835 the pirate coastal chiefs were induced to bind themselves by a maritime truce to end hostilities by sea, a truce that was made permanent in 1853. In Oman, Sulṭān Ibn Aḥmad, revolting against his uncle the *imām* in 1793 and gaining mastery of the coastal towns, had made an alliance with the British to assist them against the French. Relations between the coastal sultans of Muscat and the British developed, and the British turned Zanzibar, in East Africa, into a protectorate in 1890. The extension of British influence over Bahrain culminated in 1900 with the opening of a British political agency. The Gulf states and Zanzibar were also persuaded by the British to help suppress the slave trade.

World War I. When Turkey entered World War I it held all of western Arabia and was supported in central northern Arabia by the Rashīdīs of Ḥāʾil. Ottoman attempts to extend the empire to eastern Arabia, however, had been countered by the British, who were paramount in the Gulf and in treaty relation with the Arab shaykhdoms there. Sharīf Ḥusayn of Mecca had been corresponding with the British in Egypt before the war; with assurance of British support, he came out against the Turks in June 1916, taking Mecca but failing to capture Medina. The British also supported the Idrīsī in Asir against the Turks. The modus vivendi established in the Yemen had left the *imām* semi-independent in the north, while the Turks administered the rest; Turkish forces entered the Aden Protectorate, routing a small, badly organized British force at Laḥij, but the war settled down to a gentlemanly stalemate just outside Aden colony.

It was the British intention to align Arab against Turk in eastern Arabia and also in southern Iraq. Two sons of Sharīf Ḥusayn of Mecca, Fayṣal and ʿAbd Allāh, stirred the tribes up against the Turks and, assisted by British supplies and liaison officers, moved northward to Transjordan along the right flank of the British armies and into Damascus (1918). Fayṣal set up an Arab government there, proclaiming himself king of Syria, only to be dislodged by the French in July 1920. In 1921 he was made king of Iraq, ʿAbd Allāh *amīr* of Transjordan.

Wahhābī–sharifian dispute. During the war relations between Sharīf Ḥusayn and Ibn Saʿūd worsened, and in 1919 the dispute broke into an open clash. The Wahhābīyah won so decisive a victory that they might have advanced unopposed into the Hejaz but for pressures on Ibn Saʿūd by the British, who were supporting both him and Ḥusayn with subsidies. Instead, Ibn Saʿūd concentrated his forces against Ibn Rashīd and mastered all Shammar territory, capturing Ḥāʾil in 1921.

Meanwhile, the grand *sharīf* refused the terms of a treaty with Britain, mainly because of the Balfour Declaration, which approved a national home in Palestine for the Jews. The Wahhābīyah marched into the Hejaz in 1924, and by October Ḥusayn was ruler no longer.

Saudi Arabia. Ibn Saʿūd's zealous Wahhābī followers, arriving in the relatively lax atmosphere of Hejaz society, were exposed for the first time to the world of Islām at large and to the modern world. Ibn Saʿūd managed the resulting problems with firmness and tact. He had furthermore to enforce his rule over the tribes impatient of centralized government. Ibn Saʿūd's power was based on support from Najd, and he took tough action with the tribes; this brought its reward, and he set out to develop security, economic reform, and communications.

On Ibn Saʿūd's southern border the Idrīsī *sayyids of Asir* had risen to power in the first decade of the 20th century. After the war the British handed them the Yemeni port of al-Ḥudaydah, to the intense annoyance of Imām Yaḥyā of the Yemen. When in 1926 and 1930 Ibn Saʿūd concluded agreements with the Idrīsī that rendered Asir a virtual dependency of Saudi Arabia, Yaḥyā took al-Ḥudaydah and southern Asir. Negotiations with the *imām* failing to settle territorial questions, Saudi troops swept into the Yemeni Tihāmah; they withdrew to the present border after the Treaty of aṭ-Ṭāʾif in 1934.

In the postwar years, Britain and Saudi Arabia concluded agreements defining the frontiers with the British mandates of Jordan and Iraq, and by treaty in 1927 Ibn Saʿūd was recognized as a sovereign, independent ruler.

The Yemen. Imām Yaḥyā had had virtually to conquer the Yemen, in the Zaydī interest, after the Turkish departure; by measures often stern, he established security almost everywhere. He refused to recognize the Anglo-Turkish boundary, however, and his penetration eastward into tribal territory brought clashes with the British, who, in the later 1930s, made faint stirrings toward pacification and development of the protectorates.

Postwar Arabia, to 1962. The post-World War I settlement and centralization of power in the hands of Yaḥyā, Ibn Saʿūd, and the British, gave Arabia a large measure of peace and security, which endured until 1962. An entirely new factor was introduced in the 1930s with the discovery of immense quantities of petroleum in the deserts. In Bahrain oil was struck in June 1932. The U.S.-owned Arabian Standard Oil Company (later Aramco) discovered oil in the Dhahran area of Saudi Arabia, and the first shipments left al-Khubar in September 1938. The Kuwait Oil Company, a joint Anglo-U.S. enterprise, prospected Kuwait, where production began in June 1946. Thereafter oil was discovered in many other places, mostly in the Persian Gulf. Vast revenues from petroleum brought enormous changes to Saudi Arabia and utterly transformed Kuwait. The market for labour brought migrants from poor regions, such as the Yemen and the Ḥaḍramawt, and created an unheard-of standard of living along with great social problems.

Discovery of oil

The Arab world as a whole began to ferment under an unending stream of nationalistic, revolutionary propaganda, first from Egypt, later from Syria and Iraq, skillfully utilizing resentment of Israel. The Arab states without oil envied the wealth produced by American and British companies for Saudi Arabia, Kuwait, and other states. In the late 1950s Egypt mounted a propaganda campaign aimed at undermining "feudal" Arab kingdoms and removing "colonialism" from Arabia, with an eye to acquiring for itself control of oil in vulnerable Arabian countries with small populations.

Arabia since 1962. In 1962 Egyptian-trained Yemeni officers engineered a coup d'état, forcing the *imām*, al-Badr, to leave Ṣanʿā', and Egyptian troops invaded the Yemen. Seeing the imminent peril of Egyptian thrusts toward its southern borders, Saudi Arabia backed the *imām* during five years of war against large Egyptian

armies, which did not withdraw until after Israel defeated Egypt in the Arab–Israeli War of 1967. In 1970 Saudi Arabia mediated a conciliation between Yemeni republicans and royalists.

King Ibn Sa'ūd of Saudi Arabia died in 1953. His son and successor, Saud, an ineffective monarch, was supplanted in 1964 by another son, Faisal, who obliged Pres. Gamal Abdel Nasser of Egypt to conclude with him the Jidda Agreement (1965), by which Nasser extricated the Egyptian troops from the northern Yemen mountains. After regrouping his troops in safety (they were withdrawn entirely in 1967), Nasser switched his emphasis to political pressure to remove the British from Aden.

In Aden the British were able to contain Egyptian-supplied terrorists without serious loss or dislocation until suddenly, in 1966, the Labour government in Britain decided to abandon the base, despite American and Saudi pressures to stay. This gave a great fillip to terrorist organizations. In late 1967 chiefs in treaty relation with Britain were packed off to Jidda. An extremist group, the National Liberation Front (NLF), took control of South Yemen, which was proclaimed a people's republic and became a centre of Soviet influence, especially after 1969 when the Communist wing of the NLF attained supremacy. Aden (incorporated into the People's Democratic Republic of Yemen), previously a Western bastion, became a base for subversion against the Gulf states and aggression against northern Yemen (now the Yemen Arab Republic, or Yemen [Ṣan'ā']) and Oman. The NLF appropriated all private assets and increased Communist indoctrination and repression. Peace was patched up between the two Yemens in 1972. In 1974, as the President of Yemen (Ṣan'ā') could not maintain the internal political balance, the army made a senior officer president and the constitution was suspended.

In Oman, after a palace revolution in 1970, the new sultan, Qabus, began a program of modernization, welfare, and reform, but the country had to devote much oil revenue to repelling attacks of rebels from Yemen (Aden); these were checked by a signal defeat in late 1975.

When in 1971 the British withdrew from the Gulf, they at least left the region in good order. The former Trucial States of Abu Dhabi, Dubai, Sharjah, Ajman, Fujairah, and Umm al-Qaiwain proclaimed the United Arab Emirates, joined in 1972 by Ras al-Khaimah. Bahrain and Qatar became independent. In ensuing years the Arabian states acquired control of Western oil concerns operating in their territories. The enormous wealth thus acquired empowered them to take a prominent part in external politics, supporting the cause against Israel and subsidizing poorer Arab states.

The early 1970s witnessed major political realignments. Britain's withdrawal from the Gulf cost it all political influence outside Oman (from which, however, the last British troops were withdrawn in 1976). Common interest brought Saudi Arabia and Egypt together, especially after Anwar as-Sadat succeeded Nasser. Saudi Arabia weathered the assassination of King Faisal, who had handled the new oil strategy wisely, notably during Egypt's Sinai campaign of 1973. The Arabian states reached some understanding with Iran, now the strongest Gulf power. In 1976 Saudi Arabia "normalized" relations with and extended a subsidy to Yemen (Aden); certain other Arabian oil states, to neutralize the Communist danger, also had adopted this milder strategy.

BIBLIOGRAPHY. Reliable Western researches on the Arabian Peninsula still mostly consist of articles published in learned journals, though there is a relatively abundant travel literature in book form. The following bibliography consists almost exclusively of books in European languages, chiefly English. Some of these inevitably are not up to date with more modern research. No reference is made to works in Arabic of which translations do not yet exist.

General works: P.K. HITTI, *History of the Arabs*, 8th ed. (1964), a popularly written introduction to Arab history; R.H. KIERNAN, *The Unveiling of Arabia* (1937), an account of Western exploration of the Arabian Peninsula; E. MACRO, *Bibliography of the Arabian Peninsula* (1958); *Encyclopaedia of Islam*, articles on geography, history, and religion of Arabia; A.J. ARBERRY and C.F. BECKINGHAM (eds.), *Religion in the Middle East*, vol. 2, *Islam* (1969); Ibn Khaldūn, *The Muqaddimah* trans. by F. ROSENTHAL, 3 vol. (1958); E. DE ZAMBAUR, *Manuel de la généalogie et de chronologie pour l'histoire de l'Islam* (1955); GEORGE ANTONIUS, *The Arab Awakening* (1938).

Pre-Islāmic Arabia: ALOIS MUSIL, *Northern Neǧd* (1928); J.A. JAUSSIN and R. SAVIGNAC, *Mission archéologique en Arabie* (1909–14); A. FAKHRY, *An Archaeological Journey to Yemen* (1951–52); R. LE B. BOWEN and F.P. ALBRIGHT, *Archaeological Discoveries in South Arabia* (1958), report of the American expeditions of 1950–52 on excavations in Baiḥān and Mārib; G.L. HARDING, *Archaeology in the Aden Protectorates* (1964); F.V. WINNETT and W.L. REED, *Ancient Records from North Arabia* (1970), a field survey of primarily surface data; A. JAMME, *Sabaean Inscriptions from Maḥram Bilqîs (Mârib)* (1962); J. RYCKMANS, *L'Institution monarchique en Arabie Méridionale avant l'Islam* (1951); J. PIRENNE, *La Grèce et Saba* (1958), and with A. MARICQ, *Le Royaume sud-arabe de Qatabân et sa datation* (1961); W.H. SCHOFF, *The Periplus of the Erythraean Sea* (1912); H.V. WISSMANN, *Zur Geschichte und Landeskunde von alt-Südarabien* (1964); G. RYCKMANS, *Les Religions préislamiques* (1960); R. DUSSAUD, *Pénétration des Arabes en Syrie avant l'Islam* (1955); A. LITTMANN, *Thamūd und Safā* (1940); A. VAN DEN BRANDEN, *Histoire de Thamoud* (1960); G. OLINDER, *The Kings of Kinda of the Family of Ākil al-Murār* (1927); A. JEFFERY, *The Foreign Vocabulary of the Qur'ān* (1938).

Early Islām: A. GUILLAUME, *The Life of Muhammad* (1955, later reprints); M.J. KISTER, "Mecca and Tamīm; Aspects of Their Relations," *Journal of the Economic and Social History of the Orient*, 8:113–163 (1965); H. LAMMENS, *L'Arabie Occidentale avant l'Hégire* (1928); J. SAUVAGET, *La Mosquée Omeyyade de Médine* (1947); J. WELLHAUSEN, *The Arab Kingdom and Its Fall* (1927, reprinted 1963), still a useful work though some theories are now disputed.

Northern Arabia: GERALD DE GAURY, *Rulers of Mecca* (1951); C. SNOUCK HURGRONJE, *Mekka in the Latter Part of the 19th Century* (Eng. trans. 1931); H. ST. JOHN PHILBY, *Sa'ūdi Arabia* (1955), a discussion of modern Sa'ūdī history, and *Arabian Highlands* (1952); AMIN RIHANI, *Ibn Sa'oud of Arabia: His People and His Land* (1928); R. BAYLY WINDER, *Saudi Arabia in the Nineteenth Century* (1965).

Southern Arabia: C. VAN ARENDONK, *Les Débuts de l'Imāmat Zaidite au Yemen* (1960); H. CASSELS KAY (trans.), *Yaman: Its Early Mediaeval History* (1892, later reprints); C. NIEBUHR, *Voyage en Arabie*, 2 vol. (1774–76); H. ST. JOHN PHILBY, *Sheba's Daughters* (1939); R.L. PLAYFAIR, *A History of Arabia Felix or Yemen* (1859, reprinted 1970); J.W. REDHOUSE, *The Pearl-Strings: A History of the Resuliyy Dynasty of the Yemen* (1906–08); R.B. SERJEANT, *The Portuguese Off the South Arabian Coast* (1963); A.S. TRITTON, *The Rise of the Imams of Sanaa* (1925). Among later works may be cited DONALD HAWLEY, *The Trucial States* (1970); J.B. KELLY, *Britain and the Persian Gulf, 1795–1880* (1968); DEREK HOPWOOD (ed.), *The Arabian Peninsula: Society and Politics* (1972); and C.B. MCLANE, *Soviet Middle East Relations*, vol. 1 (1976).

(R.B.Se./M.A.G.)

Arabian Desert

One of the great desert regions of the world, the Arabian Desert occupies almost the entire Arabian Peninsula. Covering an area of about 900,000 square miles (2,331,000 square kilometres), it is bordered on the west by the Red Sea, on the north by the Syrian Desert, on the south by the Arabian Sea and the Gulf of Aden, and on the east and northeast by the Persian Gulf, the Gulf of Oman, and the Arabian Sea. (The Eastern Desert of Egypt, sometimes called the Arabian Desert, is not considered in this article. See also related entries under ARABIAN DESERT in the *Ready Reference and Index*.)

A large part of the Arabian Desert lies within the modern kingdom of Saudi Arabia. Yemen (Ṣan'ā') on the coast of the Red Sea, and Yemen (Aden), on the coast of the Gulf of Aden, border the desert to the southwest. Oman, bulging out into the Gulf of Oman, borders the desert at its eastern extremity. The sheikhdoms of the United Arab Emirates (former Trucial States) and Qatar rim the region to the north, stretching along the southern coast of the Persian Gulf. The sheikhdom of Kuwait abuts the northern Persian Gulf between Saudi Arabia and Iraq; a Neutral Zone—diamond shaped—lying to the west of Kuwait is shared by Saudi Arabia and Iraq. In the northwest the desert extends into Jordan.

Seen from the air the Arabian Desert appears as a vast

expanse of light sand-coloured terrain, with an occasional indistinct line of escarpments or mountain ranges, black lava flows, or reddish systems of desert dunes stretching to the horizon. Camel trails crisscross the surface between watering places. On the ground, features become distinctly individual and the relief seems more prominent. Vegetation at first seems nonexistent, but to the discerning eye can be seen as a minor fuzz on the surface, or as bits of green where shrubs strive to survive. There is almost always a breeze, which changes seasonally to winds of gale force. Cold or hot, these air currents chill the body or roast it. The sun and moon are bright in clear skies, although dust and humidity may cut visibility with little warning. Contrary to ideas commonly held in more northern climes, the desert is often a lovely place.

General description. Western Arabia formed part of the African landmass before a rift occurred in the earth's crust, as a result of which the Red Sea was formed and Africa and the Arabian Peninsula became separated. The southern half of the peninsula consequently has a greater affinity with the Somalia and Ethiopia regions of Africa than with northern Arabia or the rest of Asia. The northern Arabian Desert merges imperceptibly into Arab Asia through the Syrian steppe (treeless plain). The bulge of Oman contains a mountain range similar to ranges in Iran.

African affinities

The greatest length of the peninsula from northwest to southeast is about 1,600 miles (2,600 kilometres), from north to south not quite 1,400 miles (2,300 kilometres), while from east to west the maximum width is about 1,300 miles (2,100 kilometres). It is narrowest from Rābigh on the Red Sea to Manama on the Persian Gulf —a distance of about 700 miles.

Three corners are high: the southwest corner in Yemen, where Ḥaḍūr Shu'ayb reaches a height of 12,336 feet (3,760 metres); the northwest corner in Hejaz (a part of Saudi Arabia), where Jabal al-Lawz reaches a height of 8,562 feet (2,580 metres); and the southeast corner in Oman, where al-Jabal al-Akhḍar attains an altitude of 9,774 feet (2,980 metres). Much of the Yemen Plateau is over 9,840 feet (2,999 metres) high. To the north and east elevations decrease. Steep cliffs and steep canyons descend from highlands into adjacent seas south and west. Northeast slopes in Oman are short and steep, but on the southwest flanks the slopes grade gently to the Rub' al-Khali desert basin. The southern plateau is cut by great steep-walled canyons into rugged limestone masses that have held the peoples of that region in isolation for hundreds of years. On lower surfaces eastward in Dhofar grow shrubs that yield the fragrant frankincense and myrrh.

The rest of the peninsula displays a moderate relief characterized by broad plains or steppes. At least a third is covered by sand. The Red Sea escarpment, however, stands in sharp contrast to the moderate relief of the interior plateaus. South of aṭ-Ṭā'if, east of Mecca, the scarp is young, rugged, and dissected into short, steep canyons and ridges. At the foot of the scarp the Tihāmah coastal plain slopes to the sea; at Jabal as-Sawdā', near Abhā in Asir, the drop is about 9,000 feet in six miles. North of aṭ-Ṭā'if the Hejaz and Najd plateaus seldom rise above 3,600 feet (1,100 metres), except where volcanic fields occur or where remnants of the crystalline rocks that underlie the region rise to the surface. The slope to the Persian Gulf averages eight feet per mile.

Geomorphology. The Arabian Desert consists of two major regions: the first, the ancient Afro-Arabian Shield of rocks that are igneous (solidified from a molten state) and metamorphic (transformed by heat and pressure) in the west; and the second, younger sedimentary rocks that pitch gently away from the Shield toward the Persian Gulf basin. Floods of volcanic lava in the west disturbed drainage patterns and created new divides. The highlands of the Hejaz, Yemen, the Ḥaḍramawt, and Oman display structural complexes of the whole rock spectrum. Volcanic activity in Yemen and Oman dates from the Mesozoic Era (from 65,000,000 to 225,000,000 years ago), while more recent volcanic eruptions occurred in Aden, Asir, Hejaz, and Jordan between the Oligocene Epoch, which began 38,000,000 years ago, and the Recent Epoch, which began about 10,000 years ago.

The two major structural regions

Mountainous highlands rise in northwest Hejaz, in the Asir region, in Yemen, and in Oman. Lesser ranges have been uncovered by erosion in the interior. Eighteen volcanic fields are scattered through the west, mainly in Hejaz, several of them being more than 10,000 square miles in area.

Plateaus are a common desert feature. Jordan east of the Dead Sea forms a moderately elevated plateau. To the southeast, Jibal Tubaiq rises higher, standing as a mass of sandstone deeply cut into by numerous seasonal watercourses (wadis). Farther southeast the plateaus of Tabūk, Taymā', Ṭawīl, al-Hufrah, and al-Hūj reach to the west edge of an-Nafūd (Great Nafud), a sand desert in the north. Through central Najd, a highland region southeast of an-Nafūd, a series of west-facing scarps mark cuestas (low ridges with steep faces on one side and gentle slopes on the other) of limestone reaching to highlands of the Ḥaḍramawt in the south, where the plateau of al-Jawl (Jol) is located.

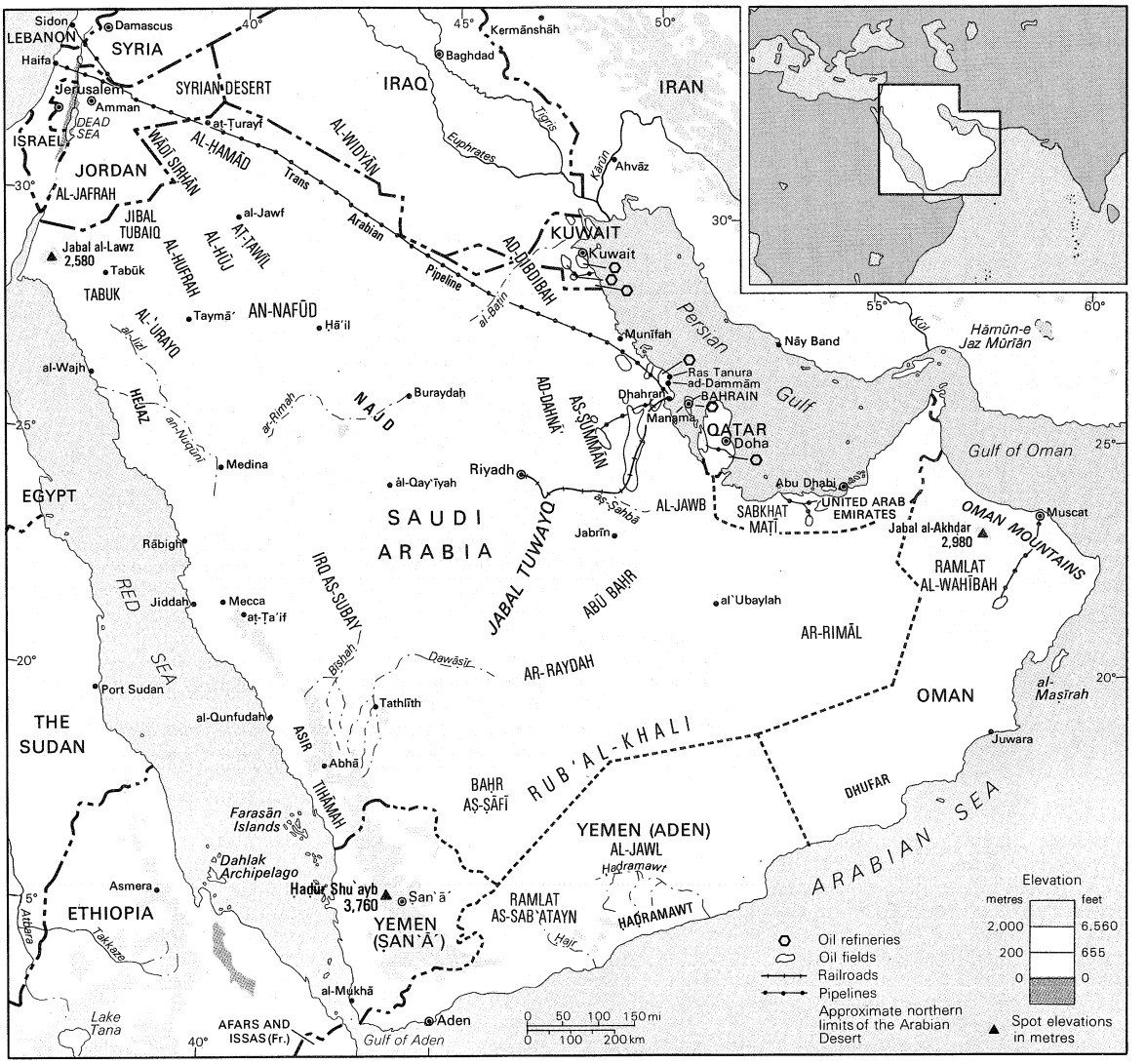
Below the plateaus spread broad plains, stony, cherty (composed of compact microcrystalline quartz), or gravelled, their surfaces well preserved under the arid climate; some plains are duricrusted (covered with a crust of soil formed by salts), having smooth, firm surfaces formed by the cementation of sandy debris at groundwater level. Typical of the stony plains is al-Ḥamād, which stretches from an-Nafūd northward into the Syrian Desert. Chert plains were formed on the post-Eocene to pre-Miocene surface (*i.e.*, between 26,000,000 and 38,000,000 years old) in al-Ḥamad, and in al-Malsūnīyah region east of the Khurayṣ oil field. The gravel plains resulted from deposits left from 10,000 to 2,500,000 years ago by ancient river systems now represented by such wadis as ar-Rummah–Bāṭin, Sahbā, and Dawāsir–Jawb, which carried vast loads of sediment from the interior toward the Persian Gulf. Ad-Dibdibah was the delta of Wādī ar-Rummah–Bāṭin, and al-Ḥaḍabah was the delta of Wādī aṣ-Ṣahbā. The gravel plains of ar-Ṛaydah and Abū Baḥr, and adjacent areas covered by sand, formed the delta of the Dawāsir–Jawb system. Several of the deltas are as large in area as the delta of the Nile. The plateau of the northern Ṣummān is smooth where it has been uncovered from under ad-Dibdibah gravels. Farther south, the surface is cut by erosion into many closed basins.

Divides and drainage basins. When the Arabian Peninsula was shifted away from Africa, the Afro-Arabian Shield was greatly elevated, thus disturbing prior drainage patterns. The uplifting of the peninsula and the appearance of cracks or faults were accompanied by large upwellings of lavas that filled valleys and covered ridges and mountains. The Arabian platform was tilted northeastward, creating a prominent regional divide along its western rim from Yemen to Jordan. In the south another regional divide separates the coastal drainage of the Ḥaḍramawt from the Wādī Ḥaḍramawt system inland, and a third system, also in the south, divides the al-Jawl region from the system draining into the Rub' al-Khali. The Oman Mountains divide short, steeply graded northeast-flowing wadis from the less steep wadis flowing southwest into the eastern Rub' al-Khali.

The effects of geological uplift

Lesser divides enclose basins in Hejaz, Najd, and Jordan. Four adjacent interior basins are to be found in the shield between latitude 21° and 26° N. Others occur in the Tabūk area, in the Wādī as-Sirḥān, at al-Jafrah in south Jordan, and in the Jordan–Dead Sea depressions. Hundreds of smaller basins, seldom larger than four to eight miles in diameter, are found in volcanic fields, in the shield, and in many poorly drained areas of sedimentary rocks. A few are saline, but the vast majority are floored by silts deposited in thin layers.

Most drainage channels in the Arabian Desert are either dry or else are intermittent, flowing only when rains are heavy. Two systems flow perennially in the region—the Tigris–Euphrates rivers and Wādī Ḥajar in the southern Ḥaḍramawt.



The Arabian Peninsula.

The main drainage systems of al-Widyān, ar-Rummah–Bāṭin, Sahbā', and the Dawāsir–Jawb were the scenes of great floods in the Quaternary Period, which began 2,500,000 years ago. Today floods are infrequent but no less destructive; they seldom, however, reach the desert sands where the channels have been dammed up. The directions taken by several large systems have been altered by stronger streams that have intercepted them; such wadis as the Jizl-Ḥamd in northern Hejaz and the Ḥaḍramawt in the south provide examples.

Where they enter large expanses of sand, wadis lose their identities. The alluvium carried in flood time is added to the sand body and redistributed by the wind. About two-thirds of the Arabian Desert is drained by complete wadi systems. The intermittent action of running water is more effective than the erosive action of the winds in shaping the landscape. Most of the exceptions are provided by the formation of sand dunes and the grooving of rocks by wind-borne sand.

Saline flats. The *sabkhah* (saline flat) of Eastern Arabia is usually a gently sloping sandy plain with concentrated salt brine at or just below the surface, located on a coastal plain. It has formed by the filling of embayments with sand—a process that continued as the sea level retreated—while the high rate of evaporation concentrated the trapped seawater to strong brine. Scouring of the sandy surface by *shamāl* (north to northwest) winds exposes the salty crust, which may then be dissolved into brine by winter rains. High tides may spread the saline surface waters inland. The salt crust does not usually attain a thickness of more than three feet. It may be interbedded with sand, silt, mud, or other deposits formed by evaporation, such as gypsum. The surface of Sabkhat Maṭī, the largest single exposed *sabkhah*, is sometimes composed of soft, wet, muddy, and salty slush, with a hard crust less than three feet deep. It is very treacherous to the unwary traveller, be he on foot, camelback, or in a motor vehicle. The *sabkhah* surface has no direct analogy with quicksand but its danger lies in the inability of the traveller to recognize its nature in time to avoid sinking into a morass. Northwest of the oil refinery on Ras Tanura, *sabkhah*s carry a thin crust of salt and sand, underlain by soft calcareous (chalky) mud that is formed in shallow tidal flats by algae and eventually is covered by sand or salt. It has the consistency of custard.

The eastern Rub' al-Khali has a broad *sabkhah* floor that rises to less than 330 feet in elevation; this appears to represent the highest altitude that *sabkhah*s attain.

There are saline basins, akin to playas (the sandy, salty, or mud-caked floors of desert basins) in the Americas, which develop in enclosed valleys, either from the evaporation of flood waters bearing dissolved minerals, or—more usually—from the evaporation of saline waters nurtured by nearby outcrops of salt. The Arabic name for this kind of salt flat is *mamlaḥah*. Arabs have quarried crude salt from both *sabkhah* and *mamlaḥah* for hundreds of years.

The sand deserts. Sand covers at least a third of the Arabian Desert, occurring as dunes of varying size and complexity or as a thin film on surfaces of lower relief. With few exceptions, sand does not accumulate in flat

Sand-dune formations

sheets but in dunes, ridges, or giant complexes. The variety of dune shapes and sizes in the Arabian Desert is legion. Many forms have not been described in print. Contrary to statements by early western explorers that there is nothing but a shapeless sea of sand, the sand desert develops along systematic lines with distinct and characteristic patterns. There are also clear-cut genetic relationships between dunes in adjacent areas. In such huge sand areas as the Rubʿ al-Khali, the evolution of dune forms can be traced from simple dunes into the more complex types.

The two largest sand bodies in Arabia are an-Nafūd in the northwest, and the Rubʿ al-Khali (ar-Ramlah) in the southeast. An-Nafūd has an area of 26,000 square miles; the Rubʿ al-Khali has an area of 229,000 square miles. Between them are two almost parallel arcs of more or less continuous dunes. The outer arc, convex to the east, is ad-Dahnāʾ, about 750 miles long and from six to 50 miles wide. The inner arc is shorter and less continuous and includes six elongated *nafūd*s situated in low areas between the west-facing limestone escarpments of the central Najd. The two major arcs are separated by the great Jabal at-Tuwayq cuesta. Ad-Dahnāʾ connects with an-Nafūd in the northwest and with the Rubʿ al-Khali in the southwest. The inner belt grows more disconnected south of 24° N and dies out before reaching Wādī ad-Dawāsir.

An-Nafūd. This sand desert fills an irregular basin surrounded by hills, ridges, and buttes (isolated hills rising abruptly from the surrounding land) or mesas (land formations having relatively flat tops and steep rock walls). Its sand was eroded and carried by winds from outcrops of sandstone that lie west and northwest. An-Nafūd is characterized by a unique dune form, a giant crescent slipface (the slope on the leeside of the dune that approximates the angle of rest of loose sand, usually about 32°) moving through a thick mass of loose sand. The lee hollow in front of the slipface is floored by bedrock, barren of sand. The impression obtained from an aerial view is of hooflike hollows, that extend from the southwest edge, well across the Nafūd. Along other margins of the desert and through the centre are linear and pyramidal dunes. The Nafūd has been a barrier to travel for ages, only rarely attempted by Western travelers. A caravan route runs from northwest to southeast connecting the oases of al-Jawf with those of Ḥāʾil. The topographic relief of dunes in the Nafūd may exceed 300 feet.

Rubʿ al-Khali. With an area of about 230,000 square miles—larger than France—this huge sand desert displays a variety of dune forms that appears to be unique, although parts of the Sahara show many similar dunes as well as a few that are different. The name Rubʿ al-Khali (the Empty Quarter) is not usually used by the Bedouins who travel over it, who refer to it instead as ar-Ramlah (the Sand). The Rubʿ al-Khali has five general types of sand terrains: (1) crescentic or transverse; (2) linear; (3) giant dune complexes, or "sand mountains," that take the shapes of domes, pyramids, huge crescents, and sigmoidal (S-shaped) ridges; (4) hook-shaped dunes; and (5) dunes with vegetation and moisture. The eastern Rubʿ al-Khali fills a broad shallow basin, floored mainly by *sabkhah*s that slope toward the southern shores of the Persian Gulf. The floor west of the 51st meridian rises to the southwest to nearly 3,600 feet (1,100 metres) and is composed largely of gravel plains. Highlands surround this sand desert on the northeast, south, and west.

"The Empty Quarter"

The greatest volumes of sand lie in the eastern basin and—as has been discovered through the use of aerial photography—are arranged in belts. These belts appear to have accumulated during the Pleistocene Epoch (from 10,000 to 2,500,000 years ago) at several successive stages that were related to the advance and retreat of sea levels.

Other sand deserts. Ramlat as-Sabʿatayn in eastern Yemen, in the lowland south of the western border of the Rubʿ al-Khali, has an area of about 60 by 150 miles, or about 10,000 square miles. It is composed mainly of transverse and linear dunes.

Ramlat Ahl Wahībah is located near the extreme eastern cape of the peninsula; it is an oblong dune field 60 by 100 miles wide, made up of linear ridges and interdune valleys oriented almost north and south.

An elongated sand body named al-ʿUrayq lies in the southern part of the Wādī ar-Rummah basin. ʿIrq as-Subbayʿ (*ʿirq* is an Arab word for linear sand dune) dams the middle Wādī al-Jarab drainage system. A small cluster of broad transverse dunes has formed on the south coast east of the port of Aden.

Climate. The Arabian Desert spreads across 22° of latitude from the 12th to the 34th parallel north; it may thus be considered as a tropical desert. Summer heat is intense, reaching a maximum 124° F, (51° C). In the interior the heat is dry and tolerable. Coastal regions and some highlands, however, attain high summer humidity with dews and fogs at night or early morning. Rainfall averages less than four inches (102 millimetres) a year but can range from zero to 20 inches. Interior skies are usually clear except for intermittent winter rains, spring hazes, or dust storms. Torrential rains flood the main drainage basins infrequently. Winters are invigoratingly cool, with the coldest weather occurring at high altitudes and in the far north. A minimum recorded at aṭ-Ṭurayf on Tapline (the Trans-Arabian Pipeline) in 1950 was 10° F (12° C), and was accompanied by several inches of snow and an inch of ice on ponds. Summer rains in the Rubʿ al-Khali accompany the monsoon winds from the Indian Ocean. Winter rains may occur in the northern Rubʿ al-Khali. The most arid part of the Arabian Desert appears to be on the western margin of the Rubʿ al-Khali, north of Wādī ad-Dawāsir.

Winds. Dominant winds blow from the Mediterranean, swinging to the east, southeast, south, and southwest in a great arc. Two semiannual windy seasons occur from December to January and from May to June. These are called *shamāl*s (from the Arabic: "north"), and last from 30 to 50 days with wind velocities averaging 30 miles per hour. *Shamāl*s, which try the patience of man and beast, are dry, transport huge loads of sand and dust, and alter the shapes of sand dunes. Millions of tons of sand are carried by each storm into the Rubʿ al-Khali. Blown sand does not rise more than a few feet, except when picked up by whirlwinds, dust devils, or regional sandstorms. Winds box the compass in central Najd and in the southeastern Rubʿ al-Khali. Strong southeast gales sweep the big sand desert for several days at a time, reversing the effect of *shamāl*s on dunes.

The power of desert winds to excavate basins is much exaggerated. The wind is important as an agent of erosion but, as mentioned above, is never as effective as running water.

"Brown rollers"

The sudden appearance on the horizon of the "brown roller" in spring or fall can be frightening. This is a frontal storm up to 60 miles wide, carrying sand, dust, and debris high in the air; it is followed by a sharp drop in temperature and often by rain. Wind velocities reach gale force for half an hour or so. Hot days produce myriads of dust devils (*jinn*) and the ill-famed mirage.

Soils. Desert soils differ from humid soils in that they undergo less chemical weathering. Mechanical weathering breaks down coarse particles into finer grains. Quartz sand abounds, covering more than a third of the desert surface. Granular debris from the crystalline basement of the Precambrian shield forms pebbly fans about the bases of hills. Sands and silts are washed down to lower levels and are then winnowed out by winds. Fine materials grade down to silt. Smaller particles, such as clays, rarely form. Limestone, when pulverized, forms silt-sized dusts. Water-borne silts eventually are deposited in *khabari*, or silt flats. Irrigated silt flats are farmed, the soils having proved to be fertile. Najd villages that once depended on November rains to raise their winter wheat, now irrigate and farm all year. Highlands of Asir and Yemen are terraced for crops. Salt flats in the desert, though too salty for many crops, if irrigated and drained properly, can be cultivated. Many of the plants that grow in saline soils, and which are called halophytes, are succulents that are suitable as camel fodder. The date palm thrives on salty soils if properly irrigated and drained.

Desert dune sands are generally dry but can hold rainfall to depths of three feet or more, thus nourishing xerophytes (plants adapted to survive under arid conditions). Shrubs unique to the area, called *ʿabl* and *ghaḍā*, send out long, shallow roots to catch the slightest bit of moisture. The roots make good firewood.

The highlands of Yemen and Asir produce forests of juniper trees. Valleys and lower slopes are extensively terraced for soil and water conservation and produce many crops, of which coffee is important in local markets. The soils are derived from crystalline rocks of high mineral content.

Vegetation and animal life. *Vegetation.* There is a great variety of desert flora and fauna. Plants are primarily xerophytic or halophytic. After spring rains long-buried seeds germinate and bloom in a few hours. The normally barren gravel plains turn green. Even chert plains produce late winter and early spring grazing for camel and sheep. The plains were once the home of the famed Arab horse, but grazing was always too poor to support a large horse population. Certainly all the grazing areas were overgrazed, thus contributing to the formation of the present widespread barren tracts. Halophytes grow on the saline flats and include many succulents and fibrous plants that can be eaten by camels. The sedge, which grows in sandy areas, is a tough plant with deep roots that help to hold down the soil. The tamarisk tree is often found on the borders of oases, where it helps to prevent the encroachment of sand.

The rare shrub *rāk*, or *arāk*, is known as the "toothbrush bush," its twigs being used by Arabs to polish their teeth. Many herbs grow throughout the desert and are well known to the Bedouins, who use them for seasoning, preserving food, perfuming clothing, and washing hair. The eastern Rubʿ al-Khali, generally thought to be dry and barren, supports much plant life on flanks of giant dunes, including a sweet grass called *naṣī* that provides the main forage for the now-rare oryx (a species of African antelope). There are no cacti in the Old World, except for those imported from the Americas. One of these imports, the prickly pear, thrives and is fed to livestock, the fruit being eaten by people. Spiny and thorny plants are common. The euphorbia, a plant with milky juice and flowers with no petals, grows in Hejaz, and the camelthorn is everywhere. Acacia trees were once abundant in Jibal Tuwayq, but the demand for charcoal decimated them. The few growing in wadis and gardens provide welcome shade. A single clump of four stands in daring isolation on the edge of al-Mijann overlooking the southern Persian Gulf and the feared Sabkhat Maṭī (the treacherous saline flat already mentioned). Junipers reach a great size in high Asir and Yemen. The trunks are cut into the beams and pillars that characterize the region's architecture. The milkweed tree (*ishar*) grows to a height of 20 feet in Wādī al-Bātin and is common in the wadis of Najd and in Wādī Bīshah.

The ubiquitous date palm

The date palm is grown in many oases, the dates themselves providing food for humans, camels, donkeys, and horses. There are many varieties. The palm supplies wood for building and for making water well frames and pulley shafts of ancient type; its fronds are used for handicrafts and for thatching roofs. The oases also produce many fruits and vegetables such as rice, alfalfa, henna (a shrub which yields a reddish-orange dye), citrus, melons, onions, tomatoes, barley, wheat, and—in higher regions—peaches, grapes, and prickly pears.

Animal life. The animal life is varied and unique. Desert insects include the fly, the malaria-carrying *Anopheles* mosquito, fleas, lice, ticks, roaches, ants, termites, beetles, and mantids (predatory insects), which camouflage themselves as leaves, twigs, or pebbles. Also found are the scavenging dung beetle, myriads of butterflies, moths, and caterpillars, and the pestiferous locust that periodically plagues the landscape and populace in great swarms.

Arachnida (a class of segmented invertebrates) include large sapulgids (scorpion-killers), scorpions, and spiders. Sapulgids grow to eight inches in length. Scorpions also range up to eight inches and are coloured black, green, yellow, red, and off-white. The scorpion's painful sting is deadly to small children.

Pools in oases contain small fish. There are a few amphibious animals, such as newts, salamanders, toads, and frogs. Reptiles include lizards, snakes, and turtles. The *ḍabb*, a fat-tailed lizard, lives on the plains and reaches a length of up to three and a half feet. It is a vegetarian with toothless jaws; its fat tail, roasted, is a Bedouin delicacy. The monitor lizard reaches lengths up to three feet; it feeds on locusts and other insects. Many lizards, including skinks, geckos, agamids, and collared lizards, are found in the sands. A salmon-coloured lizard, the *dammūsa*, is lively and pretty, seeking the black beetle for food, and literally diving and swimming in the slip-faces of the sand dunes. An agamid lizard (*ṭuḥayḥī*) scurries across the sand with its tail coiled like a watch spring, uncoiling when it stops.

Among the snakes, all of which are feared by Arabs, the sand cobra—relative of the sea snake—is slim, sand-coloured, and venomous. Vipers abound in sand and rocks but, being nocturnal, are seldom seen in the heat of day.

Birds of the Arabian Desert include local species as well as migrant groups. The migrants are from northern Europe, Africa, and India. The local birds breed from late winter to early spring. Many of the young display excellent camouflage. The bifasciated (striped) lark, the sand grouse, the Arabian courser, and the lesser bustard live in the desert all year, as also do several falcons, eagles, and vultures. The peregrine falcon is seen in Asir; saker and lanner falcons (a brown falcon with a golden cap) are found in Najd and the Eastern Province; and the kestrel is everywhere. The saker falcon (an aggressive light-brown falcon) is often captured young and trained by Bedouin falconers to hunt the bustard and sand grouse. Ravens in pairs or flocks may appear anywhere. Three eagles are known—the white-tailed, golden, and tawny eagles. Vultures were more numerous when camels were in greater use. The largest, a black species with a wing spread of up to 12 feet, has nearly disappeared. The Egyptian vulture (*ar-rakhamah*), a medium-sized bird coloured white and black with yellow, is widely distributed. The lammergeier vulture lives in Asir and Yemen. There are several owls, a burrowing species being common.

Migrant birds follow several flyways, one through the central Najd and others on each coast. Water and shore birds migrate in fall and spring between northern Europe and the tropics. Bee-eaters, warblers, babblers, carrion kites, swallows, martins, swifts, wheatears, shrikes, larks, flycatchers, hoopoes, and some exotic species may be seen alone, in pairs, or in flocks. Cranes, heron, flamingoes, ducks, and small wading birds feed on shores and in the lakes which intermittently appear. The ostrich, once abundant in the sand deserts, has been extinct since 1940.

Desert game

Mammals were probably numerous before the Arabs hunted them with rifles, shotguns, and submachine guns from motor vehicles. Gazelles, which roamed the plains in herds of hundreds before World War II, are almost extinct. The oryx had disappeared from the Rubʿ al-Khali by 1960. The ibex (a species of wild goat), which dwells on cliffs in Jibal Tuwayq, has been decimated. In desert plains, the ratel (a badger-like carnivore), the fox, and the civet cat live in territorial isolation. Wolves are much feared but, although widespread, are not numerous. The hyena lives wherever sheep are herded, preferring escarpments that provide cover. Jackals are to be seen, especially at dusk when they appear for water. There are hares, as well as golden sand rabbits. Small rodents include the jerboa (a mouselike rodent), mice, rats, and porcupines, while small hedgehogs are often found among rocks. Troops of baboons run in Asir.

Human ecology. Man has inhabited the Arabian Desert since early Pleistocene times. His artifacts have been found widely spread throughout the desert but are most abundant in the southwestern Rubʿ al-Khali, where game was prolific until about 30 years ago. Archaeological research was not encouraged by the government until re-

cently, so little is known about mankind in Arabia in earlier epochs. Remains of cultures from the last 3,000 years occur in Hejaz, Asir, Yemen, and the Ḥaḍramawt. More recent remains—perhaps 1,000 to 2,000 years old—are being examined in the Eastern Province; most of them date from the early years of the Islāmic period.

The Bedouin know little of their history beyond the dawn of Islām. They adapted to nomadic desert life by breeding camels, Arab horses, and sheep, but also conceded to necessity by growing date palms and other crops. The Bedouin seldom abandoned the nomadic life for the oasis longer than was essential to gather crops, generally hiring others for agricultural labour; finding grazing and water were their main concern, in addition to conducting raids to seize horses and camels. Grouped in hereditary tribes, the Bedouin claimed certain lands as their *dirah* (tribal territory), where their flocks could graze and water. After 1925 King Ibn Saʿūd of Saudi Arabia, who encouraged the Bedouin to settle in oases, was able to use his influence to prevent intertribal raiding, and to make the tribal *dirah* of less importance. Tribal loyalties remain strong and no Bedouin of a noble tribe will marry into a lesser tribe. Islām is their way of life, but modernization has brought much change and has lessened religious influence. Townsmen have kept their identity distinct from that of the Bedouin; each group regards the other with contempt. This feeling has lessened somewhat, however, since more Bedouins have settled in towns. Western cultural influence began with the discovery of petroleum in 1936 and led to the introduction of the motor vehicle, airplanes, telephone, radios, telegraphs, and, recently, television. The cinema has also become popular, despite disapproval by religious leaders.

Bedouin life

Western cultural influences

Schools were few and were only for boys until 1960; since then, however, girls are also taught. The veil is still worn by women, and, though the lot of women has improved, their place remains in the home.

Once crossed only by camel trails, the desert is now laced by networks of motor roads. Bedouins use trucks to travel from place to place, taking their families, belongings, and sheep with them. Instead of sheep being herded from grazing to water hole, water is now hauled to the sheep by truck. The camel caravan has practically disappeared, although the camel remains the mainstay of Bedouin culture.

Travel and exploration. Before the 18th century few travellers or would-be conquerors had succeeded in penetrating the Arabian Desert; fewer still had written about it. Scientific exploration began in 1762 with a Danish expedition led by the German surveyor Carsten Niebuhr. In the 19th century, British officers of the Indian government undertook surveys of the surrounding seas and coasts.

The first classical work on the geography of Arabia, *Travels in Arabia Deserta* (1888), was written by the English traveller C.M. Doughty. At the turn of the 20th century, the Czech explorer Alois Musil travelled through northern Hejaz and Najd, mapping topography as he went. In 1917 H. St. John Philby, an official of the British Foreign Office, paid a visit to the Sultan of Najd (later King Ibn Saʿūd of Saudi Arabia); Philby later became a Muslim, settled in Riyadh as a councillor to Ibn Saʿūd, and explored the Arabian Desert by camel and motor car, writing detailed and accurate accounts of his travels. Another British official, T.E. Lawrence, who worked with the Sultan of Hejaz against the Turks in World War I, gained fame for his romantic writings about his exploits in the region. Many individuals travelled in limited parts of the desert; the most notable among them is the British traveller Wilfred Thesiger, who crisscrossed the Rubʿ al-Khali after World War II.

The discovery of oil in 1932 on Bahrain and the subsequent American exploration of the Arab mainland that began in 1933 led, as already mentioned to the opening of the Arabian Desert to Western influence. After World War II, geographical and geological exploration intensified and was accompanied by vast aerial photographic surveys as a result of which the first accurate maps of the peninsula were prepared and published between 1956 and 1965. Since 1950, the U.S. State Department and the U.S. Geological Survey have co-operated with the Kingdom of Saudi Arabia in conducting surveys of mineral resources.

Travel today in the Arabian Desert is easy and rapid. Instead of the slow, laborious camel caravan, motor vehicles now roar across desert terrains. Jet planes fly overhead, and the railroad from ad-Dammām on the Persian Gulf to Riyadh, the political capital of Saudi Arabia, covers the distance in a few hours. Arab countries have planned rebuilding the railroad from Medina to Damascus, destroyed in World War I, but conflict with Israel appears to have halted construction. Paved roads cross the desert from the Persian Gulf to Jidda and also follow the Trans-Arabian Pipeline from Dhahran to Sidon in Lebanon. Semi-maintained dirt roads connect many Arab communities. New roads have been built into Asir, making that province and its fine agricultural produce accessible to the markets of Riyadh, aṭ-Ṭāʾif, Mecca, Jidda, and Medina. A Saudi Arabian government airline schedules flights between towns in different regions of the desert.

About 80 percent of the Arabian Desert is accessible to motor vehicles, with or without roads. Essential equipment consists of vehicles equipped with sand tires and two-way radios, and having the capacity to carry fuel, food, and water, as well as repair parts and camping equipment.

Natural resources. The greatest natural resource of the Arabian Desert is its underground water supply, which, as it remains virtually unreplenished because of low rainfall, in effect consists of Pleistocene waters that are now being tapped. Modern techniques are being used by the governments of Arab countries to develop water sources and to irrigate soils for farming.

Petroleum was found in 1936 in the Eastern Province at Dammām Dome, on which the oil-company town of Dhahran is built, but commercial production was not achieved before 1938. Since World War II many new oil fields have been brought into operation, with production exceeding 3,000,000 barrels daily; reserves are enormous and the end is not yet in sight. Prospecting for other minerals has been in progress since 1950.

The oil boom

Building materials in use before the mid-20th century were stone, adobe, a crude cement made from impure calcareous (chalky) rock taken from the floor of the Persian Gulf, the wood of the date palm, the wood of the juniper in the Asir region, and thatches of cane, palm fronds, or other vegetation. Modern construction utilizes steel, concrete, light alloys, imported lumber, local stone—especially granites from Asir and limestone from the Najd—and slates from Hejaz. Salt and gypsum are produced from saline flats.

Future prospects. The Arabian Desert, until about 1940, had remained practically unaffected by other cultures, and particularly Western culture. In the future one may anticipate a change in the influence of Islām, due to the impact of modernization. The desert, nevertheless, will remain the desert, although it will become less isolated, more comfortable, and possibly more productive. The question that remains to be answered, however, is whether or not the Arab individuality will become merged into an urban anonymity.

BIBLIOGRAPHY. GEORGE B. CRESSEY, *Crossroads* (1960), is a general account of the geography of southwest Asia. H.R.P. DICKSON, *The Arab of the Desert*, 3rd ed. (1959), is a compendium of Bedouin culture and customs prior to 1940 in northeast Arabia. C.M. DOUGHTY, *Travels in Arabia Deserta*, new ed., 2 vol. (1937), a classic of Arabian exploration, is one of the best accounts of the desert and its peoples extant. T.E. LAWRENCE, *Seven Pillars of Wisdom* (1935), is an autobiographical account of the defeat of the Turks in Hejaz during World War I. ROY LEBKICHER *et al.*, *Aramco Handbook*, rev. ed. (1960), is an illustrated handbook for company employees on the history, culture, people, and development of the petroleum venture in Saudi Arabia. R. MEINERTZHAGEN, *Birds of Arabia* (1954), contains a chapter about the desert and its biological effects on birds. ALOIS MUSIL, *Arabia Deserta* (1927), is the first of a series of geographical studies on Arabia, translated from Czech and

published by the American Geographical Society. The four titles by the same author that follow in the series are: *Northern Nĕgd* (1928), *In the Arabian Desert* (1930), *The Manners and Customs of the Rwala Bedouins* (1928), and *The Northern Hevâz* (1926). The transliteration of Arabic names into Czech were not transformed into the English system, and place-names are difficult to locate. H. ST. JOHN PHILBY travelled and described the country in a series of detailed accounts from 1917 to his death in the late 1950s. His books include *The Empty Quarter* (1933) and *Arabian Highlands* (1952). BERTRAM THOMAS, *Arabia Felix* (1932), is the account of a British political agent in Muscat and Oman, his travels in Ḥaḍramawt and across the Empty Quarter. WILFRED THESIGER, *Arabian Sands* (1959), is a modern classic on the sand desert of the Rubʿ al-Khali; F.S. VIDAL, *The Oasis of al-Hasa* (1955), a survey and analysis of a large oasis and its community in eastern Arabia.

(D.A.H.)

Arabian Religions

Under the name "Arabian religions" are grouped the cults that flourished in the Arabian Peninsula, or in adjoining areas among Arabic-speaking peoples, from earliest recorded times until the rise of Islām in the 7th century AD.

Nature and significance. In the modern world, Arabia has been so frequently and so thoroughly linked with the religious tradition of Islām that its pre-Islāmic religious history has received relatively little attention. Yet pre-Islāmic Arabia had a rich and diverse culture, in which the primitive and the cosmopolitan, the urban and the nomadic, often were found side by side.

Islām itself has done much to characterize its own background, for from its 7th-century beginnings Islām was presented as a new revelation, this time in Arabic, summing up the monotheistic tradition of the previous Jewish and Christian prophets. In so doing, it looked to revelations outside Arabia for its religious antecedents, explicitly rejecting as idols the local Arabian deities, and as superstition many of their religious practices. Classical Arab authors referred to what preceded the revelation of the Qur'ān as *jāhilīyah* ("ignorance," or, better, "barbarism"). Traditional Muslims, in short, have held Islām to be discontinuous with its Arabian past.

Some modern Western scholars, on the other hand, have attempted to see in pagan Arabian religion not only the antecedents of Islāmic religious practices but also a very early, if not original, strain of Semitic religion. The late-19th-century attempt to locate antecedent "causes" of biblical religion in the Near East promised to be nearly as threatening to Christian and Jewish orthodoxy as the location of antecedents of Islām appeared to be for that faith. More recent scholarship, however, has sidestepped such theological issues, and in historical presentations, Arabian developments have also appeared more complex than earlier they had been thought to be.

Sources. Documents from within ancient Arabia are fragmentary. Scholars must depend heavily, therefore, on other cultures' records of contact with Arabia. The earliest appear in the cuneiform annals of ancient Mesopotamia in certain scattered personal names, Arabic in form, in the 3rd millennium BC, and then in names of Arabian tribes, kingdoms, and even deities in the 1st millennium BC. A few Egyptian texts make similar mention of Arabic names.

Literary sources

Several biblical genealogies, reflecting Israelite tribal traditions before 1000 BC, record the Hebrews' sense of their close kinship and rivalry with the peoples of the desert fringe of Arabia. The account of a visit to King Solomon (died 922 BC) by an Arabian queen, reportedly Sabaean, suggests Israelite contact with the hinterland during the period of the Hebrew kingdoms.

The Greeks came to Arabia as traders, the Romans to its northern perimeter as both traders and rulers. Of the several classical authors (Arrian, Strabo, Origen, and others) who discuss Arabia, the 5th-century-BC Greek historian Herodotus is of chief interest because of his readiness to identify Arabian deities with the Greek gods of Olympus—both as to their function in the pantheon and as to their associations with the planets. In later Roman imperial times, southern Arabia was mentioned little, but northern Arabia continued to figure in literature, primarily in Syriac and Byzantine writings.

Poetry from central Arabia in the pre-Islāmic era was later recorded by Muslims, possibly with a certain degree of expurgation of pagan references; religion appears incidental in most of these pre-Islāmic odes, so that in some cases, for example, it is unclear to which community a poet belonged—pagan, Christian, or Jewish. Certain historians and geographers (Ibn Hishām, aṭ-Ṭabarī, Yāqūt, and especially al-Ḥamdānī in his *Iklīl*) wrote about pre-Islāmic affairs. Also significant in the Arabic sources are the references to, and repudiation of, polytheism in the Qur'ān itself and the treatise of the Muslim author Ibn al-Kalbī (early 9th century AD) known as *Kitāb al-aṣnām* ("Book of Idols"). Finally, an uncertain but suggestive nonliterary source for Arabian culture is the survival of ancient customs and traditions among the Bedouin.

The test of historical reconstructions based on any of the foregoing sources is the inscriptional and archaeological material of the pre-Islāmic Arabs themselves. Since the 1840s, thousands of inscriptions and graffiti have been recorded from the entire length of the Arabian Peninsula. These are in several dialects of Arabic, in addition to inscriptions by Arabs in the related Semitic languages known as Nabataean and Palmyrene. Most of the inscriptions consist only of the names and genealogies of their authors, with occasional prayers, drawings, or references to everyday activities—insignificant individually but instructive in the aggregate.

Inscriptions

Only a few archaeologists have had the opportunity to excavate in Arabia in the 20th century; a few sanctuaries are known as a result of their work. Numismatic evidence for Arabian religion is slender, since Arabian coinage for a long time copied the well-known Athenian owl with little variety in symbols or legends.

History of pre-Islāmic Arabia. *South Arabia.* The kingdoms of South Arabia, known to have flourished from the 1st millennium BC, included the kingdoms of Maʿīn, Saba', Qatabān, Awsān, Ḥaḍramawt, and Himyar. With a decline in economic fortunes after the 1st century BC because of changes of Greco-Roman trade routes to India, the centre of population and cultural influence shifted northward from the Yemen to the Hejaz, which was ruled in the early Christian centuries by Ethiopia. Though the religion of South Arabia was polytheistic, the monotheistic communities of Jews and Christians in the Hejaz gained in influence. The break of the Ma'rib Dam in the Yemen (*c.* AD 575), which brought about the collapse of irrigation in the area, was an event which later Islāmic tradition saw as dramatizing the eclipse of ancient South Arabian culture.

Northern and central Arabia. The northern and central Arabian Peninsula contained both cities, at oases and on the desert perimeter, and nomadic populations. The name "Arabs" appears in historical sources during the 1st millennium BC; the Assyrian Shalmaneser III mentions "Gindibu the Arab" in an inscription commemorating the battle of Qarqar in Syria (854 BC). In an inscription of the Assyrian king Tiglath-pileser III (died 727 BC) a kingdom of "Aribi" is mentioned. On the desert fringe of the Persian, Greek, and Roman empires, the caravan and oasis cities prospered. Most notable of these were Petra, the capital of the Nabateans, which flourished in the 1st centuries BC and AD; and the oasis of Palmyra, northeast of Damascus, which reached its zenith in the 3rd century AD. Arab kingdoms also arose on the frontiers of Persia: Characene in lower Mesopotamia and Mesene (Gerrha) on the Persian Gulf. From the 3rd to the 6th centuries the Lakhmid dynasty ruled al-Ḥīrah, a buffer state of the Sāsānian Persians and a centre of Arab Christianity, while the rival Byzantines sponsored the Ghassānid dynasty in 6th-century Syria.

Early appearance of Arabs

Pre-Islāmic deities. Pre-Islāmic Arabian religion is commonly understood to be polytheistic. Of the many Arabian deities, only some of the more important will be mentioned here.

South Arabian deities. In the official cults of the South Arabian kingdoms, the devotees venerated most highly a

triad of deities that were astral in character: the moon god, the sun goddess, and the god equated with the planet Venus. Each of these deities bore a variety of names, depending on the region, or on a particular attribute of the divinity. Chief among the triad was the moon god, who was the protector of the principal cities. The people of the various kingdoms and areas referred to themselves as his offspring, each under a different name: the Sabaeans were the children of Ilumquh ("God Is Power"), the Minaeans the children of Wadd ("Love"), the Qatabānians the children of ʿAmm ("Uncle"), and the people of Ḥadramawt the offspring of Sin (the name of the moon god in ancient Babylonia). In each region other names of the moon god appear, derived from aspects of the lunar cycle or other attributes. Next among the triad was Venus, the morning and evening star, named ʿAthtar, who also had a variety of attributes. Third was the sun goddess, whose principal name, Shams, was common to the various kingdoms, like ʿAthtar, but whose paired epithets, describing contrasting aspects, varied locally.

Astral deities

Despite the prominence of the name elsewhere among Semitic peoples, the god Il (El) appears to play a comparatively minor role in the South Arabian inscriptions. Some modern scholars have sought to explain this circumstance by equating Il with the moon god, but this opinion has not prevailed.

The remaining list of South Arabian deities is long, and many of them appear to have had a more particularized function than that of the major triad. Some were guardians of clans or of places, and two Qatabānian deities watched over boundaries and irrigation, respectively. Other deities' names, to cite only a few, were attributes, such as Yithaʿ ("Saviour"), Nasrum ("Eagle"), Raʿat ("He Who Instills Fear"), Dhū ʿAwdhān ("He Who Preserves"), and Mutībaqabt ("He Who Guarantees the Harvest").

North Arabian deities. Among the peoples around the northern perimeter of Arabia, "god," in the most generic sense, was El, or in a longer form of the same name, Ilāh. His veneration at a very early stage is attested by his appearance in theophoric names, that is, personal names of which one element is a divine name (the biblical name Gabriel is an example). Among nomadic tribes in particular, a residual sense of El as being the god par excellence remained until the time of Islām.

El, or Ilāh, and Bel

Astral or local deities, however, tended to displace El in the Nabataean and Palmyrene kingdoms. Although El was preserved in early Nabataean theophoric compounds, in Palmyra a more central place in the cult went to Bel (Baal, "Lord"), and in both Petra and Palmyra to Belshamin ("Lord of the Heavens"). With Bel, sometimes in a triad, the Palmyrenes associated Yarhibol, a solar deity, and Aglibol, a lunar deity; while Belshamin stood in a triadic relationship with the gods Malakbel, also a solar deity, and Aglibol.

Al-Lāt, Al-ʿUzzā, and Manāt. Among the Qur'ān's references to its 7th-century pagan milieu are three goddesses, called daughters of Allāh: al-Lāt, al-ʿUzzā, and Manāt; these are also known from earlier inscriptions in northern Arabia. Al-Lāt ("the Goddess") may have had a role subordinate to that of El (Ilāh), as "daughter" rather than consort, but at aṭ-Ṭā'if and a number of other sites in northern Arabia she is mentioned as al-Lāt-of-a-particular-place, as the local deity. At Palmyra she was equated with Athena. As for her two partners in the Qur'ānic triad, the goddess al-ʿUzzā ("Strong") was known among the Nabataeans, while Manāt ("Fate") was associated at Palmyra with the Greek Nemesis. Another principal goddess was Ruḍā, whose name is the feminine form of Arso, a deity whom Herodotus had mentioned in the 5th century BC along with al-Lāt as the sole recipients of worship among the Arabs.

Other indigenous deities. Other indigenous deities included the Nabataean Dhū-Sharā and Shaiʿ-ha-Qawm ("Protector of the People"), who also were invoked among neighbouring nomads. Local protectors were the deities named Gad ("Fortune"), often specified as the Gad of a particular place or, especially among the nomads, of a particular tribe.

The legacy of earlier Canaanite and Babylonian culture was noticeable on the northern perimeter of Arabia, particularly in Palmyra. There, the list of deities named or portrayed included the Babylonian gods Nergal (associated with Heracles) and Nabu or Nebo (associated with Apollo), the Syrian goddess Atargatis, and possibly the Babylonian goddesses Astarte (Ishtar) and Beltis (Beltu, feminine equivalent of Bel).

Religious objects, practices, and institutions. *Sacred stones.* A principal sacred object in Arabian religion was the stone, either a rock outcropping or a large boulder, often a rectangular or irregular black basaltic stone without representative sculptural detail. Such stones were thought to be the residences of a god—hence the term for them employed by Byzantine Christian writers in the 5th and 6th centuries: baetyl, from *bet 'el*, "house of the god."

The baetyls

Of the numerous baetyls, the best known is the Black Stone of the Kaʿbah at Mecca, which became the central shrine object of Islām. But there were others: Ibn al-Kalbī wrote of the square stone central to the cult of al-Lāt at at-Tā'if. Suidas, the compiler of a Greek lexicon and encyclopaedia of *c.* AD 1000, described the baetyl of Dhū-Sharā at Petra as a rectangular black stone on a gold base. There were a number of baetyls at Ramm, southeast of Petra, of which the baetyl of al-Lāt-of-Bostra was in the form of a pedestal.

Just how sophisticated the Arabian peoples were in distinguishing the image from the deity is a matter of evaluative judgment. Perhaps the pedestal form indicated that the baetyl was the dwelling or throne of the god; a Nabataean term meaning "seat" apparently implies so, as does the etymology of baetyl itself, while the South Arabian term "the seat of the deities" may also imply the location of their images in the temple. Some sources, however, associate the name of the deity with a word for stone or baetyl; some South Arabic texts refer to ʿAthtar as Hagar, "Stone." And a particular baetyl, or the whole form as such, might have been venerated in and for itself; northern Syrians were reported to have venerated "Symbaetylos Leon"; *i.e.*, baetyl and lion in association. An inscription on an altar at Doura-Europus (Dura-Europos) on the Euphrates indicates that it was dedicated by a Syrian soldier to the ancestral god Zeus-Baetyl.

Sanctuaries. Characteristically, a Semitic sanctuary was a rectangular enclosure known as a *ḥaram*, "reserved space." Portions of it could be covered or roofed over with porticoes. An altar was central. There often was a well or cistern with water for ablutions, and a sacred tree on which might be hung offerings of visitors or trophies of war. The *ḥaram* served as a place of asylum for all living beings, including both men and animals. The trees growing in the area were likewise inviolate. The larger temples that have been excavated had many rooms, some denoted by words that imply kneeling, others implying burning. In small shrines, and particularly in household usage, one room might suffice for all such activity

In North Arabian temples the image of the deity sometimes stood in the open air, or could be sheltered in a *qubbah*, a vaulted niche. Such a niche might be portable; a portable shelter is represented graphically on a Palmyrene relief. Not to be confused with the *qubbah* is the word *kaʿbah*, for a cube-shaped walled structure, which, though not portable, was constructed possibly after the shape of tents and served as a shelter for the sacred stones.

Considerable attention was given in Arabia to funerary practices and structures. Tombs were often as substantially built as temples, even more substantially in the case of the Nabataeans, who hollowed their tomb-facades out of the solid sandstone of the cliffs surrounding Petra. A standing stone, erected over the tomb, or at Palmyra often a mausoleum tower, was called a *nefesh* ("self" or "soul"), and symbolized the presence of the departed soul.

Ceremonies and customs. A principal public celebration of the Arabians was an annual pilgrimage, in which

Pilgrimages

tribes who shared a common bond of worship of a deity at a specific sanctuary would reunite there. A pattern of ceremonial procession around the baetyl was common, and this pattern may be seen in the surviving Islāmic custom of the pilgrimage to Mecca. Processions played a great part in ritual, and divine images were sometimes brought out of the sanctuary and carried in them.

Another practice that left its influence on Islām was ceremonial abstinence. The South Arabian festival of Ḥalfān was a moratorium on the use of weapons. Certain times were specified for fasting and abstention from sexual relations.

Burnt offerings and sacrifices were common. Some inscriptions mention numbers of animals sacrificed as high as 30 to 40. Sacrifices sometimes took place in a pilgrimage context, following the ceremonial procession around the baetyl; the blood of the animal victim, or in some instances milk as a substitute, was placed on the altar or baetyl. Incense and libations also were used.

Ritual banquets

At public ceremonies in South Arabia the congregation stood, except for sacrificial banquets, when benches were used. Ritual banquets occurred in the northern areas, too. The Palmyrenes celebrated them reclining on couches, as depicted in their tomb reliefs; and "in the presence of the god," according to one text. Stone-carved benches suitable for banquets have been found at Petra and at other Nabataean sanctuaries.

Individual devotions included worship of household gods, and devotions to the astral deities were performed on the roof terraces of South Arabia. In many cases individuals made votive inscriptions, accompanied by drawings or symbols; some of these inscriptions are stylized in form, suggesting that the public ceremonial also followed a liturgy, though its text is not preserved. Among the many Bedouin graffiti in the northern regions are prayers that the deity will grant health and safety, booty and prosperity, and vengeance on enemies, or will blind anyone who effaces the inscription. South Arabian inscriptions likewise indicate concern for personal and public prosperity, including thanksgiving, but also seek expiation for transgressions, many of them ritual. People sought answers by means of divination through the use of dice, oracles, dreams, and visions. Amulets also were used, some of them not only bearing designs but naming the deity.

Religious personages and institutions. In northern Arabia, certain tribes looked after particular sanctuaries. In the role of custodian, a man would be known as a *kāhin*, "priest," but such personnel were not, as far as is known, set apart by ordination. In some circumstances, the head of a family or the chief of the tribe served as the chief officiant in religious functions; in cities such as Palmyra, it is more likely that the priestly function was carried out by specialized personnel.

The South Arabian temples had three ranks of priests and assistants. The interrelation of the priestly with the political organization in the ancient kingdoms can be noted in the title *mukarrib*s (probably "procurators"), of Saba' who were among the first to develop a centralized political power.

As an institution, the temple received gifts both in money and in kind, including harvest offerings and tithes. The gifts in kind might be buildings, lands (which would yield revenue), slaves, animals, images and furnishings, spices and incense, and other supplies. Among the temple's functions was included a legal one: because the gods were witness to judgments and contracts, the temple was the repository of legal and regulatory documents.

Monotheism in Arabia. Despite all that has been said, present scholarly knowledge of ancient Arabia remains fragmentary and is based on arguments that fill in the unknown by an assumption of continuity with the known. These continuities often cut across long time spans and depend on assumed constant patterns of religious beliefs and practices among the Semitic peoples—an assumption that may ignore the tracing of developments, difficult when inscriptions are not easily datable.

Pre-Islāmic Arabian monotheism. Yet on one topic, namely monotheism, the views of many scholars show a marked interest in developmental change. This topic arises in two contexts: the possibility of an original monotheism anterior to Arabian polytheism, and the possibility of a monotheism as a late development out of Arabian polytheism.

Arguments for an original monotheism are based largely on the ubiquity of the word *īl, ilāh*, for "god" in Semitic languages. Each tribe started with its own single god, it has been argued, and a polytheistic pantheon was derived from a situation where tribes, failing to identify others' deities with their own, expanded the list. Whatever the speculative possibilities of such a theory, the Arabian situation was polytheistic at the time of the first surviving records of it.

External monotheistic influences

More fascinating and more tangible are the indications that in the last few pre-Islāmic centuries an Arabian monotheism developed. There were, of course, monotheistic influences from outside Arabia. Christian missionaries and traders from Syria, Mesopotamia, and Ethiopia brought a variety of teachings, and a Christian settlement flourished at Najrān near the Yemen. There were Jewish communities at Ṣan'ā' in South Arabia, and at Khaybar, Yathrib (Medina), and Taymā' farther north. In the 4th and 5th centuries AD, a few Sabaean inscriptions appear to have ascribed all supreme functions to Ilāh, calling him Merciful, and Lord of Heaven or Earth. One text calls him the Great One of Judah, and another speaks of his son Christ the Victorious. Other traditions were also in a position to have influence in Arabia, notably Zoroastrianism and Manichaeism in Mesopotamia. Despite the possibility of such influence, it has not so far been demonstrated in the inscriptions.

The ḥanīf. It may well have been the presence of Christians and Jews that set the stage in the Hejaz for the so-called *ḥanīf*s, men who found the old cult of the gods inadequate to their level of spirituality and monotheistic ideals, and who repudiated what they could not associate with Ilāh. A couple of Muḥammad's relatives, and early supporters, were reported to have been *ḥanīf*s. *Ḥanīf*s were not Christian; the word *ḥanīf* appears to be a Christian Syriac loanword meaning "heathen." But they were evidently monotheists, and, coming on the eve of Islām, they mark the passing of the old Arabian religion.

BIBLIOGRAPHY. GIORGIO LEVI DELLA VIDA, "Pre-Islamic Arabia," in N.A. FARIS (ed.), *The Arab Heritage*, pp. 25–57 (1944), is a readable treatment of ancient Arabian culture, with some references to religion. IRFAN SHAHID, "Pre-Islamic Arabia," in P.M. HOLT (ed.), *The Cambridge History of Islam*, vol. 1, pp. 3–29 (1970), stresses the displacement of the older cults by Christianity and Judaism. For further historical discussion, see ALBERT DIETRICH, "Geschichte Arabiens vor dem Islam," in B. SPULER (ed.), *Handbuch der Orientalistik*, vol. 2, sect. 4, pt. 2 (1966). Works on ancient Arabian religion, giving a review of the divinities and their functions, are based primarily on inscriptional material. Taking the simplest first: GONZAGUE RYCKMANS, "Les Religions arabes préislamiques," in M.M. GORCE and R. MORTIER (eds.), *Histoire générale des religions*, vol. 2, pp. 220–228, 593–605 (1960); JEAN STARCKY, "Palmyréniens, Nabatéens, et Arabes du nord avant l'Islam," and ALBERT JAMME, "La Religion sud-arabe pré-islamique," in M. BRILLANT and R. AIGRAIN (eds.), *Histoire des religions*, vol. 4, pp. 201–237, 239–307 (1956); and MARIA HOFNER, "Die vorislamischen Religionen Arabiens," in H. GESE, MARIA HOFNER, and K. RUDOLPH, *Die Religionen Altsyriens, Altarabiens und der Mandäer*, pp. 233–402 (1970). Extensive bibliographies may be found in these works and in YOUAKIM MOUBARAC, "Éléments de bibliographie sud-sémitique," *Revue des Études Islamiques*, 23:121–175 (1955) and 25:13–68 (1957). Readers limited to English may find the interpretations of JEAN STARCKY and ALBERT JAMME conveniently summarized in the article "Arabia," in the *New Catholic Encyclopedia*, vol. 1 (1967).

(W.G.O.)

Arabian Sea

The Arabian Sea lies in the northwestern section of the Indian Ocean. Situated within the trade wind latitudes, it forms part of the principal sea route between Europe and India. It is bounded to the east by India, to the north by Pakistan and Iran, to the west by the Arabian Peninsula and the Horn of Africa, and to the south

by the remainder of the Indian Ocean. To the north the Gulf of Oman connects the sea with the Persian Gulf via the Strait of Hormuz. To the west the Gulf of Aden connects it with the Red Sea via the Strait of Bāb el-Mandeb. Its total area is about 1,490,000 square miles (3,859,100 square kilometres); it has a mean depth of 8,750 feet (2,734 metres). In Roman times its name was the Mare Erythraeum (the Erythraean Sea).

Political units bordering the sea—apart from India, Iran, and Pakistan—are the Sultanate of Oman, Yemen (Aden), and Somalia. Islands in the sea include Socotra (a part of the Yemen [Aden]) off the Horn of Africa, the Kuria Muria Islands off the coast of Oman, and the Laccadive Islands (a part of India), the latter of which are a group of coral atolls lying between 100 and 250 miles off the southwest (Malabar) coast of India. The Indus River is the principal river draining into the sea. (For associated physical features, see ADEN, GULF OF; INDIAN OCEAN; INDUS RIVER; PERSIAN GULF; and RED SEA.)

Geography. Most of the Arabian Sea has depths that exceed 9,600 feet, and there are no islands in the middle. Deep water reaches close to the bordering lands except in the northeast, off Pakistan and India. To the southeast the Laccadive Islands form part of the submarine Maldive Ridge, which extends farther south into the Indian Ocean where it rises above the surface to form the Maldive Islands. On the western side of the sea, the plateau island of Socotra, about 70 miles long and with an area of about 1,200 square miles, is an insular extension of the Horn of Africa, lying 160 miles east of Cape Guardafui.

The Carlsberg Ridge

Submarine morphology and geology. Stretching southeastward from Socotra is the submarine Carlsberg Ridge, which coincides with the belt of seismic activity in the Indian Ocean that divides the Arabian Sea into two major basins—the Arabian Basin to the east and the Somali Basin to the west. The deepest soundings, 15,900 feet, have been recorded in the Somali Basin. The Carlsberg Ridge is longitudinally split by a central valley that reaches depths of 10,800 feet below the sea's surface. The coastal escarpments of the Gulf of Aden are formed by rift faults that converge toward the southwest to continue into Africa as the boundary scarps of the Ethiopian Rift Valley, which forms part of the East African Rift Valley.

The Arabian Basin is separated from the Gulf of Oman Basin by the Murray Ridge, a narrow submarine ridge that extends northeast to southwest to meet the Carlsberg Ridge.

A deep submarine canyon has been discovered off the mouth of the Indus River, while an abyssal cone (a cone in the lowest depths of the ocean) and an associated abyssal plain occupy much of the northeast floor of the Arabian Sea. To the east of the Somali coast is another large abyssal plain.

The coastal shelf is narrow along the coast of the Arabian Peninsula and is even narrower along the Somali coast. No true coral reefs are found along the Arabian coast. Sediments near Ra's al-Ḥadd (the easternmost headland of the Arabian Peninsula), where upwelling of deep water occurs in summer, consist of fine greenish mud with a high organic content containing hydrogen sulfide. The region, which contains many fish remains, is known as a fish cemetery. Deposits derived from the land cover the major part of the continental slope bordering the sea, down to a depth of about 8,250 feet. Below this, deposits consist of *Globigerina* (a genus of microscopic animals with calcareous shells belonging to a group known as Foraminifera), while basins lying below 12,000 feet are covered by red clay. Manganese nodules have been discovered in the midst of foraminiferal ooze; a single dredge obtained as much as 275 pounds of these nodules. Sediment thickness decreases from 8,200 feet in the north to about 1,600 feet in the south of the Arabian Basin.

The Arabian Sea is believed to have been formed during the Mesozoic and Cenozoic eras (*i.e.*, within the last 225,000,000 years). Geophysical studies conducted in the Gulf of Aden suggest that the continental blocks of Africa and Arabia separated, with new oceanic crust forming in between, in the pre-Miocene epoch (*i.e.*, more than 26,000,000 years ago).

Climate and hydrography. Minimum surface temperatures of about 75° to 77° F (24° to 25° C) occur in the central Arabian Sea in January and February, while temperatures higher than 82° F (28° C) occur both in June and in November. During the rainy season, which occurs when the southwest monsoon winds blow, from April to November, salinities of less than 35 parts per 1,000 have been recorded in the upper 150 feet of the sea, while during the dry season from November to March, when the northeast monsoon winds blow, salinities of more than 36 parts per 1,000 have been recorded at the surface over the entire Arabian Sea north of latitude 5° north, except off the Somali coast.

The Somali Current

The complex Somali Current, which attains speeds of about seven ksots (nautical miles an hour) off the coast of Socotra, becomes part of a clockwise circulation system that continues to the northeast along the coast of Arabia and thence south along the coast of India to 10° north. At this point it merges with the southwestern drift current, flowing east between 5° and 10° north, that is associated with the southwest monsoon winds. Pronounced upwelling of deeper waters occurs along the Somali and Arabian coasts in summer. Of the five water masses that have been distinguished in the upper 3,000 feet of the north Indian Ocean, three have been identified as originating in the Red Sea, the Persian Gulf, and the Arabian Sea, respectively. The paths of flow of these water masses are to the south and east.

Economic resources. High levels of inorganic nutrients, such as phosphate concentrate, which produce a rich fish life, have been observed in the western Arabian Sea and off the southeastern coast of the Arabian Peninsula. Occurring in the euphotic zone (zone of light, which is found in the upper 450 feet of the sea), this concentration of nutrients is undoubtedly due in part to coastal upwelling, which plays a major role in fertilizing the water.

Pelagic fish (*i.e.*, those living at or near the surface far from the land) in the Arabian Sea include tuna, sardine, billfish (a species having a long sharp bill or snout), wahoo (a large, swift game fish), sharks, lancet fish (a large species having dagger-like teeth), and moonfish (a species of silvery fish with very compressed bodies).

A periodic phenomenon in the Arabian Sea, however, is mass mortality of fish. In 1957, for example, in an area of approximately 77,000 square miles about 20,000,000 tons of fish were believed to have perished. Fish mortality is attributed to a subsurface layer of water of tropical origin that is poor in oxygen but rich in phosphate; under certain circumstances this layer is brought to the surface by strong upwelling, resulting in the death of fish from lack of oxygen.

Navigational history. To medieval Arabs the Arabian Sea was known as the Sea of India or as part of the "Great Sea," from which smaller gulfs such as the Sea of Faris (Persian Gulf) or Sea of Kolzum (Red Sea) were distinguished. From about the 8th or 9th centuries onward, Arab and Persian seafarers learned to use the surface currents propelled by the summer and winter monsoon winds. Detailed navigational instructions for sailing between southern Arabian, East African, and Red Sea ports, as well as ports in India, Malaysia, and China, were written down by pilots from Oman and the Ḥaḍramawt region of southern Arabia between the 9th and 15th centuries. Some of these works, entitled in Persian *rahmangs* (books of routes), contain useful information on navigating by the stars, on winds, currents, soundings, descriptions of coasts, approaches, and islands. Among flourishing medieval ports mentioned in these works are Diu and Surat in India, Hormuz in Persia, and Muscat and Aden on the Arabian Peninsula. Landmarks mentioned included Ra's (cape) al-Ḥadd and Ra's Madrak—capes on the southeast coast of the Arabian peninsula—and Cape Guardafui in Somalia, which is the cape of the Horn of Africa.

The early Arabian and Persian navigators

In recent times the Arabian Sea has been studied by several oceanographic expeditions, of which the most important has been the Mabahiss, or John Murray, Expedi-

tion of 1933 to 1934, which reported findings concerning hydrography, chemistry, currents, water masses, bottom topography, and sediments. Further information was obtained during the International Indian Ocean Programme (1960–65) in which British, U.S., Soviet, and German ships participated, studying currents, productivity, seismology, and geology.

Prospects for the future. Some of the scientific and economic data from the International Indian Ocean Programme have yet to be evaluated; further information, especially on submarine geology and the world rift system, will, it is expected, eventually be available.

Economically, fishing operations along the Arabian and Somali coasts have yet to be fully developed; the magnitude of the unexploited fish resources is indicated by the periodic mass mortalities that have occurred. The manganese nodules located on the bed of the Arabian Sea may also prove to be an exploitable mineral resource.

BIBLIOGRAPHY

General and historical: Scientific Reports, John Murray Expedition (Mabahiss Expedition, 1933–34), vol. 1-10 (1935–1960); *International Indian Ocean Expedition: Collected Reprints*, 6 vol. (1965–69); R.W. FAIRBRIDGE (ed.), *Encyclopedia of Oceanography* (1966); A.A. ALEEM, "Concepts of Currents, Tides and Wind Among Medieval Arab Geographers in the Indian Ocean," *Deep Sea Res.*, 14:459–463 (1967); and "Ahmad ibn Magid, Arab Navigator of the XV Century and His Contributions to Marine Sciences," *Bull. Inst. Oceanogr. Monaco*, Special No. 2, pp. 565–580 (1968).

Geology and hydrography: N.K. PANIKKAR and R. JAYARAMAN, "Biological and Oceanographic Differences Between the Arabian Sea and the Bay of Bengal as Observed from the Indian Region," *Proc. Indian Acad. Sci.*, 64:231–240 (1966); D.J. ROCHFORD, "Salinity Maxima in the Upper 1000 m of the Indian Ocean," *Aust. J. Mar. Freshwat. Res.*, 15:1–24 (1964); H. STOMMEL and W.S. WOOSTER, Reconnaissance of the Somali Current During the Southwest Monsoon," *Proc. Natn. Acad. Sci. U.S.A.*, 54:8–13 (1965).

Economic resources: J.H. RYTHER *et al.*, "Primary Organic Production in Relation to the Chemistry and Hydrography of the Western Indian Ocean," *Limnol. Oceanogr.*, 11:371–380 (1966).

(An.A.A.)

Arachnida

The class Arachnida, part of the invertebrate phylum Arthropoda, numbers about 65,000 species and includes forms such as spiders (Araneida), harvestmen or daddy longlegs (Opiliones), scorpions (Scorpionida), mites and ticks (Acarina), and a few lesser known groups. Only a few species are economically important, but many arachnids are conspicuous and exhibit interesting behaviour.

General features. *Structure and appearance.* The arachnids are characterized by segmented bodies, well-developed heads, a hardened (chitinized) outer (exo-) skeleton, and paired jointed appendages. Growth in size occurs by a series of molts, during which the old exoskeleton is shed as a larger one forms. The body is divided into two main parts, the cephalothorax, or prosoma (roughly equivalent to head and thorax of insects), and the abdomen, or opisthosoma. The cephalothorax has six pairs of appendages, the first of which, the chelicerae, are usually adapted for grasping (prehensile); among spiders, the second chelicerate segment is used to inject venom. The second pair of appendages, the pedipalps, may be leglike tactile organs (called palps) or large, pincerlike (chelate), predatory organs. A portion of the basal segment of each pedipalp may be modified into a cutting or crushing organ and function with the mouthparts during feeding; among spiders, the distal palpal segment is used as a special copulatory organ. Individual arachnids are variable in both size and appearance. In size, they range from tiny mites 0.1 millimetre long to the enormous black scorpion (*Pandinus imperator*) of Africa, which is about 18 centimetres (7 inches) long. Arachnids vary from small, rounded forms to long-legged runners, crablike spiders, and fast-running predators.

Size range

Distribution and abundance. The eleven living orders of arachnids may be organized into three large groups, depending on their distribution. Araneida (spiders), Opiliones (harvestmen, or daddy longlegs), Pseudoscorpiones (false scorpions), and Acarina (mites and ticks) are found throughout the world. Encountered infrequently in northern areas, but widespread within the tropical and subtropical regions of the world, are the Scorpionida (scorpions), Solifugae (solpugids, or sunspiders), Amblypygi (tailless whip scorpions), and Uropygi (vinegarroons, or whip scorpions). Of sporadic distribution are Palpigradi (palpigrades, or micro-whip scorpions), Ricinulei (ricinuleids), and Schizomida (schizomids).

With the exception of a few groups that have secondarily become aquatic, the arachnids are terrestrial, carnivorous predators. In late summer and early fall, mature spiders and harvestmen are conspicuous members of the invertebrate fauna. Even though they are most evident then, arachnids are extremely abundant throughout the year. Most, however, are seldom observed, for they inhabit leaf mold or soil and form an important part of the cryptozoic fauna; *i.e.*, animals that live in crevices or under objects. Most abundant of the arachnids are the acarines found in soil, in fresh and marine waters, and as parasites of animals, including man. In warm areas, scorpions, solpugids, and whip scorpions are numerous. Tiny false scorpions, micro-whip scorpions, and ricinuleids, sometimes considered rare, may be abundant under favourable circumstances.

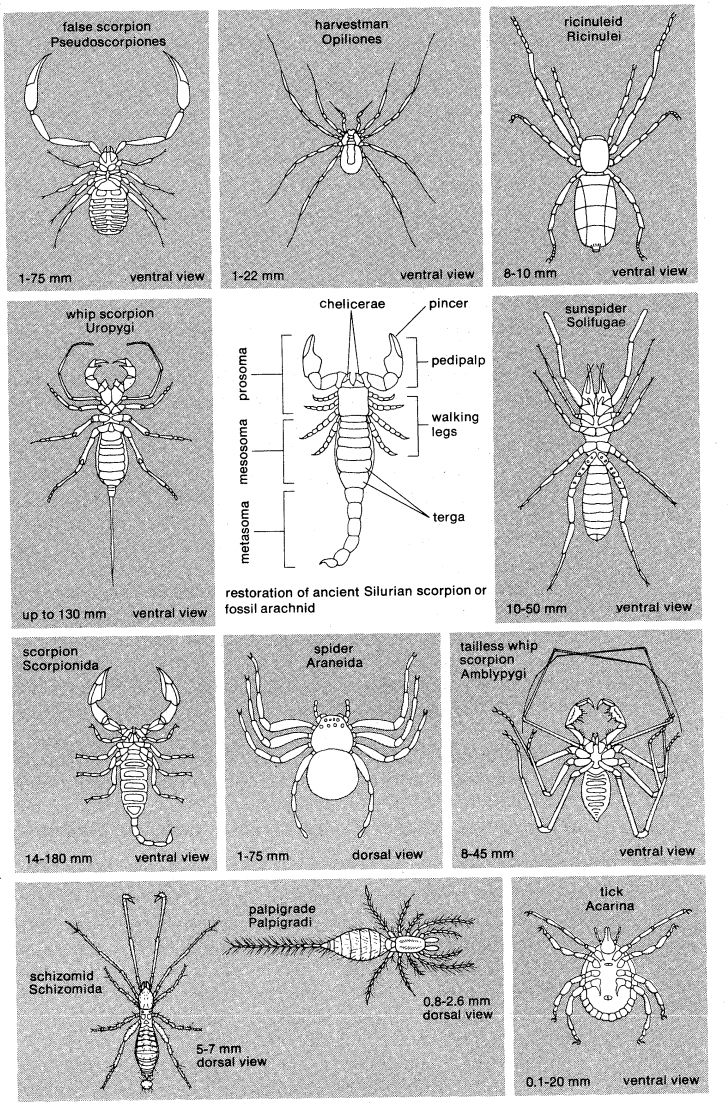


Figure 1: Diversity of arachnids.

From (restoration) R.R. Shrock and W.H Twenhofel, *Principles of Invertebrate Paleontology*, copyright 1953, used with permission of McGraw-Hill Book Company; (Araneida) *Invertebrate Identification Manual* by Richard A. Pimentel, © 1967 by Litton Educational Publishing, Inc., reprinted by permission of Van Nostrand Reinhold Company; (Ricinulei, Uropygi, Solifugae, Scorpionida, Amblypygi) *Transactions of the Connecticut Academy of Arts and Sciences*, vol. XXXI (1933); (Schizomid) after *American Museum Novitates*, The American Museum of Natural History, New York City (1957), in A. Kaestner, *Invertebrate Zoology*, John Wiley & Sons, Inc.

Importance. The most economically important arachnids are the mites and ticks (Acarina). Many of the latter are intermediate hosts of serious diseases; *e.g.*, relapsing fever, African tick fever, and Rocky Mountain spotted fever. Common mange of cattle, horses, and dogs is caused by a mite; plants and stored grains also are attacked by mites. The importance of other arachnid groups cannot be assessed accurately. Since arachnids are predators, they destroy large numbers of smaller invertebrates, but the total effect of this predation has not yet been established with certainty. In general, most arachnids probably are neither particularly harmful nor beneficial to man.

Natural history. *Reproduction and life cycle.* The courtship and mating behaviour of arachnids is of interest. Although generalizations are difficult, the male usually does not transfer spermatozoa directly to the female. Among the tailless whip scorpions, false scorpions, scorpions, and schizomids, the male grasps the female with palpal or chelicerate appendages, and the pair moves forward and backward until the male produces the spermatophore, a gelatinous structure with a drop of fluid containing spermatozoa; after the female is led to the spermatophore, the spermatozoa enter her body. The transfer of spermatozoa in spiders often is preceded by a brief courtship, during which the male usually must identify himself to the female in some way in order to avoid being attacked by her. In some species, recognition is based on a chemical sense; among those with good vision, males may display special colouring. Among web builders, the male may touch the web before approaching the female. After recognition occurs, the male transfers spermatozoa to a sac within the female reproductive system called a spermatheca; the eggs are fertilized as they are laid among arachnids that transfer a spermatophore to the female spermatheca.

Mating among harvestmen is casual and may occur whenever a male and female encounter one another; the male has a hardened copulatory organ that is inserted into the genital opening of the female as the partners face each other. Mating in solpugids is more active, occurring at dusk or during the night; during courting the male strokes the female, seizes her, lays her on her side, massages her under surface with his jaws, opens her genital orifice, and forces a mass of spermatozoa into her spermatheca. Reproductive behaviour in mites is highly variable; sperm usually are produced in a spermatophore and transferred to the female either by the chelicerae or by the third pair of legs of the male. The mating behaviour of ricinuleids is not yet understood, but a complex structure on the third pair of legs of the male presumably is used during mating.

Growth and development

Although details of early development are not yet known for all forms, that of egg-laying spiders is considered typical. The two major divisions of the body (prosoma and opisthosoma) become apparent at an early stage in development, and the appendages first appear as knobs; in many arachnids, the organism is wrapped around the yolk, a situation altered by a process termed inversion or reversion, after which hatching usually occurs. In most cases, there is enough yolk to nourish the young through one molt. Maternal care is minimal, although the eggs of some spiders are contained in a cocoon, and those of some phyrnichids, schizomids, thelyphonids, and false scorpions are attached beneath the abdomen.

Scorpions bear living young; the yolky eggs of ovoviviparous forms develop in the ovarian tubules. The eggs of viviparous species, however, which lack yolk, develop in blind sacs of the ovary; each sac has absorptive cells and a tubular appendage to which the embryo is attached and through which nutritive material passes. After hatching, young scorpions crawl to the mother's back and remain there for at least one molt. The eggs of the false scorpion remain in a membranous sac attached to the female's genital opening; development occurs within the sac, and nutritive material is provided by the female.

Growth occurs among arachnids by molting, or ecdysis, at varying intervals. In many species, the first molt occurs while the animal is still within the egg. The newly hatched arachnid, small and less sclerotized (sclerotization is a hardening process analogous to the tanning of leather) than the adult, is, however, recognizable as an arachnid. With the exception of acarines and ricinuleids, which have three pairs of legs when hatched, most young arachnids have four pairs. In most orders, young are not considered larval forms. The number of molts required to attain maturity varies; some spiders, for example, have 12. Before molting, arachnids usually seek a protected site; spiders, false scorpions, and some mites produce a cocoon to protect themselves. Most arachnids secrete a molting fluid between the old and new exoskeletons, and some species regenerate lost limbs during the molting process.

Development and growth among mites differ in some respects from that of other arachnids. The mite life cycle consists of an egg that hatches into the six-legged, or hexapod, larva, which passes through one or several immature (nymphal) stages before becoming an adult. Although most mites lay eggs, some are ovoviviparous; *i.e.*, the female retains the eggs in her body, and the young develop within the egg. If the female dies before the young are born, fully formed larvae usually emerge from the dead female. Acarines also are able to reproduce from unfertilized eggs (parthenogenesis). The life cycle of ticks is similar to that of mites.

The life-span of arachnids in temperate areas apparently is a single season; therefore, they rely on eggs to perpetuate the species from year to year. In warm regions, certain groups (*e.g.*, whip scorpions, tailless whip scorpions, scorpions, solpugids, and tarantulas) appear to live more than a single year; captive tarantulas, in fact, have survived for as long as 20 years.

Behaviour. Arachnid behaviour in searching for food has not yet been thoroughly investigated; many species apparently search in a random way, using tactile organs to recognize suitable prey. Harvestmen, for example, run rapidly over bushes, herbs, and other vegetation; false scorpions move slowly until they encounter prey, which is grasped by the palps and passed to the chelicerae for mastication. The solpugids, inhabitants of dry areas and very active predators, use both tactile and visual responses in prey recognition as they run at random seeking prey; often at night, however, solpugids congregate in areas of high prey density. The American whip scorpion (*Mastigoproctus giganteus*) hunts mostly at night, moving slowly and cautiously with palps extended, touching objects with sensitive first legs; prey is seized rapidly when encountered.

Among many arachnids, especially scorpions and spiders, quiet waiting is the most important aspect of capturing prey. Spiders, most of which use silk in some way, exhibit the widest variation in feeding habits among arachnids. Web-building forms rest quietly beside their webs, waiting for prey to entangle itself; active spiders often have remarkable adaptations for capturing prey.

A few arachnids (*e.g.*, some scorpions, solpugids, spiders) are capable of producing audible rasping sounds by rubbing together horny ridges or other special sound-producing structures. Such sound production may be used to warn enemies or, by males, during courtship.

Preening

Arachnids typically preen themselves, a process that usually consists of cleaning the legs by passing them through the chelicerae. Some tailless whip scorpions have brush-like structures on the pedipalps that are used to clean the appendages; similar structures occur on the legs of certain spiders (family Gnaphosidae), the pedipalps of some schizomids, and the chelicerae of some whip scorpions.

Because arachnid behaviour during courtship, hunting, and, in the case of spiders, web building is complex, it sometimes appears that these invertebrates are very intelligent. Arachnid behaviour, however, is considered to be largely instinctive; the complexities of courtship, for example, are a series of instinctive actions that follow each other in a predetermined succession.

Ecology. Although most arachnids, usually inconspicuously brown or gray in colour, are free-living terrestrial animals, some acarines are parasitic, a few spiders live on

or near water, and some mites are aquatic. Most arachnids lead solitary lives and come together only for mating; such encounters are necessarily brief, however, for the larger female may kill the male if he does not escape rapidly. A few acarines, spiders, and harvestmen are gregarious.

As carnivores, most arachnids feed chiefly upon smaller arthropods; exceptions occur among parasitic acarines and plant-feeding harvestmen. Acarines are nourished principally by fluids obtained either from living animal or plant material or from decaying organic matter. Parasitic forms have mouthparts modified for sucking blood or juice. The feeding habits of micro-whip scorpions are not yet known; large numbers of adult ricinuleids living in Mexican caves feed upon millipede feces and, possibly, other fecal deposits. In most cases, food is partially liquefied before entering the digestive tract; harvestmen apparently are the only arachnids capable of ingesting small particles. Usually either prey is torn into small pieces as digestive fluids flow over it, or a hole is made in the body of the prey, digestive fluids injected, and the digested contents absorbed.

Habitats

Even though they possess a chitinous exoskeleton, most arachnids are subject to damage by desiccation. Many, especially small species, are most abundant in protected niches; *i.e.*, the favourable environment of the relatively constant, moisture-containing microclimate provided by soil litter, burrows, or caves is the reason these species constitute a high proportion of the cryptozoic fauna. Cave-dwelling species often have special cave adaptations; *e.g.*, long extremities, light colour, and no eyes. Most arachnids avoid excessive heat by adopting a cryptozoic habitat; the scorpions, however, adapt to high temperatures by straightening the legs to allow air to circulate more freely. Desert groups (*e.g.*, scorpions, solpugids) tend to be found in protected sites during the hottest parts of the day. Most arachnids remain in or near the areas in which they were hatched; a few spiders and false scorpions, however, may leave their hatching areas. The latter often use their chelicerae to attach themselves to larger animals, chiefly other arthropods; this behaviour, called phoresy, also occurs in a few mites.

Form and function. *External features.* The body of arachnids (except acarines), as noted above, is divided into two parts, the cephalothorax, or prosoma, and the opisthosoma. The prosoma is formed of six somites, or segments, the opisthosoma of 12; the first somite of the opisthosoma may be reduced or absent. The dorsal region of the prosoma is covered by a solid covering called a carapace, the underside by one or more sternal plates or the coxal segments of the appendages. The soft body is protected by an exoskeleton composed of chitin, and the somites are united by a soft membrane. Segmentation is most evident in the opisthosoma of the scorpions, least evident in the acarines. Among acarines and some harvestmen, the prosoma and abdomen have fused into a single region.

The six pairs of appendages, which arise from the prosoma and often are modified for special function, include chelicerae, pedipalps, and four pairs of legs. The number of segments comprising each leg varies, although the legs terminate in claws in most species.

The abdomen may be broadly joined to the cephalothorax (*e.g.*, opilionids, scorpions, false scorpions, and mites) or narrowly, by means of the pedicel, a modified opisthosomatic somite (*e.g.*, palpigrades, tailless whip scorpions, and spiders).

At certain times, one or both sexes, often difficult to distinguish, may develop some distinctive feature; *e.g.*, a special structure, a brighter colour, larger spines, or larger size. The genital orifice of both sexes (except mites) is on the undersurface of the second abdominal segment; in mites it may be dorsal. The sex organs, or gonads, found in the abdomen, may be single or paired and are associated with various ducts and glands.

Sense organs

The three types of sense organs common among arachnids include tactile hairs, simple eyes (ocelli), and slit sense, or lyriform, organs. Specialized structures, perhaps serving as special tactile organs, include malleoli, or

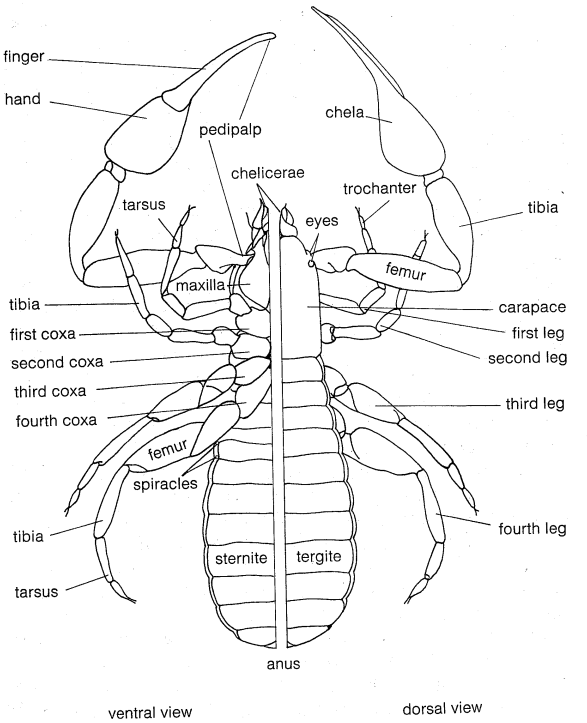


Figure 2: Anatomy of a pseudoscorpion.
From A. Petrunkevitch, *Treatise on Invertebrate Paleontology*, part P (1955); courtesy of the Geological Society of America and the University of Kansas Press

racket organs, of solpugids and comblike appendages (pectines) of scorpions. The number of ocelli, which are found on the carapace, varies; scorpions, for example, may have as many as five pairs of simple eyes on the sides of the carapace in addition to a median pair, while opilionids have only median eyes, and many cryptic (soil or leaf-mold inhabitants) or cave-dwelling species have either reduced eyes or none at all. The most abundant sense organs, tactile hairs, are scattered over the body. The lyriform, or slit sense, organs, which appear as slits in the cuticle, may function in the sense of smell; slit sense organs on the legs of harvestmen, however, function in the reception of internal stimuli (proprioception). Tarsal organs, small, round holes in the upper surface of the last segment of the legs, may act as chemoreceptors.

Functional features. The arachnid nervous system, similar to that of other arthropods in that it consists of a brain and a chain of paired ganglia, has been highly modified by ganglionic fusion and migration forward toward the head region.

The food of most arachnids is partially digested as it is held by the chelicerae. After the prey has been liquefied by digestive enzymes, a pumping action by the hardened pharynx forces the food into the foregut; most digestion and absorption occur in blind sacs extending from the midgut. Undigested materials are excreted through the anus. Three types of excretory organs occur in arachnids, coxal glands, malpighian tubules, and an orifice located on the pedipalpal somite (in Palpigradi and Solifugae). The coxal glands consist of three parts, a large excretory sac lying opposite the coxal segment of the first pair of legs, a long coiled tubule, and a short exit tube that opens to the exterior through orifices behind the first and third coxal leg segments. Although the malpighian tubules of arachnids have a different embryological origin from those of insects, they appear to be similar in function; *i.e.*, excretory granules are formed within them and excreted into the midgut. The nitrogen-containing waste material usually is the organic compound guanine.

Arachnids respire by structures known as book lungs and tracheae. Book lungs are found in hardened pockets; diffusion of gases occurs between the blood circulating within thin, leaflike structures (lamellae) stacked like pages in a book and the air in spaces between these thin structures. The tracheal system, which consists of a num-

ber of tubes that open to the exterior by paired respiratory pores (spiracles), is similar to that of insects; diffusion of gases occurs within small fluid-filled tubes that ramify over the internal organs. Pseudoscorpions, solpugids, harvestmen, ricinuleids, and acarids have only tracheae; scorpions, uropygids, and amblypygids have only book lungs; most spiders have both; and the small palpigrades and acarines have only cutaneous respiration.

The arachnid circulatory system is an open one; that is, blood circulates in tissue sinuses. Special venous channels conduct blood from the sinuses to the heart, from which it is pumped through a series of blood vessels back to the sinuses. The respiratory pigment usually is hemocyanin. Some very small forms lack a circulatory system. Since most conspicuous arachnid movements involve legs, mouthparts, gut, and some glands, most of the musculature is within the prosoma, and the muscles usually are attached to a plate of chitin within the prosoma, called an endosternite.

Glands

The glands of arachnids usually are peculiar to individual orders, making generalizations difficult. The silk glands of spiders, in which silk is stored as a viscous liquid, are at the end of the abdomen, and the number of spinnerets, which are modified abdominal appendages, is variable. The viscous liquid, chemically a protein, passes through minute tubes at the tips of the spinnerets; the change from liquid to thread occurs by a structural rearrangement of the protein molecules after the silk is drawn from a silk droplet. The elastic silk of spiders has considerable strength and is used to snare prey, to build cocoons, and for transportation. The chemical composition of silk produced by false scorpions and some mites is not yet known. The spinneret of false scorpions is located on a movable finger of the chelicerae; and the few mites that produce silk have glands in the mouth area. False scorpions and mites apparently use silk chiefly to construct cocoons.

Toxic substances are secreted by special glands found in the chelicerae of spiders, in the pedipalps of false scorpions, and in the last opisthosomal segment of scorpions. The poisonous substance (and its effects) produced by each group differs sometimes within species of the same order. The strongly alkaline poison of spiders is much less poisonous to mammals than to other arthropods; bites of spiders of the genera *Loxosceles* (brown recluse) and *Latrodectus* (black widows), for example, may cause human discomfort, but rarely death. The venom of most scorpions, sufficiently toxic to kill most invertebrates, usually is not dangerous to man; the venoms of some genera (*Androcthonus* and *Buthus* of Africa, various species of *Centruroides* of Mexico, Arizona, and New Mexico), however, are highly toxic nerve poisons. The venom of false scorpions, probably very toxic, has not yet been studied chemically. The paired glands of whip scorpions (Uropygi), located near the anus, secrete specific acids (formic, acetic), which, although not particularly harmful, serve as irritants and apparently are used for defense. The openings of paired odoriferous glands are found on the prosoma of opilionids; when irritated, these animals secrete a fluid containing cyanic acid that probably serves to repel predators.

A number of endocrine glands have been identified among arachnids; most important, possibly, are those involved in the molting process.

Evolution and paleontology. Although the number of known arachnid fossils is impressive, the evolution of individual groups is difficult to trace. The earliest known forms recognizable as arachnids are scorpions dating from the Silurian Period (about 395,000,000 years ago); even though the Silurian fauna was primarily marine, a generally accepted opinion is that these scorpions were terrestrial, rather than aquatic, and that they possibly were among the earliest animals to leave water for land. All arachnid orders apparently were represented in the Carboniferous (about 345,000,000 years ago), although the palpigrades have been described as fossils only from the Jurassic Period of Europe (about 190,000,000 years ago), and the schizomids are known from the Late Cenozoic Era of Arizona (about 7,000,000 years ago). The Mesozoic (about 65,000,000–225,000,000 years ago) is poor in arachnid fossils; on the other hand, the Cenozoic Era (from about 65,000,000 years ago to the present) is rich in them. Fossil evidence implies that arachnids must have appeared in the Early Paleozoic Era (about 570,000,000 years ago) as a distinct group.

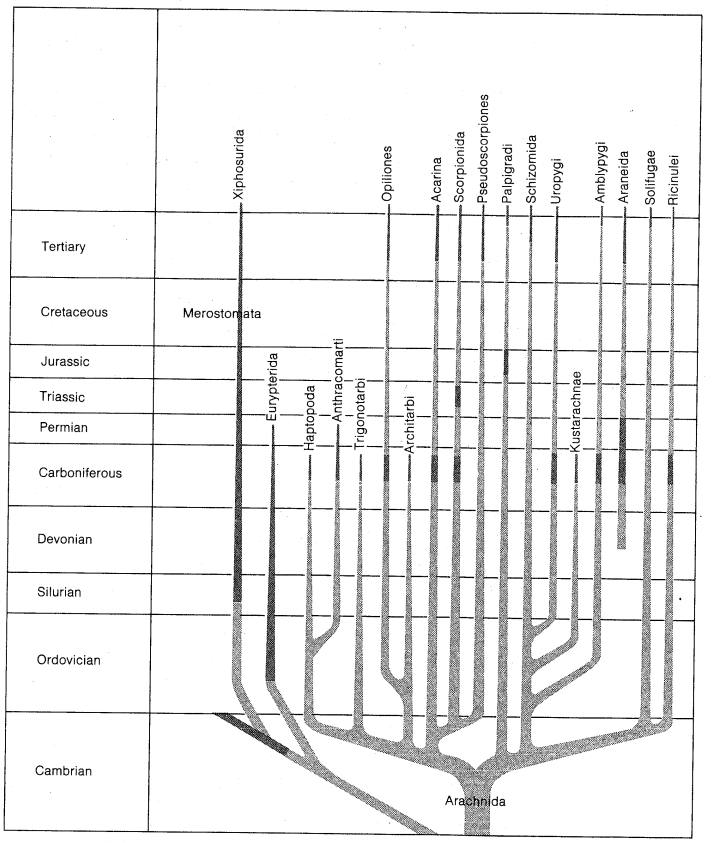


Figure 3: Family tree of arachnids. Solid regions represent discovery of individual groups.

Since the Silurian scorpion closely resembled modern forms, the hypothetical arachnid ancestor must antedate any known fossil form. The oldest arthropods (the trilobites), abundant from the Cambrian Period (about 570,000,000 years ago), are too specialized to be considered ancestral forms of other arthropods. Another group abundant during Paleozoic times was that of the extinct eurypterids, large animals that resembled present-day scorpions; they are not considered ancestors of arachnids.

Theory of the "protochelicerate"

Theodore Savory, an English arachnologist, postulated in 1964 that an animal called a protochelicerate arose among the ancestral arthropods but differed morphologically from other arthropods; *i.e.*, it had a mouth on the underside of its body, an opisthosoma with 12 or 13 somites, and a reduced middle portion of the cerebral ganglion. From marine organisms such as these, evolution could have resulted in four different groups: the sea spiders (Pycnogonida) have an extremely reduced opisthosoma; the eurypterids, now extinct, had body changes that included modification of the last pair of legs into paddles for swimming; the horseshoe crabs (Merostomata), related to the eurypterids, crawled on mud and sand; and a group of eurypterids that may have migrated from shallow water to land may have evolved into present-day arachnids. This terrestrial branch of the chelicerates had to evolve a method for respiration, and it is possible that ancestral gill books represented a preadaptation for land life, for they were easily converted into lung books.

After adaptation to land was achieved, primitive arachnids evolved a more efficient carapace to strengthen the eight legs; in addition, segmentation disappeared in some groups, posterior appendages vanished, and the position of the genital orifice became stabilized in most forms. It

is generally agreed that scorpions have the largest number of primitive arachnid features and that spiders and mites are the most highly evolved. Differences between these two extremes include the disappearance of visible segmentation from the bodies of spiders and mites, the narrowing of the first opisthosomatic somite to form the pedicel in spiders, the secretion of silk, and the varying life patterns.

Classification. *Distinguishing taxonomic features.* In classifying arachnids, taxonomists rely mostly upon external structures, including such features as general body form, the degree of visible external segmentation, structural modifications of prosoma and opisthosoma, characteristics of appendages, and special structures involved in sperm transfer. Internal anatomical features, developmental traits, and serological characteristics are used to a limited extent; conceivably, however, as information becomes available, traits such as these could play a more important role in arachnid classification.

Annotated classification. The following classification of arachnids is based chiefly on work by L. Störmer (1944) and by A. Petrunkevitch (1949). In the classification below, the groups marked with a dagger (†) are wholly extinct and known only from fossils.

CLASS ARACHNIDA (arachnids)

Chelicerate arthropods with adult body composed of 18 somites, divided into anterior prosoma (6 somites) and posterior opisthosoma (12 somites); prosoma with 6 pairs of appendages, 1st pair preoral, last 2 pairs secondarily lost in some mites; genital opening in both sexes on 2nd abdominal somite (except some mites); heart usually in abdomen; respiration by means of book lungs, tracheal tubes, a combination of the two, or cutaneous (some cryptozoic forms); excretion by coxal glands and/or malpighian tubules; digestive system with a pumping foregut, midgut with blind sacs, and a short hindgut opening with an anus on ventral side of opisthosoma; nervous system with a tendency to fuse; not more than 12 simple eyes (ocelli); sexes dimorphic; courtship frequent before sperm transfer; usually terrestrial, carnivorous, nocturnal, and cryptozoic; social organization rare; instinctive behaviour highly developed; distribution variable; about 65,000 species described.

Subclass Latigastra

Prosoma and opisthosoma broadly joined.

Superorder Pectinifera

Pair of comblike abdominal appendages (pectines) on 3rd somite.

Order Scorpionida (Scorpiones; scorpions). Prosoma compact, opisthosoma subdivided, taillike portion ending in stinger; chelicerae chelate (pincerlike), 3-jointed; size from 14–180 mm; ovoviviparous; earliest fossil records from Silurian; widespread in tropical and subtropical regions; about 700 species described (see SCORPIONIDA).

Superorder Epectinata

Pectines absent.

Order Pseudoscorpiones (false or book scorpions). Length, 1–7.5 mm; carapace entire, abdominal somites often subdivided dorsally; chelicerae chelate, movable finger with spinneret; oviparous; worldwide distribution; about 1,700 species described.

Order Opiliones (harvestmen, daddy long-legs). Size 1–22 mm; 3 to 8 opisthosomatic dorsal regions of segments (tergites) fused with carapace, remainder free; single pair of eyes, usually on median tubercle; oviparous; worldwide distribution; about 3,400 species described.

Order Acarina (Acari; mites and ticks). Size from 0.1 to 20 mm in some fully engorged ticks; prosoma and opisthosoma fused, somites not apparent; mouthparts adapted to piercing, sucking, biting, grating, or sawing; highly variable in appearance; many parasitic and economically important forms; worldwide distribution; about 20,000 species described (see ACARINA).

†*Order Architarbi.* Fossil forms with segmented opisthosoma; carapace complete, with 2 to 6 eyes; chelicerae chelate; Upper Carboniferous of Europe, Pennsylvanian of North America.

Subclass Stethostomata

Basal joint of chelicerae wedged between coxae of pedipalps and mouth, situated ventrally behind mouth.

†*Order Haptopoda.* Fossil forms with carapace complete; chelicerae 3-jointed, chelate, from Upper Carboniferous, England.

†*Order Anthracomarti.* Fossil forms with complete carapace; abdomen with 10 segments; from Upper Carboniferous of Europe, Pennsylvanian of North America.

Subclass Soluta

Opisthosomal tergites divided by 2 longitudinal lines; labium, sternum, 5 pairs of coxae in same plane.

†*Order Trigonotarbi.* Fossil forms; with complete carapace, chelicerae 2-jointed, pedipalps leglike, arrangement of coxae as in spiders; from Devonian of Scotland, Upper Carboniferous of Europe, Pennsylvanian of North America.

Subclass Caulogastra

1st opisthosomal somite attenuated, narrowly joined to prosoma.

Superorder Latisterna

Coxae of pedipalps separated by anterior sternite.

Order Palpigradi (palpigrades). Size, 0.8 to 2.6 mm; carapace subdivided into 3 parts; eyes absent; coxae movable; chelicerae chelate; 3-jointed, pedipalps leglike; telson a many-segmented flagellum; no book lungs or tracheae; sporadic distribution; about 36 species described.

Superorder Camarostomata

Coxae of pedipalps fused medially.

Order Uropygi (whip scorpions, vinegarroons). Size, to 130 mm; carapace complete; coxae movable, pedipalps predatory; 1st tarsus tactile; long, whiplike telson; order coextensive with 1 family, Thelyphonidae; fossils nearly identical with living forms; known from Carboniferous of Illinois, Czechoslovakia, and England; widespread in tropical and subtropical regions; about 75 species described.

Order Schizomida (schizomids). Size, 5–7 mm; prosoma with segmented carapace; chelicerae 2-segmented; telson short; sporadic distribution; about 35 species described.

†*Order Kustarachnae.* Fossil forms known only from Pennsylvanian of North America; sternum small, coxae radiate from it.

Superorder Labellata

Coxae of pedipalps contiguous but free.

Order Amblypygi (tailless whip scorpions). Size, 8–45 mm; prosoma broader than long, covered with carapace; 8 eyes; chelicerae not chelate; pedipalps very long, strong; 1st pair of legs tactile; members widespread in tropical and subtropical regions. About 60 species described.

Order Araneida (spiders). Size, 1–75 mm; carapace complete; chelicerae 2-jointed, with ducts of poison glands; pedipalps leglike in males, with special copulatory apparatus; opisthosoma nonsegmented, bearing spinnerets; group large and variable in appearance and habits; worldwide distribution; about 30,000 species described (see ARANEIDA).

Superorder Rostrata

Mouth opening at end of a beak or rostrum; with malleoli (racket) organs on fourth coxae and trochanters.

Order Solifugae (sunspiders, solpugids). Size, 10–50 mm; anterior part of 3-part prosoma swollen into head; chelicerae powerful; opisthosoma 10 or 11 segments; widespread in tropical and subtropical regions; about 800 species described.

Superorder Podogona

Movable cucullus in front of carapace.

Order Ricinulei (ricinuleids). Size, 8–10 mm; carapace complete, with cucullus; abdomen of 9 segments, last 3 forming caudal region pygidium, development with a 6-legged larval form; of sporadic distribution; about 25 species described.

Critical appraisal. Many schemes for the classification of the Arachnida have been proposed. The system used here recognizes the name Arachnida as that of a class of terrestrial chelicerates. This class differs from the aquatic Merostomata (horseshoe crabs) in the absence of gill books and compound eyes, while resembling them in possession of six pairs of prosomatic appendages. This definition of Arachnida results in the elimination of the pycnogonids from the class and is in accord with taxonomic practices of subdividing and separating groups in order to clarify relationships. The classification probably will change and improve in the future; one possible change might be the separation of the acarines from the arachnids, since mites differ greatly in development, structure, and habits from other arachnids.

Arachnologists eventually may utilize serological or behavioral studies to reduce the number of species; the taxonomic importance given spines, tubercles, and other

often highly variable chitinous structures has resulted in a profusion of described species, many of which may be simply variations of a single species.

BIBLIOGRAPHY. R.D. BARNES, *Invertebrate Zoology* (1963), a particularly good discussion of arachnids; J.L. CLOUDSLEY-THOMPSON, *Spiders, Scorpions, Centipedes, and Mites*, new ed. (1968), a readable and well-illustrated account of the ecology and natural history of the sowbugs (Crustacea), myriapods, and arachnids; J.H. COMSTOCK, *The Spider Book*, rev. and ed. by W.J. GERTSCH (1948), a classic and still usable book for information on arachnids; W.J. GERTSCH, *American Spiders* (1949), an informative discussion of arachnids, with emphasis on spiders; P.P. GRASSE (ed.), *Traité de zoologie*, vol. 6 (1949), a very detailed and scholarly discussion, in French, of some arthropod classes, including arachnids; T.H. SAVORY, *The Arachnida* (1964), an excellent, well-illustrated summary of information on arachnids; P. WEYGOLDT, *The Biology of Pseudoscorpions* (1969), an in-depth study of this little known, but fascinating group.

(C.J.Go./M.L.Go.)

Aral Sea

Covering an oval-shaped area that, containing 25,659 square miles (66,458 square kilometres), is almost the size of Sri Lanka, the shallow Aral Sea (Aralskoye More in Russian; Aralskoje More in the transliteration system of the Akademiya Nauk) is the world's fourth largest body of inland water. It nestles in the climatically inhospitable heart of Soviet Central Asia, to the east of the Caspian Sea, and about equidistant from Moscow, Baghdad, and Karachi. Of great interest to scientists because of the remarkable changes in area and volume it has experienced in geologically recent times, the sea is also notable for the changes introduced by modern man, who has diverted some of the waters of the great Syrdarya and Amu Darya (*darya*, "river"), which discharge into it, for purposes of irrigation.

Scientific significance

The Aral Sea lies 174 feet above sea level, straddling the boundary between the Kazakh S.S.R. to the north and the westernmost dependency of the Uzbek S.S.R. Its greatest north to south extent is almost 270 miles, while that from east to west is just over 180 miles. Although the average depth is a shallow 70 feet or so, it descends to a maximum of 226 feet off the western shore. The total volume has been calculated at 255 cubic miles. The Aral Sea derives its name from the Kirgiz Aral-denghiz, "Sea of Islands"—an apt designation, as there are no fewer than 1,130 islands of a size of 2.5 acres or more strewn across its waters.

Physiography. The sea's northern shore—high in some places, low in others—is indented from west to east by the bays of Chernyshev, Tushchibas, Shevchenko, Butakov, and Saryshiganak. The low-lying and irregular eastern shores are interrupted in the north by the huge delta of the Syrdarya and in the south are bordered by a wide tract of shallow water. The equally vast Amu Darya Delta lies on the southern shore, and along the sea's western periphery extends the almost unbroken eastern edge of the 820-foot-high Ustyurt Plateau (Ustyurt Plato). The largest islands are, from west to north, Vozrozhdeniye, Barsakelmes, and Kokaral.

Geology. The shoreline region is made up of marine and continental deposits laid down from the Cretaceous Period (136,000,000–65,000,000 years ago) onward. Cretaceous outcrops in the coastal areas have produced mainly limestone, sandstone, and sand and clay deposits in the northern, southern, and western parts of the sea. Tertiary (65,000,000–2,500,000 years ago) deposits are composed mainly of marl, limestone, compact clay, and clayey sand. Quaternary (2,500,000–10,000 years ago) deposits of various origins are widely distributed along the shores, especially in the east and south: the marine and eolian deposits consist mainly of sandy sediments, the alluvial ones mainly of sand and clay. The seabed is mainly smooth and unbroken and descends from east to west. Bottom deposits, comprising quartz and limestone sands and clays, and clay–limestone silts are the products of the rivers' powerful flow and the erosion of the shores by wind and ice. Stony sediments from rock deposits occur only in the coastal areas.

The Aral Sea depression was formed at the end of the Pliocene Epoch (7,000,000–2,500,000 years ago), when the Earth's crust subsided and the hollow was filled with water—some of which came from the Syrdarya. In the early and middle Pleistocene epochs (2,500,000–10,000 years ago), the depression was subject largely to eolian processes, and in the late Pleistocene Epoch it was inundated for a second time by the Syrdarya and for the first time by the Amu Darya, which had temporarily changed its course from the Caspian to the Aral Sea. At the turn of the 3rd and 2nd centuries BC and at the beginning of the 1st century BC the Aral depression was again inundated by both rivers. Between these inundations dry periods with water occurring only in the western hollow resulted largely from changes in the behaviour of the Syrdarya. But from the Pleistocene Epoch onward, the rivers' combined flow maintained a high water level, which was almost unaffected by climatic changes.

Formation of the Aral Sea depression

Climate. The Aral Sea area is characterized by a desert-continental climate of which the typical features are wide-ranging air temperatures, cold winters, hot summers, and sparse rainfall. Average monthly air temperatures from January to February vary from 10° F (−12° C) in the north to 21° F (−6° C) in the south and in July from 73.9° F (23.3° C) to 79° F (26.1° C), respectively. The rate of precipitation—an average of four inches (100 millimetres) in all and occurring mainly in spring and autumn—equals one-twelfth the rate of evaporation. The average water temperature for July is 73°–77° F (23°–25° C), and, when ice forms in November and December, the mercury falls to 30.8° F (−0.7° C). Northeasterly winds prevail in autumn and winter and westerly and southwesterly winds in spring and summer.

Hydrology. Factors affecting the water balance of the Aral Sea are precipitation, river flow (accounting for about 90 percent of inflow), and evaporation, which takes out each year about the same amount of water that the rivers bring in. Climate may quite considerably influence the long-term variation in water level. Over the centuries, variations have exceeded 20 feet, while annual and seasonal variations of between ten feet and less than a foot have been recorded. In the short term, similar variations have been caused by oscillations of barometric pressure and by a combination of condensation and distillation. From 1960 onward the water level has been systematically reduced by climatic changes and by the diversion of water from the Amu Darya and Syrdarya to agricultural irrigation. The mean salt level of the sea is from ten to 11 parts per thousand rising to 14 parts per thousand on the southeastern shores and, in the shallow bays from 8 to 15 percent.

Variations in water level

Human geography. Aralsk in the northeast and Muynak in the south are important economic centres and ports between which ships steam for seven months of the year. Fishing villages are scattered along the northern and eastern shores in the Aralsk region, to which catches of such freshwater fish as sturgeon, carp, barbel, zander, rapfen (*Aspius aspius*), bream, and roach are conveyed. Newly introduced strains of Atlantic herring, striped mullet, and goby are also caught in coastal waters. The inhabitants of the settlements near Muynak are similarly engaged in fishing but also—though to a lesser extent—in stock raising, muskrat breeding, and the cultivation of melons. In many of the shallow bays and coves—particularly in the east—there are large accumulations of sodium sulfate and sodium chloride.

If controls are not instituted, the continuing diversion of the Amu Darya and the Syrdarya to irrigation, the increase in agricultural land, and the formation of new sodium chloride and sodium sulfate deposits will ever more quickly reduce the volume of the sea. (V.Y.L.)

Araneida

Spiders (order Araneida) are common terrestrial members of the invertebrate arthropod class Arachnida. They differ from another arthropod group, the insects, in having eight legs rather than six and in having the body

divided into two parts rather than three. Spiders attract attention by the webs they spin and by their hunting habits, both of which are instinctive traits. About 34,000 species of spiders have been described thus far; the world's spider fauna outside of northern Europe, however, has not yet been thoroughly studied.

Spiders are predators, feeding almost entirely on other arthropods, especially insects. Many spiders that are active hunters and overpower their prey have a well-developed sense of sight. Others construct silk snares, or webs, to capture prey, sometimes in larger numbers than can be consumed; webs, constructed instinctively, effectively trap flying insects. Many spiders inject venom into their prey to kill it.

From P.A. Meglitsch, *Invertebrate Zoology;* Copyright © 1967 by Oxford University Press, Inc.; reprinted by permission

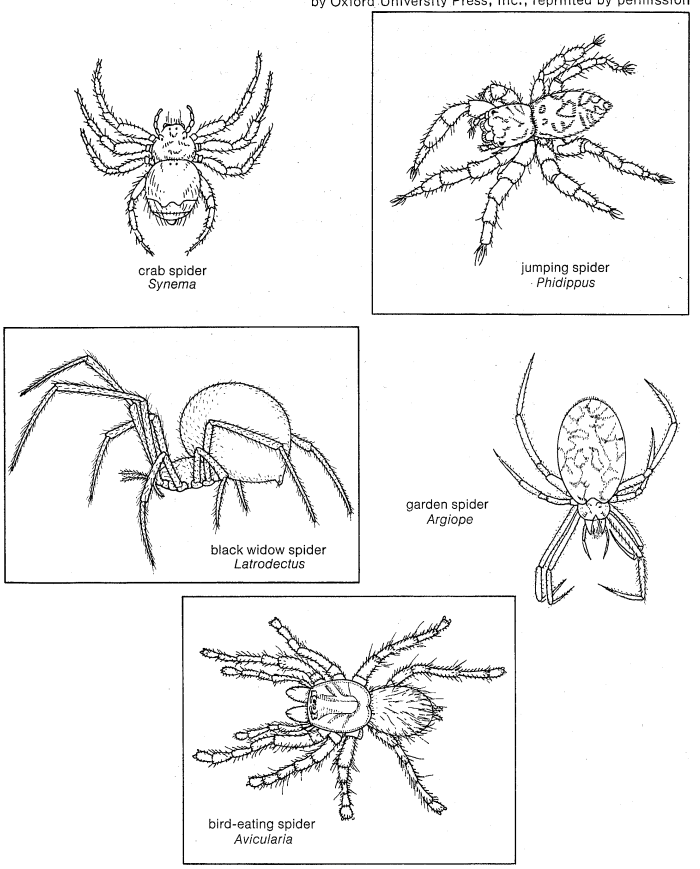


Figure 1: Representative spider body plans.

GENERAL FEATURES

Importance. The value of spider predation is difficult to assess, but spiders probably play an important role in the control of populations of certain insect species. The web-building process is of considerable scientific interest and has practical application in tests of certain drugs that affect the nervous system—*i.e.*, webs deviate from normal when built by spiders treated with drugs, reflecting drug-induced changes in their brains. A few spiders affect man—*i.e.*, the venom of the black widow and other widow spiders (*Latrodectus*) acts as a nerve poison; the bite of the brown recluse (*Loxosceles*) may cause localized death of living tissues; and the hairs of some mygalomorph spiders may cause itching and other allergic reactions.

Size range. Spiders range in body length from less than one millimetre to about 90 millimetres (about 3½ inches). The smallest, the dwarf spiders, are common in the temperate Northern Hemisphere; oonopid and symphytognathid spiders also are small. The largest spiders, the theraphosids or hairy mygalomorphs, found in warm climates, are most abundant in the Americas. The largest species, *Theraphosa leblondi*, is found in Guiana.

Distribution. Spiders have been found on all continents except Antarctica (although spider fragments have been reported) and at elevations as high as 5,000 metres (16,400 feet) in the Himalayan Mountains. More species occur in the tropics than in temperate regions. Though most spiders are terrestrial, one Eurasian species is aquatic and lives in slow-moving freshwater, and a few species live along shores or on the surfaces of either freshwater or oceans.

Small spiders are transported through the air, sometimes for long distances, by wind action on long silk strands they secrete; the process, called ballooning, expedites their distribution. Ballooning spiders, which probably occur in all families, drift through the air at heights that range from three metres or less to more than 800 metres (2,600 feet).

NATURAL HISTORY

Reproduction and life cycle. *Courtship.* In male spiders, the second pair of appendages (pedipalps) is modified to form a complex structure for holding sperm; the enlarged tips of the pedipalps form copulatory organs. When the time for mating approaches, the male constructs a special web, called the sperm web. The silk comes from silk glands and emerges from the spinnerets; epigastric silk glands also supply silk, which emerges from spigots in the epigastric area (between the book lungs). A drop of fluid containing sperm is deposited through a structure located on the underside of the abdomen (gonopore) onto the sperm web. The male draws the sperm into his pedipalps; the process of charging the pedipalps, known as sperm induction, may require minutes or hours for completion. Sperm induction may occur before a male seeks a mate or after the mate is found; if more than one mating occurs, the male charges his pedipalps between copulations.

Charging the pedipalps

The way in which a male finds a female varies among different species. The males of many species, for example, wander more extensively than the females, and males of the relatively sedentary web weavers generally have longer legs than females. The wandering males of some species follow silk threads; some of them may recognize both the threads produced by a female of the same species and the female's condition—*i.e.*, whether she is mature and receptive. Other species (*e.g.*, jumping spiders) use visual senses to recognize mates.

Males in a few species locate a female and then unceremoniously run to her and mate. In most species, however, elaborate courtship patterns have evolved, probably to protect the male from being mistaken for prey. The male web spider courts by rhythmically plucking the threads of a web; after the female approaches, he pats and strokes her before mating. When male wolf spiders or jumping spiders see a female, they wave the pedipalps, conveying a visual message characteristic of the species; an appropriate response from a female encourages the approach of the male. Some male wolf spiders tap dry leaves, perhaps to attract a female; an aggregation of tapping males produces sound that can be heard some distance away. A crab spider male quickly and expertly wraps his intended mate with silk. Although the female is able to escape, she does not do so until mating is completed. After the male of the European nursery web spider locates a suitable mate, he captures a fly, wraps it in silk, and presents it to the female; while the female is occupied with eating the fly, the male mates with her. If no fly is available, the male may wrap a pebble. Some male spiders use specialized structures to immobilize the jaws of the female during mating.

Mating. After a male has successfully approached a female, in most groups he inserts his left pedipalp into the left opening of her genital structure and the right pedipalp into her right opening.

The female genital structure, or epigynum, is a hardened plate on the underside of the abdomen anterior to the gonopore. After the sperm are transferred into the epigynum, they move into receptacles (spermathecae) that connect to the oviducts; in this way, eggs are fertilized as they pass through the oviducts and out through the gonopore. In some primitive spiders (*e.g.*, haplogynes, mygalomorphs) and some Tetragnathidae lacking an epigynum, the male inserts both pedipalps simultaneously

into the female's genital slit, the left pedipalp on her right side and the right pedipalp on her left side.

The force that causes the injection of sperm from the pedipalp of the male into the receptacle of the female has not yet been established with certainty but may involve blood pressure. Increases in blood pressure expand a soft vascular tissue (hematodocha) between the hard plates of pedipalps, causing a bulbous structure containing a duct to twist and to hook into the epigynum of the female.

Epigynal plug

Mating may require seconds in some species, hours in others. Some males recharge their pedipalps and mate again with the same female. After mating, the males of some species smear secretion over the epigynum; called an epigynal plug, the secretion prevents the female from mating a second time. Male spiders usually die soon after, or even during, the mating process. The female of one European orb weaver species bites into the abdomen of the male and holds on during mating. Although some females eat the male after mating, the practice is not common. The male of the black widow, *Latrodectus mactans*, for example, usually dies days after mating; occasionally he is so weak after mating that he is captured and eaten by the female.

Females of some species mate only once; others accept several males, one after another. The long-lived females of mygalomorph spiders must mate repeatedly because they shed their skins once or twice a year, including the lining of the sperm receptacles.

Eggs and egg sacs. Female spiders produce either one egg sac containing several to a thousand eggs, or several egg sacs, each with successively fewer eggs. Females of some species usually die after producing the last egg sac or after the young have been taken care of; these females live one or, at most, two years. On the other hand, the females of the mygalomorph spiders may live 20 years, and those of the primitive haplogyne spiders ten.

The protective egg sac surrounding the eggs of most spiders is made of silk. Although a few spiders tie their eggs together with several strands of silk, most construct elaborate sacs of numerous layers of thick silk; the colour of the silk may vary. Eggs, which often have the appearance of a drop of fluid, first are deposited on silk and then wrapped in silk and covered, so that the finished egg sac is spherical or disk-shaped. The females of many species place the egg sac on a stalk, attach it to a stone, or cover it with smooth silk before abandoning it. Other females guard the egg sac or carry it either in their jaws or attached to the spinnerets. If a female loses an egg sac, she makes searching movements and may pick up a pebble or a piece of paper and attach it to the spinnerets. The European cobweb spider, *Achaearanea saxatile,* lowers its egg sac from a protective thimble-shaped structure or pulls it inside, depending on the temperature. Female wolf spiders carry their egg sacs attached to the spinnerets and instinctively bite the egg sac to permit the young to emerge after a certain time.

Fate of the egg sac

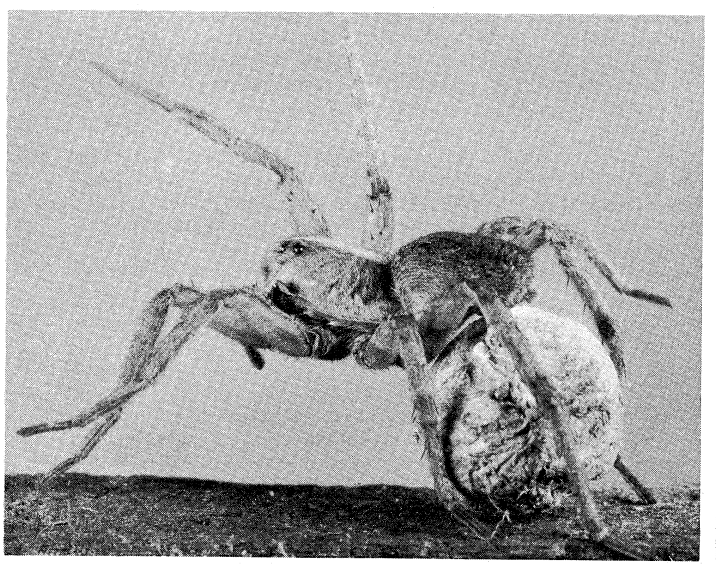
Maturation. After hatching, young wolf spiderlings, usually numbering from 20 to 100, climb onto the back of their mother and remain there about ten days before dispersing. If they fall off, they climb back up again, seeking contact with bristlelike structures (setae) or something similar in texture. Some female spiders feed their young; the female of one European species dies at the time the young are ready to feed, and they eat her carcass. When food has been sufficiently liquefied by females of other species (in spiders, digestion occurs outside the mouth), the young also feed on their mother's prey. The mother of one web spider species plucks threads of the web to call her young, both to food sources and to warn them of danger. The young of many species are independent as they emerge from the egg sac.

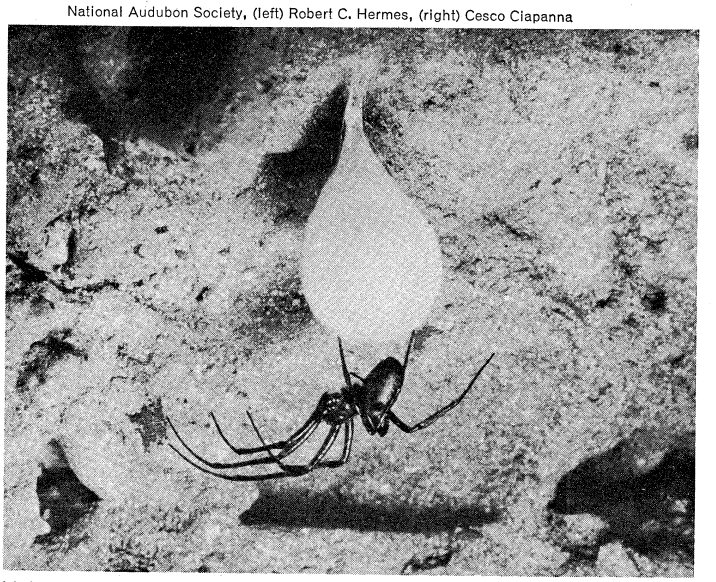
Molting

Young spiderlings, except for size and undeveloped reproductive organs, resemble adults. They shed their skins (molt) as they increase in size. The number of molts varies among species, within a species, and also among related young of the same sex. Males have fewer molts than females and mature earlier; females may molt six to 12 times, males two to eight. Males of some species may emerge mature from the egg sac, one or two molts having occurred before emergence. Some spiders mature a few weeks after hatching, but many overwinter in an immature stage. Mygalomorph spiders require three to four years to become sexually mature in warm climates.

Before molting many spiderlings hang by the claws in some inconspicuous place; mygalomorph spiders turn on their side or back. The protective cover (carapace) surrounding the cephalothorax breaks below the eyes, and the spider laboriously extracts its legs from their casings as the old abdominal cuticle (skin) shrinks. The cast off cuticle, or exuviae, remains hanging, and the newly emerged spider dangles on a thread, helplessly soft. Molting is hazardous for spiderlings: they may dry out before emerging from the old cuticle or fall victim to a predator. Even a small injury during a molting period usually is fatal. Emergence from the exuviae is accompanied by wide fluctuations of blood pressure; these raise and lower the setae (bristlelike hairs) and gradually force the legs free. The first break around the carapace edge also is caused by increased blood pressure. Growth and molting are believed to be under the control of hormones; some spiderlings fail to molt, others undergo delayed molts, perhaps because of faulty hormone balance, and may die.

National Audubon Society, (left) Robert C. Hermes, (right) Cesco Ciapanna





Eggs of spiders are enclosed in silk sacs: (left) female wolf spider with her portable egg sac attached to spinnerets; (right) cave orb weaver (*Meta menardii*) with egg sac suspended from ceiling.

Many spiderlings eventually disperse by ballooning, usually in the fall. After a spiderling climbs a grass blade, fence post, or shrub, it lifts its abdomen and lets out strands of silk until the wind carries it away. Even though ballooning is an important means of distribution among many spider families (especially Linyphiidae), little is known about it thus far; large mygalomorph spiders probably do not balloon.

Feeding behaviour. As predators, spiders feed mainly on insects, other spiders, sowbugs, centipedes, and millipedes; relatively few, however, are specialized predators. Many hunt at random, stalking prey encountered accidentally; others use silk to snare prey.

Stalking of prey. Most hunting spiders locate prey either by searching randomly or by responding to vibrations; wolf spiders and jumping spiders have keen eyesight. Jumping spiders stalk their prey within five or ten centimetres, then pounce when it moves. Many crab spiders wait for prey on flowers similar in colour to their own; the spider uses his legs to grasp an unsuspecting insect and then gives it a lethal bite.

Spitting spiders

Unique among the hunters are the spitting spiders (family Scytodidae). When these spiders encounter an insect, they touch it, back off, and shoot a zigzag stream of sticky material over it. The sticky material, produced by modified venom glands in the cephalothorax, emerges from pores near the tips of the fangs, which are located on the chelicerae; as the victim struggles, the spider approaches cautiously and bites the entangled insect.

Trapping of prey. Spiders that use silk to capture prey utilize various techniques. Ground-dwelling trap door spiders construct silk-lined tubes, sometimes with silk trapdoors, which they open at night, capturing passing insects. Other tube-dwelling spiders place silk threads around the mouth of the tube and trip insects; vibrations of the threads inform the spider of a victim's presence. Funnel-weaving spiders live in silk tubes with a narrow end that extends into vegetation or into a rock wall and an expanded sheetlike end that vibrates when an insect walks across it. Many web spiders construct several sheets in vegetation, sometimes one above the other, and often add anchor threads, tripping unsuspecting insects. The irregular three-dimensional web of cobweb spiders (Theridiidae) has anchoring threads of sticky silk. An insect caught in the web or touching an anchor line becomes entangled, increasingly so if it struggles; if a thread breaks, the force of elasticity in it pulls the insect toward the centre of the web.

Web construction

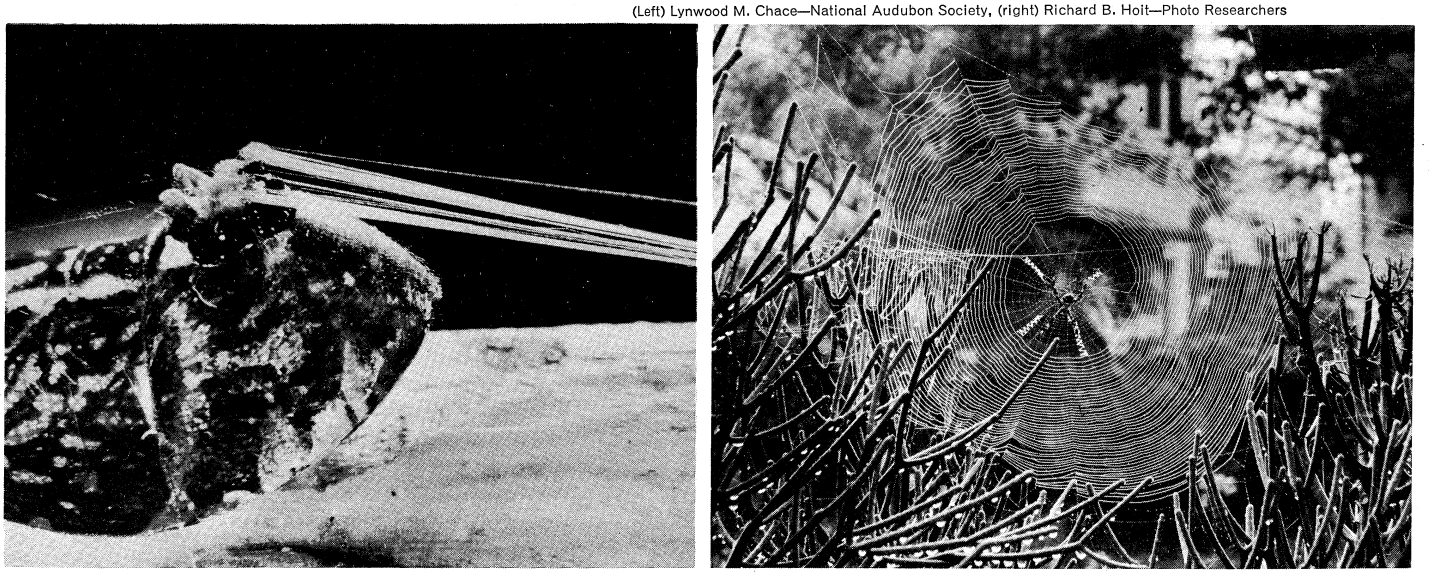
The most elaborate webs (orb webs), which are conspicuous on mornings with dew, are constructed by several spider families, suggesting that they are efficient traps, covering the largest area with the least possible silk. The web acts as an air filter, trapping weak-flying insects who cannot see the fine silk. Although web construction is a complex process, it usually is completed in about an hour, generally just before daybreak. A web is damaged during capture of prey, and most spiders construct new webparts frequently. The ways by which spiders keep from becoming entangled in their own webs are not yet completely understood nor is their mechanism for cutting the extremely elastic silk threads used in web construction.

During orb web construction, the spider releases a silk thread which is carried by wind. If the free end does not become attached to some object, the spider may pull it back and feed on it; if it becomes firmly attached—for example, to a twig—the spider secures the thread and crosses the newly formed bridge, reinforcing it with additional threads. The spider then descends from the centre of the bridge, securing a thread on the ground or on a twig. The centre, or hub, of the web is established when the spider returns to the bridge with a thread, carrying it part way across the bridge before securing it; this thread is the first radius, or spoke. After all the spokes are in place, the spider returns to the hub and constructs a few temporary spirals of dry silk toward the outside of the web. The spider then reverses direction, deposits ensnaring silk, and removes the spiral. The ensnaring threads form a dense spiral. Some species attach a signal thread from the hub to a retreat in a leaf so that they are informed (by vibrations) of trapped insects; others remain head down in the centre of the web, locating prey by sensing tensions or vibrations in individual spokes. Webs of the family Uloboridae have a woolly (cribellate) ensnaring silk; those of the families Araneidae and Tetragnathidae have spirals constructed of a sticky material that dries after several days. As a result, spiders frequently remove the silk and rebuild the sticky part of the web.

Spiders usually cover a captured insect in silk and turn it efficiently, as on a spit, before biting it and carrying it either to a retreat or to the hub of the web for feeding or storage. Although the scales on butterfly and moth wings facilitate their escape from the web, spiders have evolved a counterstrategy; *i.e.*, they bite the moths before wrapping them, rather than afterward.

The webs of uloborid spiders of the genus *Hyptiotes* consist of only three sections of the orb web. The spider, attached by a thread to vegetation, holds one thread from the tip of the hub until an insect brushes the web; then the spider alternately relaxes and tightens the thread, and the struggling victim becomes completely entangled. Ogre-faced spiders of the family Dinopidae build small flat webs during the evening hours and then cut the attachments and spread the web among their four long an-

(Left) Lynwood M. Chace—National Audubon Society, (right) Richard B. Holt—Photo Researchers



Web building involves (left) drawing out of silk from spinnerets as is shown for the golden argiope and (right) the weaving of strands into geometric traps like the orb of the silver argiope spider.

terior legs. During the night, an insect flying against the web is readily entangled. The spider abandons or eats the web in the morning and passes the day resting on a branch before constructing a new web. The bolas spider, *Mastophora*, releases a single thread with a sticky droplet at the end (bolas) and holds it with one leg; some species swing the bolas, others throw it when a moth approaches.

Communal webs

Several species of spiders are social; *e.g.*, *Agelena consociata* from Gabon and *Stegodyphus* from Pakistan construct large communal webs containing hundreds of spiders. Prey is killed by a group of spiders hunting on the web in a pack, and feeding is communal.

From P.A. Meglitsch, *Invertebrate Zoology*, copyright © 1967 by Oxford University Press, Inc.; reprinted by permission

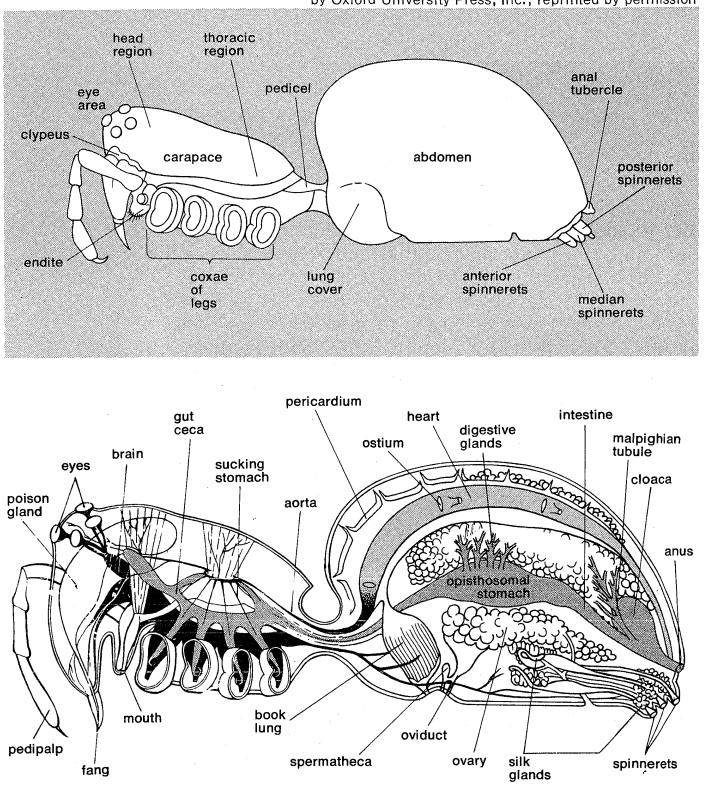


Figure 2: (Top) External organization; (bottom) internal organization of a spider.

FORM AND FUNCTION

External features. The bodies of spiders, like those of other arachnids, are divided into two parts, the cephalothorax (prosoma) and the abdomen (opisthosoma). The legs are attached to the cephalothorax, which contains the stomach and brain. The top of the cephalothorax is covered by a protective structure, the carapace; the underside is covered by a structure called the sternum, which has an anterior projection, the labium. The abdomen contains the gut, the heart, the reproductive organs, and the silk glands. Spiders (except the primitive Mesothelae) differ from other arachnids in lacking external segmentation of the abdomen and in having the abdomen attached to the cephalothorax by a narrow stalk, the pedicel. The gut, the nerve cord, the blood vessels, and sometimes the respiratory tubules (tracheae) pass through the narrow pedicel, which allows the body movements necessary for efficient use of the spinnerets during web construction. Among arachnids other than spiders, the tailless whip scorpions (Amblypygi) have a pedicel but lack spinnerets. Spiders, like other arthropods, have an outer skeleton; inside the cephalothorax is the endosternite, to which some jaw and leg muscles are attached.

Spiders have six pairs of appendages. The first pair, the jaws, or chelicerae, have a fang containing the opening of a poison gland near the tip. The chelicerae move forward and down in the Orthognatha, sideways and together in the Labidognatha. The venom ducts pass through the chelicerae, which sometimes also contain the venom glands. The second pair of appendages, the pedipalps, are modified in the males of all adult spiders to carry sperm (see above, *Reproduction and life cycle*). In females and immature males, however, the leglike pedipalps are used to handle food and must also function as sense organs. The pedipalpal segment attached to the abdomen (coxa) usually is modified to form a structure (endite) that is used in feeding.

The pedipalps are followed by four pairs of walking legs. Each leg consists of eight segments—the coxa, attached to the cephalothorax; a small trochanter; a long, strong femur; a short patella; a long tibia; a metatarsus; a tarsus, which may be subdivided in some species; and a small pretarsus, which bears two claws in running spiders and an additional median claw in web-building ones. The young of two-clawed running spiders often have three claws. The legs, covered by bristle-like hairs (setae), contain several types of sense organs and may have accessory claws. A few species use the first pair of legs as feelers. Spiders can amputate their own legs (autotomy); new but shorter legs appear at the next molt.

Sense organs. The nervous system, unlike that of other arachnids, is completely concentrated in the cephalothorax. The masses of nervous tissue (ganglia) are fused with a ganglion found under the esophagus and below and behind the brain (subesophageal ganglion). The shape of the brain, or epipharyngeal ganglion, somewhat reflects the habits of the spider; *i.e.*, in the web weavers, which are sensitive to touch, the posterior part of the brain is larger than is that in spiders with good vision.

Eyes

The simple eyes (ocelli) of spiders, which number eight or less, consist of two groups, the main or direct eyes (called the anterior medians) and the secondary eyes, which include anterior laterals, posterior laterals, and posterior medians. The rhabdoms, which receive light rays, face the lenses in the main eyes; in the other eyes, the rhabdoms turn inward. Both the structure of the secondary eyes and eye arrangement are characteristic for each family.

Other sense organs are long fine hairs (trichobothria) on the legs, which are sensitive to air currents and vibrations. Slit sense organs are minute slits, or several parallel slits, located either near the leg articulations (or hinges) or scattered over the body. The slit is closed toward the outside by a thin membrane and on the other side by a very delicate membrane that may be penetrated by a nerve. Slit sense organs are sensitive to stresses on the cuticle; others act as vibration receptors or hearing organs. Internal receptors (proprioceptors) provide information about body movement and position. Olfactory organs of spiders may be certain hollow hairs at the tips of pedipalps and legs.

Internal features. *Digestion and excretion.* The mouth leads into a narrow passage, the pharynx, which leads to a stomach. The midgut has a variable number (usually four pairs) of blind extensions, or ceca, that extend into the coxae. Additional ceca and a branched digestive gland are located at the front of the abdomen. At the end of the gut a cecum (stercoral pocket, or cloaca) connects with the hindgut before opening through the anus.

Food is digested outside the mouth (preorally). Some spiders chew their prey as they cover it with enzymes secreted by the digestive tract; others bite the prey and pump digestive enzymes into it before sucking up the nutrient liquid.

The excretory system includes large cells in the prosoma that concentrate nitrogen-containing wastes (nephrocytes); a pigment-storing layer (hypodermis); the ends of the abdominal gut ceca, which are filled with a white excretory pigment (guanine); coxal glands; and tubular glands (malpighian tubules) in the abdomen. The excretory compounds include various nitrogen-containing compounds; *e.g.*, guanine, adenine, hypoxanthine, and uric acid; silk may have evolved from an excretory product.

Respiration. The respiratory system, located in the abdomen, consists of book lungs and tracheae. Book lungs are paired respiratory organs composed of ten to 80 hollow leaves that hang into a blood sinus, are separated by

small hardened columns, and open into chambers (atria), which open to the outside through one or several slits (spiracles). Book lungs frequently transform into tracheae, which are tubes that conduct air to various tissues. The two respiratory organs at the very front end of the abdomen usually are book lungs; the posterior two are tracheae; in some groups, however, both pairs are book lungs (Orthognatha) or tracheae (Symphytognathidae). Because book-lung covers, which are hardened plates, may be present in species in which the book lungs have transformed into tracheae, it is impossible to determine from fossils whether a spider had lungs, tracheae, or both.

Circulation. The circulatory system is best developed in spiders with book lungs and is least developed in spiders with bundles of tracheae going to various parts of the body. The abdomen of all spiders contains a tube-shaped heart, which usually has a variable number of openings (ostia) along its sides and two arteries to carry blood forward and backward when the heart contracts. The ostia close during contraction. The forward-flowing artery is branched in spiders with book lungs; the blood eventually empties into spaces, flows into the lung sinus, and travels into a cavity (pericardial cavity) from which it enters the heart through ostia. The blood, or hemolymph, contains various kinds of blood cells and a respiratory pigment, hemocyanin. Changes in blood pressure function to extend the legs and break the skin at molting time.

Heart

The sex organs (gonads) of male and female spiders are in the abdomen. The eggs are fertilized as they pass through the oviduct to the outside. The fertilized egg (zygote) develops in the manner typical of arthropod eggs rich in yolk.

Specialized features. *Venom.* Venom glands, present in most spiders but absent in the family Uloboridae, are located either in the chelicerae or under the carapace; venom ducts cross the chelicerae and open near the tips of their fangs. A muscle layer covers the venom glands in *Latrodectus.* Venom glands probably originated as accessory digestive glands whose secretions aided in the external digestion of prey. Although the secretions of many spiders may consist entirely of digestive enzymes, those of some species effectively subdue prey, and venoms of a few species are effective against predators, including vertebrates. The spitting spiders (*Scytodes*) secrete a sticky substance that glues potential prey to a surface. The structure of the carapace of the spitting spider is modified as a result of the large glands it contains.

Characteristics of the venom of widow spiders (*Latrodectus*), especially the black widow (*Latrodectus mactans*), have been determined. The various protein components comprising the venom affect specific organisms; *i.e.*, different components affect mammals and insects. Widows (*Latrodectus*) exhibit a warning colouration in the red hourglass-shaped mark on the underside of the abdomen and often have a red stripe; because the spider hangs upside down in its web, the hourglass mark is turned conspicuously upward. The venom, a nerve toxin, causes severe pain in man, especially in the abdominal region, but is usually not fatal. *Latrodectus mactans* is probably the same species as the katipo of New Zealand, the redback of Australia, the button spider of South Africa, the malmignatte of the Mediterranean region, and the karakurt of southern Russia; there are no widows in central and northern Eurasia. The brown widow (*L. geometricus*), found throughout the world in the tropics, is venomous but reluctant to bite man.

Effect of spider bites

The bite of the brown recluse (*Loxosceles reclusa*) results in a localized region of dead tissue (necrotic lesion) that heals slowly; the larger *Loxosceles laeta* of South America causes a more severe lesion in man. The bites of several other species may occasionally cause necrotic lesions; *e.g.*, *Lycosa raptoria, Mastophora, Phidippus formosus, Chiracanthium mildei,* and other species of *Chiracanthium.* Knowledge of the effects of spider bites on man is limited usually because either the bites, including those of the black widow, are not noticed or the spider is never identified.

Silk. Although silk is produced by some insects, centipedes, and millipedes, and a similar substance is produced by some crustaceans (ostracods and amphipods), mites, and pseudoscorpions, only the spiders are true silk specialists. Spider silk is a protein called fibroin, which has chemical characteristics similar to those of insect silk. Spiders utilize about six different types of glands in the abdomen to produce silk. Ducts from the glands enter structures called spinnerets, which open to the outside through spigots. Abdominal pressure forces the silk to flow outward; the rate of flow is controlled by muscular valves in the ducts. The primitive Mesothelae have only two types of silk glands, but orb weavers have at least six, each of which produces a different kind of silk; *e.g.*, aciniform glands produce silk for wrapping prey, ampullate glands produce the sticky globules in the web spiral, and cylindrical glands produce parts of the egg sac. Epigastric silk glands of male spiders produce silk that emerges through spigots in the abdomen between the book-lung covers; silk from these glands provides a surface for the sperm drop in sperm induction.

Threads of *Nephila* silk have a very high tensile strength and great elasticity. Silk probably changes to a solid in the spigot, perhaps as a result of tension forces. Strands usually are flat or cylindrical as they emerge and are of surprisingly uniform diameter. The glob of silk that binds or anchors strands presumably emerges from the spigot as a liquid.

The movable spinnerets, which consist of hardened tubes, are modified appendages; two pairs are from the tenth segment, two pairs from the eleventh. *Liphistius* (Mesothelae) is the only spider with a full complement of four pairs of spinnerets in the adult; the Palpimanidae have only one pair. Most spiders have three pairs; usually the anterior median pair is lost or reduced either to a nonfunctional cone, called the colulus, or to a flat plate, the cribellum, through which open thousands of minute spigots. Spiders with a cribellum also have a comb, the calamistrum, on the metatarsus of the fourth leg. The calamistrum combs the silk that flows from the cribellum, producing a characteristic, woolly (cribellate) silk.

PALEONTOLOGY AND EVOLUTION

Spiders are recorded in red sandstone from the Devonian Period (about 395,000,000 to 345,000,000 years ago) and have been described from the Carboniferous Period of Europe and America (about 345,000,000 to 280,000,000 years ago). Oligocene amber (about 26,000,000 to 38,000,000 years old) is rich in spiders.

The characters of living spiders useful in taxonomic identification usually are not available in fossil spiders; as a result, evolutionary histories based on fossil spiders may not be meaningful.

The oldest fossils, from the Devonian Period, were placed by Alexander Petrunkevitch in the family Arthromygalidae of the suborder Liphistina (Mesothelae in the annotated classification below); Theraphosina (suborder Orthognatha in the classification below) are not known before the Oligocene Epoch (about 38,000,000 to 26,000,000 years ago). Many families of spiders of the Recent Epoch (the last 10,000 years) are known from Oligocene amber; numerous spider specimens described from Baltic Oligocene amber are believed to have been destroyed during World War II.

Spiders with primitive features, relatively simple genital structures, and four book lungs usually have a long life span, are often large in size, and do not make elaborate use of silk. Small spiders appear to have been more successful than large ones, both in numbers of species and individuals; evolution, therefore, has tended to result in smaller size, an annual life cycle, and complex behaviour, including elaborate use of silk.

CLASSIFICATION

Distinguishing taxonomic features. The Araneida may be separated into three suborders—Mesothelae, Orthognatha (mygalomorph spiders), and Labidognatha (araneomorph spiders)—on the basis of the type of movement

of the two jaws; *i.e.*, movement forward and down is orthognath (paraxial), movement sideways and together is labidognath (diaxial). Other features that distinguish suborders include the structure of the male pedipalps, presence or absence of an epigynum in the female, presence and number of book lungs, number of small openings (ostia) in the heart, and extent of fusion of nerve ganglia in the prosoma. Families are distinguished on the basis of such characteristics as number and spacing of simple eyes (ocelli), number of tarsal claws, number of spinnerets, habits, structure of chelicerae, and fusion of an anterior projection (labium) to the underside of the cephalothorax (sternum).

Annotated classification. The classification presented here is that of Ulrich Gerhardt and Alfred Kaestner (1937), used by Alfred Kaestner (1968). Of the numerous schemes published in the 1930s in response to that of Petrunkevitch, the one below is most widely accepted. The nomenclature of spiders starts with C. Clerck in 1757, rather than with Linnaeus (1758), whose tenth edition is the basis for the nomenclature of other animals. Spiders are not well known, and new species are continuously described; few comparative revisionary studies, however, have been made. As a result, the classification presented below is incomplete and out-of-date. Families and species of the Recent Epoch only are considered. The number of species in each group is a rough estimate, extrapolated from figures given for 1935.

ORDER ARANEIDA (ARANEAE) (spiders)

Members of the class Arachnida; unsegmented abdomen (opisthosoma) attached to cephalothorax (prosoma) by pedicel; prosoma with poison glands whose ducts open through large chelicerae with a movable fang, no pincers; leglike pedipalps usually with modifications (endites) on coxae to aid in feeding, modified as copulatory organs in males; abdomen usually with 2 pairs of respiratory organs, anterior book lungs and posterior tracheae; spinnerets at posterior end of abdomen; 4 pairs of walking legs; 1 or 2 pairs of coxal glands; about 30,000 species described; worldwide distribution.

Suborder Mesothelae

Abdomen clearly segmented dorsally from 7th segment (pedicel) to 18th; 8 spinnerets at middle of abdomen; chelicerae orthognath (paraxial), attached in front of carapace; pedipalpal coxae without endites; male pedipalps relatively complicated; epigynum absent in females; 18 ganglia in prosoma more distinctly separated than those in other suborders; heart with 5 pairs of ostia in segments 8 to 12; 2 pairs of book lungs; 9 species from Japan to southeastern Asia; inhabit trapdoor tubes in ground.

Family Liphistiidae. Habitat, tube in ground as deep as 60 centimetres, widened at end to form chamber; silk threads from opening trip potential prey; trapdoor held shut by spider; body length up to 3.5 centimetres.

Suborder Orthognath (mygalomorph spiders)

Chelicerae orthognath (paraxial), attached in front of carapace; pedipalpal coxae without endites (except in Atypidae); 2 pairs of book lungs; heart with 4, rarely 3 ostia; bulb of male pedipalp simple; female without epigynum; 13th through 18th ganglia lost, others fused; most species large and long-lived in warm climates; about 2,000 species.

Family Ctenizidae (ctenizid trap-door spiders). Chelicerae with structure (rake or rastellum) used to dig; 3 tarsal claws; labium and sternum separate; eyes closely grouped; most species at least 3 centimetres or more in length; inhabit silk tubes in ground; entrances covered by hinged silk lids; about 700 species.

Family Antrodiaetidae (antrodiaetid trap-door spiders). Chelicerae with rastellum; 3 tarsal claws; eyes closely grouped; abdomen with tergites (dorsal hardened plates); anal projection, or tubercle, behind 4 to 6 spinnerets; door of burrow a low collar or turret; distributed in Japan and North America, from Alaska to southern U.S.; only a few species.

Family Actinopodidae (actinopodid trap-door spiders). Chelicerae with rastellum; 3 tarsal claws; labium fused to sternum; eyes widely spaced across carapace; found in Argentina, Africa, Australia; only a few species.

Family Migidae (migid trap-door spiders). No cheliceral rastellum; 3 tarsal claws; labium fused to sternum; 4 spinnerets; found in temperate Southern Hemisphere; about 50 species.

Family Dipluridae (diplurid funnel-web mygalomorphs). No cheliceral rastellum; 3 tarsal claws; 4 to 6 spinnerets, posterior (lateral) pair very long; body hairy; web similar to funnel webs of Agelenidae; venom of *Atrax* (southeastern Australia) feared; some North American representatives only 3 millimetres long; about 300 species.

Family Mecicobothriidae (mecicobothriid funnel-web mygalomorphs). Chelicerae lack rastellum; 3 tarsal claws; abdomen with hard plates (sclerites); 6 spinnerets, posterior pair long; anal tubercle well separated from spinnerets; inhabits leaves, liter, and logs on North American Pacific coast; only a few species.

Family Atypidae (purseweb spiders). No rastellum; 3 tarsal claws; pedipalpal coxae with endites; 6 spinnerets; less than 3 centimetres long; live in closed silk tubes partly below ground; bite prey through tube and pull it in; found in Europe, North America, Japan, Burma, and Java; about 20 species.

Family Theraphosidae (hairy mygalomorphs, tarantulas of North America; monkey spiders of South Africa; bird-eating spiders of Australia). No cheliceral rake; 2 tarsal claws; labium not fused to sternum; large in size, hairy, nonsocial, nocturnal; some burrow, others inhabit trees; mostly tropical; *Dugesiella hentzi* in south central U.S.; *Theraphosa* of Amazon River Basin up to 9 centimetres in body length with 28 centimetres leg spread; about 800 species.

Family Pycnothelidae (pycnothelid mygalomorphs). Two tarsal claws, no claw hair tufts; tarsi with scopulae (a brush); labium fused to sternum with seam; eyes grouped on a projection; 4 spinnerets; found from Brazil to Patagonia; only a few species.

Family Paratropidae (paratropid funnel-web mygalomorphs). Two tarsal claws with hair tufts; labium fused to sternum; 2 to 4 spinnerets; funnel web; limited to Amazon River Basin; about 10 species.

Family Barychelidae (barychelid mygalomorphs). Two tarsal claws; chelicerae with rastellum; inhabit tubes in soil or under tree bark; found in South America, Australia, Africa, and India; about 200 species.

Suborder Labidognatha (araneomorph spiders)

Chelicerae labidognath (diaxial), attached below carapace; pedipalpal coxae with endites; usually 1 pair of book lungs, sometimes replaced by tracheae; heart with 3, sometimes 2, ostia; 13th through 18th ganglia lost, others fused; slightly more than 28,000 species described.

Infraorder Ecribellatae (ecribellate spiders)

Labidognath spiders without cribellum or calamistrum; subdivided into 2 groups; group Haplogynae—male pedipalp a comparatively simple bulb rigidly attached to tarsus, females without epigynum, life-span several years, includes superfamilies Dysderoidea, Scytodoidea, and Pholcoidea; group Entelegynae—male pedipalps more complex, females usually with epigynum (copulatory pores of female on plate separate from birth pore), annual or biannual life cycle, includes superfamilies Zodaroidea, Hersilioidea, Araneoidea, Lycosoidea, Clubionoidea, and Thomisoidea.

Superfamily Dysderoidea. Usually no long plate (lamella) on cheliceral fang margin; respiratory tracheae with 4 spiracles (holes) in 2 pairs, one behind the other (except Telemidae and Leptonetidae); often only 6 eyes; wide distribution.

Family Gradungulidae (gradungulids). Four pairs of lungs; 8 eyes; only a few species in New Zealand and Australia.

Family Oonopidae (oonopids). Most with 6 eyes, closely grouped; book lungs sometimes lacking; less than 3 millimetres long, with short legs; inhabit leaf litter, worldwide mostly in tropical regions; about 250 species.

Family Caponiidae (caponiids). Usually 2, rarely 8, eyes, closely grouped; tracheae replace lungs; chelicerae with lamella opposite fang; found in tropical and subtropical regions; about 50 species.

Family Leptonetidae. Three spiracles; 6 eyes, 4 in an anterior semicircle, 2 close together some distance behind; long-legged sheet web weavers, often inhabit caves; about 30 species found in Mediterranean Sea region, southern U.S., Japan; placement in this superfamily doubtful.

Family Telemidae. Four spiracles, no lungs; may lack eyes; long-legged cave species; a few species found in Europe and Africa; placement in this superfamily doubtful.

Superfamily Scytodoidea. Chelicera with lamella on fang margin; usually 6 eyes; 2 spiracles next to each other, other spiracles posterior.

Family Scytodidae (spitting spiders.) Six eyes arranged in 3 groups; high, domed carapace, slopes anteriorly; catch prey with well-aimed squirt of saliva; mostly in tropical and subtropical regions; about 200 species.

Family Loxoscelidae (brown spiders). Six eyes arranged in 3 groups; carapace low; overpower web-entangled prey; *Loxosceles reclusa* (brown recluse) and *Loxosceles laeta* venomous to humans; about 20 other species found in North and South America and Mediterranean region.

Family Sicariidae (6-eyed crab spiders). Six eyes, low carapace; legs extended toward sides; bury in sand in deserts of Southern Hemisphere; about 40 species.

Family Diguetidae (diguetids). Six eyes; long carapace covered by scales; abdomen covered by scales and hairs; live beneath a vertical silk tube with a horizontal sheet of threads; known only from the North American Southwest and Argentina; only a few species.

Family Plectreuridae (plecteurids). Eight eyes in 2 rows; heavy legs; hangs upside down in loose web under stones in North American Southwest.

Family Ochyroceratidae (ochyroceratids). Six eyes closely grouped in a transverse row; small with long, slender legs; mostly cave inhabitants; about 30 species.

Superfamily Pholcoidea (pholcids). Three claws; male pedipalp complex, female without epigynum; may belong to the entelegyne group of spiders.

Family Pholcidae (daddy longlegs spiders). Tarsi of legs with many false articulations; no tracheae; web loose and tangled; *Pholcus*, cellar spider of Europe and America; about 500 species.

Superfamily Zodarioidea. Three tarsal claws; reduced number of spinnerets.

Family Zodariidae (zodariids). Two, 4, or 6 spinnerets; some species with only 2 claws; hunt on ground; about 300 species.

Family Palpimanidae (palpimanids). Usually 2 spinnerets; 1st legs greatly enlarged, held up when walking; inhabit warm climates; about 100 species.

Superfamily Hersiloidea. Three tarsal claws; endites of pedipalps converge distally; posterior spinnerets with long distal projection.

Family Hersiliidae (hersiliids). Posterior (lateral) spinnerets very long; legs long; tropical spiders, found against bark or walls, head down; about 50 species.

Family Urocteidae (urocteids). Eyes closely grouped far back on carapace; short legs; large anal tubercle; often with brightly marked abdomen; live in silk tubes in Old World; only a few species.

Superfamily Araneoidea. Three tarsal claws; web spinners; anterior and posterior spinnerets short and conical, arranged in a circle; males with complex pedipalps.

Family Theridiidae (cobweb weavers or comb-footed spiders). Often with a ventral comb of serrated bristles on tarsus of each 4th leg, used to throw silk over prey; chelicerae pointed on proximal end under carapace, few teeth; male pedipalp with accessory arm (paracymbium) absent or present as a small hook on side of cuplike modification of tarsus (cymbium); no stridulating (sound-producing) structures on side of chelicerae; nonfunctional cone between anterior spinnerets (colulus) frequently absent; web irregular; sometimes hunt on ground; *Latrodectus* (widows) poisonous to mammals; more than 2,500 species; worldwide distribution.

Family Linyphiidae (sheet-web weavers and dwarf spiders). Chelicerae with many teeth, blunt on proximal end, often with stridulating file on side; legs with many strong setae; colulus always present; paracymbium a separate sclerite in male pedipalps; about 3,500 species; worldwide distribution.

Family Symphytognathidae (symphytognathids). Minute spiders; females may lack pedipalps, or if present, they lack claws; book lungs usually absent; usually with colulus; common in Southern Hemisphere; about 100 species.

Family Hadrotarsidae (hadrotarsids). Minute spiders; book lungs usually absent; anterior median eyes most developed, posterior median eyes bean-shaped; no teeth on chelicerae; found in Australia; only a few species.

Family Nesticidae (nesticids). May have comb on 4th pair of legs; teeth on chelicerae; large colulus; paracymbium of male pedipalp a sclerite attached to cymbium; construct cobwebs; females carry egg sac attached to spinnerets; many live in caves; about 100 species.

Family Araneidae (*Argiopidae*) (orb weavers). Legs with strong setae; chelicerae strong, truncated proximally, with many teeth; colulus always present; paracymbium fused to cymbium of male pedipalp; usually construct orb webs; *Araneus* widespread; *Mastophora*, the bolas spider, probably venomous to man; many species make no webs; wait for prey to stumble into them; over 2,500 species in the world.

Family Theridiosomatidae (ray spiders). Globular abdomen; dark anterior median eyes; high clypeus (area below eyes); chelicerae lack boss; orb webs; probably a subfamily of Tetragnathidae; about 50 species.

Family Tetragnathidae (tetragnathid orb weavers). Males with long chelicerae; epigynum often secondarily lost; pedipalp different in structure from Araneidae; relationship of the two families uncertain; distribution worldwide; about 100 species.

Family Archaeidae (archaeids). Front of carapace pushed up by enormously elongated chelicerae; chelicerae work independently of each other; very small in size; prey on other spiders in Southern Hemisphere; about 50 species.

Family Mimetidae (pirate spiders). Characteristic row of large setae on 1st pair of legs; colulus present; prey on other spiders; distribution worldwide; about 100 species.

Superfamily Lycosoidea. Tarsus with 3 claws, ctenid adults may lose 3rd claw; 1 or 2 rows of trichobothria (sensory hairs) on dorsal side of each tarsus; spinnerets usually neither cone-shaped nor arranged in a circle.

Family Agelenidae (funnel weavers). Eyes in 2 rows; anterior (lateral) spinnerets long; most make funnel web (flat web in vegetation with a tube-shaped retreat at the side); *Argyroneta*, the Eurasian water spider, constructs an air bell beneath surface of slow-moving waters; about 1,000 species; worldwide distribution.

Family Hahniidae (hahniids). Six spinnerets in a single transverse row; small in size; make sheet webs in depressions in ground; worldwide distribution; about 100 species.

Family Pisauridae (nursery web spiders). Hunting spiders; eyes arranged in 3 rows; anterior row with 4 eyes; female carries egg sac in chelicerae; medium to large in size; hunt near or over water; about 400 species.

Family Lycosidae (wolf spiders). Eyes in 3 rows; anterior row with 4 eyes, smaller than large anterior medians of second row; hunting spiders; carry egg sac attached to spinnerets; carry young on abdomen; may dig tubes in soil; numerous species in the Arctic and on high mountains; true tarantula, *Lycosa tarentula* of southern Europe, probably not poisonous to man; about 2,500 species throughout world.

Family Toxopidae (toxopids). Eight eyes in 2 rows; posterior eyes very large; hunting spiders of Australian and New Zealand forests; may be related to the family Selenopidae; a few species.

Family Oxyopidae (lynx spiders). Eyes arranged in a hexagon; hunt on vegetation, pounce on prey; worldwide distribution; about 500 species.

Family Senoculidae (senoculid spiders). Anterior median eyes far anterior to remaining 6; small family of hunting spiders; found in South America; about 40 species.

Family Ctenidae (wandering spiders). Eyes in 3 rows; median eyes face same direction, form a trapezoid wider behind than in front; eyes of posterior row largest, anterior laterals smallest; 1st 2 legs armed with strong ventral setae; some adults with only 2 tarsal claws; aggressive *Phoneutria fera* of South America venomous to man; mainly large tropical spiders; may be related to Clubionidae; about 400 species.

Superfamily Clubionoidea. Two tarsal claws; chelicerae with teeth; eyes in 2 rows; usually no colulus.

Family Gnaphosidae (*Drassodidae;* gnaphosids). Anterior (lateral) spinnerets cylindrical and separated; posterior median eyes often oval and diagonal; nocturnal hunters; common and widespread; about 2,000 species.

Family Clubionidae (sac spiders). Anterior pair of spinnerets conical and touching; *Chiracanthium* species may be venomous to man; common and widespread hunting spiders; about 2,000 species.

Family Prodidomidae (prodidomids). Resemble gnaphosids but differ in having eyes arranged in a circle, with largest gap between posterior median eyes; chelicerae extended; found in tropical regions; about 70 species.

Family Platoridae (platorid crab spiders). Flattened; legs extend sideways; sternum wider than long; anterior spinnerets separated, middle spinnerets large; a few tropical Asian and American species.

Family Homalonychidae (homalonychids). Eyes in 2 rows; 2 smooth claws; a few sluggish species under stones in southwestern U.S.

Family Heteropodidae (*Sparassidae*) (huntsman spiders, tarantulas in Australia). Eyes in 2 rows; legs extended sideways; large, slightly flattened body; found in most tropical regions; about 1,000 species.

Family Selenopidae (selenopid crab spiders). Six of 8 eyes in a single row; legs extend to sides; much flattened body; found under bark or rocks in tropics; about 100 species.

Family Amaurobioididae (amaurobioidids). Eyes in 2 rows; front row strongly curved forward on sides; tarsi and metatarsi with scopulae (brushes); colulus present; 1 species; found in New Zealand, Australia, and South Africa; lives in silk nests in intertidal zone of ocean.

Superfamily Thomisoidea. Two claws; 8 eyes in 2 rows; legs extend sideways; chelicerae usually without teeth opposing fang; colulus present.

Family Thomisidae (crab spiders). May sit on flowers awaiting insects; some change colour; some live on or under bark; many species; common; worldwide distribution; about 3,000 species.

Superfamily Salticoidea. Two claws; anterior median eyes very much larger than others.

Family Salticidae (jumping spiders). Hunt during day; stalk and jump on prey; best vision of all spiders; *Phidippus* common in North America; about 4,000 species; mostly tropical, but many in northern and Arctic regions.

Superfamily Ammoxenoidea. Two tarsal claws; spinnerets as in Urocteidae, anterior far apart, but no colulus; chelicerae fused near base and lacking teeth; legs slender; male pedipalps simple.

Family Ammoxenidae (ammoxenids). Rapid-running red spiders; few species; found in Africa.

Infraorder Cribellatae (cribellate spiders)

Labidognath spiders with specialized structures involved in silk making, the cribellum and the calamistrum.

Family Hypochilidae (hypochilid spiders). Four book lungs, or 2 book lungs and 2 sets of tracheae; pedipalpal coxae with endites; no epigynum in female; *Hypochilus*—several species in North America, orthognath chelicerae, lampshade-shaped web; *Ectatosticta* found in China—chelicerae in intermediate position, now placed in Ectatostictidae; South American species, *Thaida peculiaris*—labidognath chelicerae, tracheae as posterior respiratory organs, now placed in *Thaididae*; *Hickmania troglodytes* of Tasmania—orthognath chelicerae, labium not fused to sternum, now placed in Hickmaniidae; only about a dozen species.

Family Filistatidae (filistatids). Eyes closely grouped; 3 tarsal claws; labium fused to sternum; male pedipalp simple; female without epigynum; chelicerae fused at base; builds tube into crevice with threads radiating from opening to trip prey; inhabits warmer parts of North America and Europe; perhaps 50 species.

Family Acanthoctenidae (acanthoctenids). Adults with 2 tarsal claws; eyes in 3 rows as in Ctenidae; strong leg setae; a few large tropical species.

Family Zoropsidae (zoropsids). Two tarsal claws; eyes in 2 rows; adults with 2 or 3 tarsal claws and claw tufts; about 30 species in Mediterranean area, Mexico, and Central America.

Family Oecobiidae (oecobiids). Eyes closely grouped away from anterior carapace margin; 3 tarsal claws; labium not fused to sternum; anal tubercle large and hairy; small in size, with bodies resembling large Urocteidae; less than 2.5 millimetres in size; found in small webs under stones, on walls, and in houses; inhabit warmer climates; about 50 species.

Family Eresidae (eresids). Three tarsal claws; large, hairy spiders; domed cephalothorax; make funnel-like webs; *Stegodyphus* colonial; about 100 species, all Old World.

Family Tengellidae (tengellids). Three tarsal claws; scopulae (brushes) on 1st tarsi and metatarsi; a few species in Madagascar and Africa.

Family Dictynidae (dictynids). Three tarsal claws; tarsi lack trichobothria and brush of setae; small in size; make irregular webs under leaves or in branches of herbs; common in temperate areas; about 500 species.

Family Psechridae (psechrids). Three tarsal claws with brush of setae between; large in size; make an irregular funnel web; found in moist habitats of India and Ceylon to New Zealand; about 40 species.

Family Amaurobiidae (amaurobiids). Three tarsal claws without brush of setae; tarsi with dorsal row of trichobothria; resemble Agelenidae; make an irregular funnel web between stones; common; about 350 species, worldwide distribution.

Family Uloboridae (uloborids). Lack poison glands; 3 tarsal claws; eyes in 3 rows; anal tubercle large; make orb webs; *Uloborus*, feather-legged spiders, mostly subtropical; *Hyptiotes*, triangle spiders, north temperate; about 200 species, worldwide distribution.

Family Dinopidae (ogre-faced spiders). Three tarsal claws; eyes in 3 rows; anal tubercle large; *Dinopis*, huge anterior eyes, holds web, throws it over prey; about 60 species, found throughout tropical world.

Critical appraisal. The classification given above relies heavily on the work of Eugène Simon. Since the 1930s spiders of the Southern Hemisphere have become better known. New families have been established for some new species; *e.g.*, Hadrotarsidae, Symphytognathidae, Toxopidae, and Grandungulidae. Newly described spiders are more frequently placed in well-known families, which often makes them meaningless as a taxon (*e.g.*, Zodariidae); this procedure, however, may have advantages over that of naming numerous new families of uncertain relationships.

Gerhardt and Kaestner assumed that the presence of a spinning plate (cribellum) and a calamistrum comb indicated common ancestry; Petrunkevitch did not. Evidence indicates that the cribellum was lost in several cribellate lines and that a number of cribellate families have close relatives among spiders lacking this structure (ecribellate spiders). Examples include Acanthoctenidae and Ctenidae; Amaurobiidae and Agelenidae; Oecobiidae and Urocteidae. Filistatids appear to be closely related to the Haplogynae. The Petrunkevitch system of classification has shortcomings, however, because it did not consider the functional relationships of taxonomic features; *i.e.*, he failed to realize that complicated circulatory system is always present when the respiratory system consists of book lungs and that tracheae indicate a simpler circulatory system. In addition, book lungs have transformed into a tracheal system in spiders that, regardless of their affinities, have specialized needs; *i.e.*, those that are small in size and vulnerable to drying, those that are sun loving (Salticidae), or those that live in water. R. Forster, in volumes on New Zealand spiders, pointed out the striking similarities of the complex genitalia in certain cribellates and ecribellates and also suggested that tracheal development might be used to establish relationships. Other characters that might reveal phylogenetic relationships include details of secondary-eye structure, cheliceral structure, musculature, and perhaps the structure of the spinnerets.

BIBLIOGRAPHY. P. BONNET, *Bibliographia Araneorum*, 7 vol. (1945–61), an indispensable work, listing all references to spiders and spider species up to 1938; W.S. BRISTOWE, *The Comity of Spiders*, 2 vol. (1939–41), *The World of Spiders* (1958), both about the habits of spiders; R.R. FORSTER, *The Spiders of New Zealand* (1967), a beautifully illustrated technical guide to New Zealand spiders with a look at relationships; W.J. GERTSCH, *American Spiders* (1949), a natural history of American spiders; A. KAESTNER, *Invertebrate Zoology*, vol. 2, trans. and adapted by H.W. and L.R. LEVI (1969), an expanded textbook of arachnology and myriapodology; B.J. KASTON, *Spiders of Connecticut* (1948), a useful reference on the biology of spiders, especially in the northeastern United States, and *How to Know the Spiders*, 2nd ed. (1972), keys to North American spiders; H.W. and L.R. LEVI, *A Guide to Spiders and Their Kin* (1968), coloured illustrations of spider families; G.H. LOCKET and A.F. MILLIDGE, *British Spiders*, 3 vol., (1951–71), a guide to British spiders; J. MILLOT in P.P. GRASSE (ed.), *Traité de zoologie*, vol. 6 (1949), a good summary of spiders (in French); A. PETRUNKEVITCH, "Chelicerata," in *Treatise on Invertebrate Paleontology*, pt. *P* (1955), on the paleontology of spiders; C.F. ROEWER, *Katalog der Araneae*, 3 vol. (1943–55), a catalogue to spider species, arranged by family (in German); E. SIMON, *Histoire Naturelle des Araignées* (1892–1903), one basis for classification system used above; P.N. WITT, C.F. REED, and D.B. PEAKALL, *A Spider's Web* (1968), an excellent introduction to problems presented by orb-web building spiders.

(H.W.L./L.R.L.)

Arbitration

Arbitration is a legal technique for resolving disputes by referring them to a third party for a binding decision, or "award," as an arbitrator's findings are usually described. The arbitrator may be a single person or an arbitration board, usually of three members. Arbitration is most commonly resorted to for the resolution of commercial disputes and must be distinguished from mediation and conciliation, which are common in the settlement of labour disputes between management and labour unions. In such cases, the parties resort to a third person to offer a recommendation for a settlement or to help them to reach a compromise. Such intervention by a third party, which also occurs in international disputes between states in the form of diplomatic intervention and good offices, has no binding force upon the disputants, as has the arbitrator's decision, the award. Commercial arbitration, labour arbitration, and international arbitration have common features only in their procedural aspects and will be treated separately in this article.

COMMERCIAL ARBITRATION

Commercial arbitration is a means of settling disputes by referring them to a third person, an arbitrator, selected by the parties for a decision based on the evidence and arguments presented to the arbitration tribunal. The

parties agree in advance that the decision will be accepted as final and binding upon them.

Historically, commercial arbitration was used in resolving controversies between medieval merchants, in fairs and marketplaces in England and on the European continent, and in the Mediterranean and Baltic sea trade. The increased use of commercial arbitration became possible after courts were empowered to enforce the parties' agreement to arbitrate. The first such statute was the English Arbitration Act of 1889, now consolidated into an act of 1950 and adopted by arbitration statutes in most countries of the Commonwealth. It was followed in the United States by an arbitration statute of the state of New York in 1920 and a Federal Arbitration Act of 1925. Codified in 1940, the latter deals with the enforcement in federal courts of arbitration agreements and awards in maritime transactions and those involving interstate and foreign commerce. Most states of the United States adopted, sometimes with minor changes, the Uniform Arbitration Act of 1955, as amended in 1956, which had been promoted by the Commissioners on Uniform State Laws and recommended by the American Bar Association. This act provides for the judicial enforcement of an agreement to arbitrate existing and future disputes, thereby making the arbitration agreement no longer revocable, as it had been under common law. It also provides for the substitution of arbitrators in the event of a party failing to select an arbitrator and for a suspension of any court action instituted in contravention of a voluntary arbitration agreement. The courts thereby play an important role in implementing arbitration agreements, making judicial assistance available against a recalcitrant party. This concept of modern arbitration law, which recognizes the irrevocability of arbitration agreements and the enforceability of awards prevails also in the arbitration statutes of nearly all countries of Europe and Asia. Latin American procedural laws generally provide only for court enforcement of agreements to arbitrate existing disputes and do not provide for the enforcement of subsequent disputes that may arise under the arbitration agreement.

Statutory development of arbitration

Function and scope. Arbitration has been used customarily for the settlement of disputes between members of trade associations and between different exchanges in the securities and commodities trade. Form contracts contain a standard arbitration clause referring to the arbitration rules of the respective organization. Numerous arrangements between parties in industry and commerce also provide for arbitration of controversies arising out of contracts for the sale of manufactured goods, for terms of service of employment, for construction and engineering projects, for financial operations, for agency and distribution arrangements, and for many other undertakings.

The usefulness and significance of arbitration is demonstrated by its increasing use by the business community and the legal profession in many countries of the world. The primary advantage is the speed with which controversies can be resolved by arbitration, compared with the long delays of ordinary court procedure. The expert knowledge of arbitrators of the customs and usages of a specific trade makes testimony by others and much documentation unnecessary, thereby eliminating expenses connected with court procedures. The privacy of the arbitration procedure is also much valued by parties to the controversy; situations unfavourable to the party's credit or deficiencies in manufactured goods revealed in arbitration proceedings do not become known to outsiders. There are, however, also disadvantages in the arbitration process. The fact that in Anglo-American practice no reasons are given by the arbitrator to accompany his award prevents the development of a guideline for the further conduct of business relations. This uncertainty resulting from lack of reasoned precedents, moreover, makes the arbitral decision less predictable. Further obstacles to the wider use of commercial arbitration are the divergencies in municipal laws and court decisions that result in different interpretations of similar arbitration questions and the fact that awards are not published: publication of awards, even without identification of the parties, might assist in the establishment of precedents useful in discouraging future disputes on similar issues in a specific branch of industry or commerce.

Advantages of arbitration over court procedure

Procedure. The method of selecting arbitrators is an important aspect of the arbitration process, for the arbitrator's ability and fairness is the decisive element in any arbitration. The general practice is for both parties to select an arbitrator at the time a conflict arises or at the time the arbitration agreement is concluded. The two arbitrators then select a chairman, forming a tribunal. Selection of arbitrators is also often made by agencies administering commercial arbitration under pre-established rules of procedure. These organizations—various trade associations, produce exchanges, and chambers of commerce in many countries—maintain panels of expert arbitrators. The parties may either make their own selection or entrust the appointment of the arbitrators to the organization.

Challenges to the arbitration process are not uncommon. A party may claim, for example, that no valid arbitration agreement came into existence because the person signing the agreement had no authority to do so or that a condition precedent to arbitration had not been fulfilled. More often, arbitration is contested on the ground that the specific controversy is not covered by the agreement. In such cases, the issue of whether the arbitrator has authority to deal with the conflict is usually determined by a court. Further challenges to the arbitration process may be directed against an arbitrator, on grounds, for example, of alleged lack of impartiality. Any such challenge can generally only be maintained after the arbitration has been concluded, as courts are reluctant to interfere with the arbitration process before an award has been rendered.

The arbitration process is governed by the rules to which the parties referred in their agreement; otherwise, the procedure will be determined by the arbitrators. The arbitration proceedings must be conducted so as to afford the parties a fair hearing on the basis of equality. The arbitrator generally has the authority to request the parties and third persons to produce documents and books and to enforce such a request by issuing subpoenas. If a party fails to appear at a properly convened hearing, without showing a legitimate cause, the arbitrator in most instances will proceed in the absence of that party and render an award after investigation of the matter in dispute.

Under the law and arbitration practice of most countries, the award is valid and binding upon the parties when rendered by a majority of the arbitrators, unless the parties expressly request a unanimous decision of the arbitrators, which they seldom do. The statutory law of various countries and the rules of agencies administering commercial arbitration contain provisions on the form, certification, notification, and delivery of the award, with which requirements the arbitrator has to comply.

A much-disputed question in commercial arbitration concerns the law to be applied by the arbitrators. Generally, the award must be based upon the law as determined by the parties in their agreement. This failing, the arbitrator must apply the law he considers proper in accordance with the rules of conflict of laws (*q.v.*). In both cases, the arbitrator will have to take account of the terms of the contract and the usage of the specific trade. If, during any arbitration proceeding, a compromise is reached by the parties, that compromise may be recorded as an award by the arbitrator.

Appeal of arbitration decisions to the courts

Appeals to the courts from the award cannot be excluded by agreement of the parties, since the fairness of the arbitration process as a quasi-judicial procedure has to be maintained. Any court control is, however, confined to specific matters, usually enumerated in the arbitration statutes, such as misconduct of the arbitrator in denying a party the full presentation of its claim or not granting a postponement of the hearing for good cause. A review of the award by a court generally will not deal with the arbitrator's decisions as to facts or with his application of the law. The competence of the

courts is restricted in order not to make the arbitration process the beginning of litigation instead of its end. Recognition of an award and its enforcement will be denied when it appears to be contrary to public policy, as might be the case, for example, in cases involving trusts (monopolies), industrial property rights, or violation of foreign-currency restrictions. An arbitration award has the authority of a court decision and may be enforced by summary court action according to the procedural law of the country in which execution is being sought.

International commercial arbitration. International commercial arbitration between traders of different countries has long been recognized by the business community and the legal profession as a suitable means of settling trade controversies out of court. The procedure in international commercial arbitration is basically the same as in domestic arbitration. In order to establish more uniformity in procedure and to make access to arbitration facilities more easily available, the United Nations economic commissions in 1966 published new rules applying to international arbitration. Those for Europe are contained in the "Arbitration Rules of the United Nations Economic Commission for Europe" and for Asia and the Far East in the "Economic Commission for Asia and the Far East Rules for International Commercial Arbitration."

The development of international commercial arbitration has been furthered by uniform arbitration legislation prepared by the United Nations Conference on International Commercial Arbitration in 1958 and by the Council of Europe and the Inter-American Juridical Committee of the Organization of American States. One particularly difficult problem of international commercial arbitration is the enforcement of awards in a country other than the one in which they were rendered. Statutory municipal laws do not usually contain provisions for the enforcement of foreign awards, and parties are faced with uncertainty about the law and practice of enforcement procedure in a country other than their own.

The aforementioned international agreements, to which most of the trading nations of the world adhere, facilitate the enforcement of foreign awards to the extent that no further action is necessary in the country in which the award was rendered: the opposing debtor must establish that the award has been set aside or that its effects have been suspended by a competent authority, thus shifting the burden of proof of the nonbinding character of the award to the losing party.

Further development of international commercial arbitration is encouraged by the United Nations Commission on International Trade Law, which aims at promoting the harmonization and unification of laws in the field of international commercial arbitration.

LABOUR ARBITRATION

Labour arbitration—the reference of disputes between management and labour unions to an impartial third party for a final resolution—is usually the last step under a collective-bargaining agreement after all other measures to achieve a settlement have been exhausted. Labour arbitration is not, as is commercial arbitration, an auxiliary avenue of justice and thereby a substitute for ordinary court procedure but a technique also for settling or avoiding strikes.

Two major aspects of labour arbitration

Two major aspects of labour arbitration are usually distinguished: arbitration of rights and arbitration of interests. Arbitration of rights refers to the arbitration of an existing labour contract when a dispute over the application of that contract arises between labour and management. Arbitration of interests refers to arbitration between labour and management during the negotiation of a new labour contract.

Arbitration of rights. Arbitration of rights under the terms of a collective-bargaining agreement is employed in the United States far more than in most other countries. Outside the United States, labour courts, industrial courts, or conciliation and arbitration commissions perform the function of arbitrating rights. These bodies are usually appointed by the government, and recourse to them is frequently compulsory.

More than 90 percent of the collective-bargaining agreements in the United States provide for arbitration as a last step in the grievance procedure. Employees, for example, through their union, may present for arbitration complaints concerning such matters as discipline, discharge, and violations of working conditions. Other issues frequently submitted to arbitration customarily concern premium payments and incentive rates, overtime and vacations, Christmas bonuses, seniority rights, and fringe benefits, such as pension and welfare plans.

The arbitrator's decision must be based on the collective-bargaining agreement, which provides for the application of an existing contract to the grievance presented. The question of whether the various steps in the grievance procedure have been complied with before the initiation of the arbitration is usually left to the determination of the arbitrator and not of a court. The question, however, of whether the disputed issue is covered by the collective-bargaining agreement has to be determined by a court and not by the arbitrator. This authority of the courts was upheld by the Supreme Court of the United States in 1960.

Selection of labour arbitrators

The choice of arbitrator is made either by naming him in the agreement or, more often, by leaving the choice open until a dispute has arisen. Frequently, only a single arbitrator is appointed—usually an expert in the field of industrial relations. Alternatively, tripartite arbitration boards are established, both parties appointing their own arbitrator, who acts somewhat as advocate. A neutral chairman is selected either by the parties or by the two party-appointed arbitrators.

A further technique of arbitration of rights is the appointment of a single permanent arbitrator, or "umpire," to resolve disputes for the duration of the collective-bargaining agreement. The umpire will be intimately acquainted with the various economic, financial, and other aspects of the particular industry and will be familiar with the relationship between management and union developed in the past. He sometimes follows precedents, especially those established by his predecessor. This permanent umpire system originated in the United States in the anthracite-coal industry at the beginning of the 20th century and has been employed in such other important industries as newspaper printing and clothing.

Labour arbitrators render binding decisions on the disputes submitted to them. They are not bound by strict rules of court procedure, especially as regards burden of proof and the presentation of evidence. As arbitrators, they have the power to subpoena persons and written evidence. Labour arbitrators tend to evaluate factual evidence rather freely and often reduce penalties imposed upon employees by the management for breach of the labour contract. Even minor questions, such as the use of company time by employees for washups or coffee breaks, are submitted to arbitration, in order to establish precedents in the operation of the plant. Generally, however, arbitrators are not bound to follow previous decisions.

Decisions of labour arbitrators are seldom reviewed by the courts, as awards are usually fully complied with by both parties.

Arbitration of interests. Arbitration of the terms of a new contract, referred to as arbitration of interests, may be instituted if management and the labour union are unable to agree on the conditions of a new contract. In some industries, such as hotels, printing, transit, and utilities, such disputes are submitted to arbitration. In most countries, however, management and union are seldom inclined to forgo resort to lockouts and strikes in an attempt to obtain favourable new contracts, and interest arbitration is thus seldom used.

Compulsory arbitration, directed by legislative fiat, has been a controversial issue in the settlement of industrial disputes. It has been favoured in disputes in the transportation industry, which may involve great public inconvenience, and in disputes in the public-utilities sector

when an immediate danger to public health and safety might occur. Compulsory arbitration has been declared unconstitutional in some states of the United States. More recently, however, it has been adopted as a regular procedure for the settlement of disputes with municipal employees in some U.S. cities.

INTERNATIONAL ARBITRATION

Controversies between sovereign states that are not settled by diplomatic negotiation or conciliation are often referred, by agreement of both parties, to the decision of a third disinterested party, who arbitrates the dispute with binding force upon the disputant parties. Such arbitration between states has a long history: it was used between city-states in ancient Greece and also in the Middle Ages, when the pope often acted as the sole arbitrator.

Historical development. The modern development of international arbitration can be traced to the Jay Treaty of 1794 between Great Britain and the United States, which established three arbitral commissions to settle questions and claims arising out of the American Revolution. In the 19th century many arbitral agreements were concluded by which arbitration tribunals were established ad hoc to deal with a specific case or to handle a great number of claims. Most significant was the "Alabama" claim arbitration under the Treaty of Washington (1871), by which the United States and Great Britain agreed to settle claims arising from the failure of Great Britain to maintain its neutrality during the U.S. Civil War.

Commissions consisting of members drawn from both disputant countries ("mixed arbitral commissions") were often used in the 19th century to settle pecuniary claims for compensation of injuries to aliens for which justice could not be obtained in foreign courts. Such was the purpose of a convention of 1868 between the United States and Mexico, by which claims of citizens of each country arising from the Civil War were settled. Boundary disputes between states were also often settled by arbitration.

Development of international arbitration procedures

International arbitration was given a more permanent basis by the Hague Conference of 1899, which adopted a convention on the pacific settlement of international disputes, revised by a Conference of 1907. The convention stated that:

> International arbitration has for its object the settlement of disputes between States by judges of their own choice and on the basis of respect for law. Recourse to arbitration implies an engagement to submit in good faith to the award.

A Permanent Court of Arbitration, composed of a panel of jurists appointed by the member governments, from which the litigant governments may select the arbitrators, was established at The Hague in 1899.

Twenty cases were arbitrated between 1902 and 1932, but, from that year until 1972, only five cases were dealt with. This was largely because the importance of the Permanent Court of Arbitration was lessened by the Permanent Court of Justice (established in 1922) and its successor, the International Court of Justice. More recently, in 1960, the court, which was originally devised for the settlement of disputes between states, has offered its services for the arbitration of controversies between states and individuals or corporations. Such a dispute was arbitrated in 1970 between a British company and the government of the Democratic Republic of Sudan. The case concerned the repudiation of a contract for building houses in the irrigation zone of the Khashm Al Qirbah Dam in The Sudan.

Arbitration provisions of international treaties. There are several multilateral treaties that provide for the pacific settlement of international disputes by arbitration, including the Geneva General Act for the Settlement of Disputes of 1928, adopted by the League of Nations and reactivated by the General Assembly of the United Nations in 1949, which provides for the settlement of various disputes, after unsuccessful efforts at conciliation, by an arbitral tribunal of five members. Other such treaties include the General Treaty of Inter-American Arbitration, signed in Washington in 1929, and the American Treaty on Pacific Settlement of Disputes, signed in Bogotá in 1948. More recently, the Council of Europe adopted the European Convention for the Peaceful Settlement of Disputes (1957).

Arbitration is also mentioned as a proper method of settling disputes between countries in the Charter of the United Nations, as it was in the Covenant of the League of Nations.

Model rules of the International Law Commission

The International Law Commission of the United Nations submitted to the General Assembly in 1955 a Convention on Arbitral Procedure. Its model rules would not become binding on any member-state of the United Nations unless they were accepted by a state in an arbitration treaty or in a special arbitral agreement. The model rules, however, have not yet been adopted in any arbitration arrangement between disputant governments, although in 1958 the General Assembly recommended the model rules for use by member-states when appropriate. It seems clear that states prefer flexibility in the resolution of their disputes by arranging the rules and proceedings of an arbitration according to circumstances.

Great impediments, indeed, still exist in the acceptance of international arbitration, especially in cases in which disputes between governments and foreign private parties are involved. In such cases, the state will often insist that its own local remedies—administrative and court proceedings—be exhausted. Generally, the government of the national who advances a claim against a foreign government will require evidence that the injured party has pursued all remedies in the foreign country before it presses a claim for international negotiation and adjudication, if, indeed, it decides to take up the case at all. Contracting parties may agree in their contract that they need not exhaust local remedies before resorting to arbitration, and one 1965 instrument, the Convention on the Settlement of Investment Disputes, states:

> Consent of the parties to arbitration under this Convention shall, unless otherwise stated, be deemed consent to such arbitration to the exclusion of any other remedy. A Contracting State may require the exhaustion of local administrative or judicial remedies as a condition of its consent to arbitration under this Convention.

The arbitration agreement in a general multilateral treaty, a bilateral convention, or in a specific contractual arrangement between two states often does not deal with the selection of the arbitrators and the appointment of an umpire, the procedure to be followed in the arbitration, the subject matter of the dispute, the specific issues to be submitted, the presentation of evidence, the place of the hearings, the law to be applied by the arbitrators, and the time when the award has to be rendered. These questions are usually dealt with in an agreement between the parties to the dispute known as compromis. If the compromis fails in some particular—to define the applicable law, for example—it is generally understood that the arbitrator shall apply international law.

An award rendered by an arbitral tribunal is customarily complied with by states: it is, in fact, invariably the case that unless a state is prepared to comply with an adverse decision, it will not submit the dispute to arbitration. The difficulties in the use of international arbitration thus consist less in the enforcement of arbitral awards than in persuading states involved in disputes to submit them to a third party, an arbitrator, or an arbitration tribunal.

BIBLIOGRAPHY

Commercial arbitration: The leading treatises and handbooks on commercial arbitration in English are FRANCIS RUSSELL, *On the Law of Arbitration*, 18th ed. by ANTHONY WALTON (1970); MARTIN DOMKE, *The Law and Practice of Commercial Arbitration* (1968); NRISINHADAS BASU, *The [Indian] Arbitration Act*, 5th ed. by S.K. BOSE (1965); and the INTERNATIONAL ASSOCIATION OF LAWYERS, *International Commercial Arbitration: A World Handbook*, ed. by PIETER SANDERS, 3 vol. (1956–65). The history of commercial arbitration is dealt with admirably by WILLIAM C. JONES in "An Inquiry into the History of the Adjudication of Mercantile Disputes in Great Britain and the United States," *University of Chicago Law*

Review, 25:445–464 (1958). Current information on arbitration agreements, proceedings, and awards, on legislation, and court decisions may be found in the following quarterly arbitration journals: *Arbitration* (Institute of Arbitrators, London), the *Arbitration Journal* (American Arbitration Association, New York), *Arbitration Quarterly of the India Council of Arbitration* (New Delhi), and the *Quarterly of the Japan Commercial Arbitration Association* (Tokyo). Documentation on the activities of the United Nations Commission on International Trade Law in the field of international commercial arbitration is included in its *Yearbook*.

Labour arbitration: FRANK and EDNA ELKOURI, *How Arbitration Works*, rev. ed. (1960); and CLARENCE M. UPDEGRAFF and WHITLEY P. MCCOY, *Arbitration of Labor Disputes*, 2nd ed. (1961), are good accounts of the law and practice of labour arbitration. The historic development is described by E.E. WITTE, *Historical Survey of Labor Arbitration* (1952). Representative of labour arbitration practice are the *Proceedings* of the National Academy of Arbitrators (published annually since 1955); and MORRIS STONE, *Labor Grievances and Decisions* (1965, and *New Series* 1970). Foreign industrial arbitration is dealt with in an article by BENJAMIN AARON, "Labour Courts: Western European Models and Their Significance for the United States," *U.C.L.A. Law Review*, 16:847–882 (1969) and in his edition of *Labour Courts and Grievance Settlement in Western Europe* (1971).

International arbitration: Various aspects of international arbitration are dealt with by J.L. SIMPSON and HAZEL FOX, *International Arbitration: Law and Practice* (1959); in a *Report of a Study Group on the Pacific Settlement of International Disputes* (David Davies Memorial Institute 1966); and in *International Arbitration: Liber Amicorum for Martin Domke*, ed. by PIETER SANDERS (1967). "The Changing Role of Arbitration in the Settlement of International Disputes" is the topic of a well-documented report by LOUIS B. SOHN in the Helsinki Conference Report of the International Law Association, pp. 325–356 (1967). There are valuable references on the history of international arbitration in A.M. STUYT, *Survey of International Arbitrations, 1794–1938* (1939); and in JOHN BASSETT MOORE, *History and Digest of the International Arbitrations to Which the United States Has Been a Party*, 6 vol. (1898). The *Reports of International Arbitral Awards*, 16 vol., have been published by the United Nations (1948–69). Further important UN publications are the *Systematic Survey of Treaties for the Pacific Settlement of International Disputes, 1928–1948* (1949) and *A Survey of Treaty Provisions for the Pacific Settlement of International Disputes, 1949–1962* (1966).

(M.Do.)

Archaeology

The word archaeology comes from the Greek *archaia* ("ancient things"), and *logos* ("theory or science"). It has been used in varying ways, but from the late 18th century onward has come to mean that branch of learning that studies the material remains of man's past. This includes man's artifacts, from the very earliest stone tools, of perhaps 2,000,000 years ago, to the man-made objects that are buried or thrown away at the present day: everything made by human beings—from simple tools to complex machines, from the earliest houses and temples and tombs to palaces, cathedrals, and pyramids.

The archaeologist is first a descriptive worker: he has to describe, classify, and analyze the artifacts he studies. An adequate and objective taxonomy is the basis of all archaeology, and many good archaeologists spend their lives in this activity of description and classification. But the main aim of the archaeologist is to place the material remains in historical contexts, to supplement what may be known from written sources, and, thus, to increase understanding of the past. Ultimately, then, the archaeologist is a historian: his aim is the interpretive description of the past of man.

Increasingly, many scientific techniques are used by the archaeologist, and he uses the scientific expertise of many persons who are not archaeologists in his work. The artifacts he studies must often be studied in their environmental contexts; and botanists, zoologists, soil scientists, and geologists may be brought in to identify and describe plants, animals, soils, and rocks. Radioactive carbon dating, which has revolutionized much of archaeological chronology, is a by-product of research in atomic physics. But although archaeology uses extensively the methods, techniques, and results of the physical and biological sciences, it is not a natural science; some consider it a discipline that is half science and half humanity. Perhaps it is more accurate to say that the archaeologist is first a craftsman, practicing many specialized crafts (of which excavation is the most familiar to the general public), and then a historian.

The nature of archaeological work

The justification for his work is the justification of all historical scholarship, namely, to enrich the present by our knowledge of the experiences and achievements of our predecessors. Because it concerns itself with things people have made, the most direct findings of archaeology bear on the history of art and technology; but, by inference it also yields information about the society, religion, and economy of the people who created the artifacts. Also, it may bring to light and interpret previously unknown written documents, providing an even more certain form of evidence about the past.

But no one archaeologist can cover the whole range of man's history, and there are many branches of archaeology divided by geographical areas (such as classical archaeology, the archaeology of ancient Greece and Rome; or Egyptology, the archaeology of ancient Egypt) or by periods (such as medieval archaeology and industrial archaeology). Writing began 5,000 years ago in Mesopotamia and Egypt; its beginnings were somewhat later in India and China, and later still in Europe. The aspect of archaeology that deals with the past of man before he learned to write has, since the middle of the 19th century, been referred to as prehistoric archaeology, or prehistory. In prehistory the archaeologist is paramount, for here the only sources are material and environmental.

The scope of this article is to describe briefly how archaeology came into existence as a learned discipline; how the archaeologist works in the field, museum, laboratory, and study; and how he assesses and interprets his evidence and transmutes it into history.

THE HISTORY OF ARCHAEOLOGY

No doubt there have always been people who were interested in the material remains of the past, but archaeology as a discipline has its earliest origins in 15th- and 16th-century Europe, when the Renaissance Humanists looked back upon the glories of Greece and Rome. Popes, cardinals, and noblemen in Italy in the 16th century began to collect antiquities and to sponsor excavations to find more works of ancient art. These collectors were imitated by others in northern Europe who were similarly interested in antique culture. All this activity, however, was still not archaeology in the strict sense. It was more like what would be called art collecting today.

The Mediterranean and the Near East. Archaeology proper began with an interest in the Greeks and Romans and first developed in 18th-century Italy with the excavations of the Roman cities of Pompeii and Herculaneum. Classical archaeology was established on a more scientific basis by the work of Heinrich Schliemann, who investigated the origins of Greek civilization at Troy and Mycenae in the 1870s; of M.A. Biliotti at Rhodes in this same period; of the German Archaeological Institute under Ernst Curtius at Olympia from 1875 to 1881; and of Alexander Conze at Samothrace in 1873 and 1875. Conze was the first person to include photographs in the publication of his report. Schliemann had intended to dig in Crete but did not do so, and it was left to Arthur Evans to begin work at Knossos in 1900 and to discover the Minoan civilization, ancestor of classical Greece.

Egyptian archaeology began with Napoleon's invasion of Egypt in 1798. He brought with him scholars who set to work recording the archaeological remains of the country. The results of their work was published in the *Description de l'Égypte* (1808–25). As a result of discoveries made by this expedition, Jean-François Champollion was able to decipher ancient Egyptian writing for the first time in 1822. This decipherment, which enabled scholars to read the numerous writings left by the Egyptians, was the first great step forward in Egyptian archaeology. The demand for Egyptian antiquities led to organized tomb robbing by men such as Giovanni Battista Belzoni. A new era in systematic and controlled archae-

The development of Egyptology

ological research began with the Frenchman Auguste Mariette, who also founded the Egyptian Museum at Cairo. The British archaeologist Flinders Petrie, who began work in Egypt in 1880, made great discoveries there and in Palestine during his long lifetime. Petrie developed a systematic method of excavation, the principles of which he summarized in *Methods and Aims in Archaeology* (1904). It was left to Howard Carter and Lord Carnarvon to make the most spectacular discovery in Egyptian archaeology, that of the tomb of Tutankhamen in 1922.

Mesopotamian archaeology also began with hectic digging into mounds in the hopes of finding treasure and works of art, but gradually these gave way in the 1840s to planned digs such as those of the Frenchman Paul-Émile Botta at Nineveh and Khorsabad, and the Englishman Austen Henry Layard at Nimrud, Kuyunjik, Nabī Yūnus, and other sites. Layard's popular account of his excavations, *Nineveh and Its Remains* (1849), became the earliest and one of the most successful archaeological best sellers. In 1846 Henry Creswicke Rawlinson became the first man to decipher the Mesopotamian cuneiform writing. Toward the end of the 19th century, systematic excavation revealed a previously unknown people, the Sumerians, who had lived in Mesopotamia before the Babylonians and Assyrians. The most impressive Sumerian excavation was that of the Royal Tombs at Ur by Leonard Woolley in 1926.

First steps to archaeology. The development of scientific archaeology in 19th-century Europe from the antiquarianism and treasure collecting of the previous three centuries was due to three things: a geological revolution, an antiquarian revolution, and the propagation of the doctrine of evolution. Geology was revolutionized in the early 19th century with the discovery and demonstration of the principles of uniformitarian stratigraphy (which determines the age of fossil remains by the stratum they occupy below the earth) by men like William Smith, Georges Cuvier, and Charles Lyell. Lyell, in his *Principles of Geology* (1830–33), popularized this new system and paved the way for the acceptance of the great antiquity of man. Charles Darwin regarded Lyell's *Principles* as one of the two germinal works in the formation of his own ideas on evolution. Man's early stone tools had first been discovered and identified in the late 18th century. The location of these genuine artifacts in certain strata demonstrated, on the principle of uniformitarianism, that man had been in existence long before 4004 BC, the date calculated according to the chronology of the Bible. The discoveries of Jacques Boucher de Perthes in the Somme Valley in France, and of William Pengelly in the caves of South Devon in England, were used to demonstrate the antiquity of man in 1859, the same year that saw the publication of Darwin's revolutionary *Origin of Species*. Approximate dates for the Paleolithic Period (Old Stone Age) of man's prehistoric past were thus established, although the expression "Palaeolithic" was not used until John Lubbock coined it in his book *Pre-historic Times* (1865).

The age of the human species

Half a century before this, Scandinavian archaeologists had created a revolution in antiquarian thought by postulating, on archaeological grounds, successive technological stages in man's past. C.J. Thomsen classified the material in the Copenhagen Museum, opened to the public in 1819, on the basis of three successive ages of Stone, Bronze, and Iron. His pupil and successor, J.J.A. Worsaae, showed the correctness of this museum arrangement by observed stratigraphy in the Danish peat bogs and barrows (funerary mounds). Low lake levels in Switzerland in the mid-1850s permitted the excavation of the prehistoric Swiss lake dwellings, and here again the theory of a succession of technological stages was confirmed.

Darwin's *Origin of Species* implied a long past for man, and the acceptance of the idea of human evolution in the last four decades of the 19th century created a climate of thought in which archaeology flourished and that led to great advances in the unfolding of the full story of man's development.

In his *Pre-historic Times*, Lubbock expanded the three-age system of Thomsen and Worsaae to a four-age system, dividing the Stone Age into Old and New periods (Paleolithic and Neolithic). In the last quarter of the 19th century remarkable Paleolithic discoveries were made in France and Spain; these included the discovery and authentication of actual works of sculpture and cave paintings from the Upper (later) Paleolithic Period (*c.* 30,000–*c.* 10,000 BC). When Marcellino de Sautuola discovered the cave paintings at Altamira, Spain (1875–80), most experts refused to believe they were Paleolithic; but after similar discoveries at Les Eyzies in France around 1900, they were accepted as such and were recognized as one of the most surprising and exciting archaeological discoveries. A succession of similar finds has continued in the 20th century. The most famous of these was at Lascaux, France, in 1940.

During the last quarter of the 19th century, Gen. A.H. Pitt-Rivers' excavations of prehistoric and Roman sites at Cranborne Chase, Dorset, laid the foundations of modern scientific archaeological field technique, which was later developed and improved in England and Wales by men such as Sir Mortimer Wheeler and Sir Cyril Fox.

20th-century developments. The 20th century has seen the extension of archaeology outside the areas of the Near East, the Mediterranean, and Europe, to other parts of the world. In the early '20s, excavations at Mohenjo-Daro and Harappā, in present Pakistan, revealed the existence of the prehistoric Indus civilization. In the late '20s, excavations at An-yang in eastern China established the existence of a prehistoric Chinese culture that could be identified with the Shang dynasty of early Chinese records.

Studies of the Stone Age

The Stone Age has been described and studied throughout the world; among the most sensational discoveries are those of L.S.B. Leakey, who found stone tools and skeletal remains of early man dating back 2,000,000 years in the Olduvai Gorge in Tanzania. Intensive work of great importance has brought to light early Neolithic sites at Jericho in Palestine; Hassuna, Iraq; Çatalhüyük, Turkey; and elsewhere in the Near East, establishing the origins of agriculture in that region.

Serious archaeological work began later in America than Europe, but as early as 1784 Thomas Jefferson had excavated mounds in Virginia and made careful stratigraphical observations. The 20th century has seen a great increase in archaeological knowledge about prehistoric America: two startling advances were the discovery of the origin of domesticated crops (including maize) in Central America and of the Olmec civilization of Mexico (1000–300 BC)—the oldest of the New World civilizations and probably the parent of all the others.

The enormous growth of archaeological work has meant the establishment of archaeology as an academic discipline; few important universities anywhere in the world are now without professors and departments of archaeology. There are now a very large number of scholarly journals in the field, as well as a considerable body of popularized books and journals that attempt to bridge the gap between professional and layman.

FIELDWORK

Preliminary work. Some archaeologists call everything they do out-of-doors fieldwork, but others distinguish between fieldwork, in a narrower sense, and excavation. Fieldwork, in the narrow sense, consists of the discovery and recording of archaeological sites and their examination by methods other than the use of the spade and the trowel. Sites hitherto unknown are discovered by walking or motoring over the countryside: deliberate reconnaissance is an essential part of archaeological fieldwork.

In Europe, a study of old records and place-names may lead to the discovery of long-forgotten sites. The mapping of new and old sites is an essential part of archaeological survey. This process has been brought to a very high standard of perfection, both in the marking of archaeological sites on ordinary topographical maps and in the production of special period maps. The distribution

map of artifacts, especially when studied against the background of the natural environment, is a key method of archaeological study.

Aerial photography in archaeology

The formerly earthbound archaeologist has been greatly helped by the development of aerial photography. The application of aerial photography to archaeological investigation began in a small way during World War I, as a side effect of military reconnaissance, and was given further impetus by World War II; the photographic intelligence departments of all the combatant nations were extensively staffed by archaeologists, who then carried their expertise and enthusiasm into the postwar years. The University of Cambridge now has its own department of air photography under J.K.S. St. Joseph: using its own pilot and aircraft, it flies photographic missions over Ireland, Great Britain, Denmark, and The Netherlands. The number of new sites discovered each year by aerial photography is very large. Some of these are surface sites, especially partly destroyed sites that show up well in special conditions of light, as in early morning or late evening. But many are sites that could not be found on the ground and that show up in aerial photographs as variations in soil colour or in the density of crop.

Archaeological reconnaissance may be advanced from ordinary surface or aerial methods in a wide variety of ways. A very simple method is tapping the ground to sound for substructures and inequalities in the subsoil. Deep probes have made it possible to trace walls and ditches. The Lerici Foundation of Milan and Rome has had great success with this method since its development of the Nistri periscope, first used in 1957 in an Etruscan tomb in the cemetery of Monte Abbatone. The periscope is inserted into the burial chamber and can photograph the walls and contents of the whole tomb.

Other modern techniques that have been applied to archaeological prospecting employ electricity and magnetic fields (geophysical prospecting). A method of electrical prospecting had been developed in large-scale oil prospecting: this technique, based on the degree of electrical conductivity present in the soil, began to be used by archaeologists in the late 1940s and has since proved very useful. Magnetic methods of prospecting detect buried features by locating the magnetic disturbances they cause: these were introduced in 1957–58 and use such machines as the proton magnetometer, the proton gradiometer, and the fluxgate gradiometer. An American expedition discovered the site of Sybaris in Sicily by magnetic prospecting. Electromagnetic methods have been in use only since 1962; they employ developments of the concepts used in mine detectors. Instruments such as the pulsed-induction meter and the soil-conductivity meter detect magnetic soil anomalies, but only if the features are fairly shallow.

Excavation. Excavation is the surgical aspect of archaeology: it is surgery of the buried landscape and is carried out with all the skilled craftsmanship that has been built up in the last hundred years since Schliemann and Flinders Petrie. Excavations can be classified, from the point of view of their purpose, as planned, rescue, or accidental. Most important excavations are the result of a prepared plan—that is to say, their purpose is to locate buried evidence about an archaeological site. Many are project oriented: as, for example, when a scholar studying the life of the pre-Roman, Celtic-speaking Gauls of France may deliberately select a group of hill forts and excavate them, as Sir Mortimer Wheeler did in northwestern France in the years before the outbreak of World War II. But many excavations, particularly in the heavily populated areas of central and northern Europe, are done not from choice but from necessity. Gravel digging, clearing the ground for airports, quarrying, road widening and building, the construction of houses, factories, and public buildings frequently threaten the destruction of sites known to contain archaeological remains. Emergency excavations then have to be mounted to rescue whatever knowledge of the past can be obtained before these remains are obliterated forever. Partial destruction of cities in western Europe by bombing during World War II allowed rescue excavations to take place before rebuilding. A temple of Mithras in the City of London, Viking settlements in Dublin and at Århus, Denmark, and the original 6th-century-BC Greek settlement of Massalia (Marseille) were discovered in this way. An extension of the runways at London Airport led to the discovery of a pre-Roman Celtic temple there.

Archaeology in urban areas

The role of chance in the discovery of archaeological sites and portable finds is considerable. Farmers have often unearthed archaeological finds while plowing their fields. The famous painted and engraved Upper Paleolithic cave of Lascaux in southern France was discovered by chance in 1940: four French schoolboys were hunting for rabbits when their dog fell down a hole; they climbed down to retrieve him and found themselves in the middle of this remarkable pagan sanctuary. Similarly, the first cache of the Dead Sea Scrolls was discovered in 1947 by a Bedouin looking for a stray animal. These accidental finds often lead to important excavations. At Barnénès, in north Brittany, a contractor building a road got his stone from a neighbouring prehistoric cairn (burial mound) and, in so doing, discovered and partially destroyed a number of prehistoric burial chambers. The French archaeologist P.R. Giot was able to halt these depredations and carry out scientific excavations that revealed Barnénès to be one of the most remarkable and interesting prehistoric burial mounds in western Europe.

Excavation method

All forms of archaeological excavation require great skill and careful preparation. Years of training in the field, first as an ordinary digger, then as a site supervisor, with spells of work as recorder, surveyor, and photographer, are required before anyone can organize and direct an excavation himself. Most museums, universities, and government archaeological departments organize training excavations. The very words dig and digging may give the impression to many that excavation is merely a matter of shifting away the soil and subsoil with a spade or shovel; the titles of such admirable and widely read books as Leonard Woolley's *Spadework* (1953) and *Digging Up the Past* (1930) and Geoffrey Bibby's *Testimony of the Spade* (1956) might appear to give credence to that view. Actually, much of the work of excavation is careful work with trowel, penknife, and brush. It is often the recovery of features almost indistinguishable from non-archaeological aspects of the buried landscape: one example of this is the recovery of mud-brick walls in Mesopotamia; another is the tracing of collapsed walls of dry stone slabs in a cairn in stony country in the southwest Midlands of England. Sometimes it is the recovery of features of which only ghost traces remain, like the burnt-out bodies from the buried city of Pompeii, or the strings of a harp included among the furnishings of Mesopotamian tombs at Ur.

Because of the damage he may cause by inexperience and haste, the untrained amateur archaeologist often hinders the work of the professional. Amateur archaeology is forbidden in many countries by stringent antiquity laws. At the same time, it is certainly true that nonprofessionals have made important contributions in many areas of archaeology. Occasionally, an amateur does make an important discovery the further excavation of which can then be taken over by trained professionals. Such was the case at Sutton Hoo in Suffolk in 1939, when work begun by a competent amateur was taken over by a team of experts who were able to uncover a great Anglo-Saxon burial boat and its treasure, without doubt the most remarkable archaeological find ever made in Britain.

There are, of course, many different types of archaeological sites, and there is no one set of precepts and rules that will apply to excavation as a whole. Some sites, such as temples, forts, roads, villages, ancient cities, palaces, and industrial remains, are easily visible on the surface of the ground. Among the most obvious archaeological sites that have yielded spectacular results by excavation are the huge man-made mounds (tells) in the Near East, called in Arabic *tilāl*, and in

Turkish *tepes* or *hüyüks*. They result from the accumulation of remains caused by centuries of human habitation on one spot. The sites of the ancient cities of Troy and Ur are examples. Another type consists of closed sites such as pyramids, chambered tombs, barrows (burial mounds), sealed caves, and rock shelters. In other cases there are no surface traces, and the outline of suspected structures is revealed only by aerial or geophysical reconnaissance as described above. Finally, there are sites in cliffs and gravel beds, where many Paleolithic finds have been made.

The wide range of techniques employed by the archaeologist vary in their application to different kinds of sites. The opening of the tomb chamber in an Egyptian pyramid is, for example, a very different operation from the excavation of a tell in Mesopotamia or a barrow grave in western Europe. Some sites are explored provisionally by sampling cuts known as sondages. Large sites are not usually dug out entirely, although a moderate-sized round barrow may be completely removed from the earth. Whatever the site and the extent of the excavation, one element of the technique is common to all digs, namely, the use of the greatest care in the actual surgery and in the recording of what is found by word, diagram, survey, and photography. To a certain extent all excavation is destruction, and the total excavation of a site subsequently engulfed by a housing estate or gravel digging is total destruction. This is why the archaeologist's field notes and his published report become primary archaeological documents. They are not themselves, strictly speaking, archaeological facts: they are the excavator's interpretation of what he saw, or thought he saw, but this is the nearest the discipline can ever get to archaeological facts as established by excavation. The really great excavators leave such a fine record of their digs that subsequent archaeologists can re-create and reintrepret what they saw and found. To delay publishing the results of an excavation within a reasonable time is a serious fault from the point of view of archaeological method. An excavation is not complete until the printed report is available to the world. Often the publication of the report takes as long as, or much longer than, the actual work in the field.

Publication of results

When a site like the Palace of Minos at Knossos or the city of Harappā in Pakistan has been excavated, and the excavations are over, the excavator and the antiquities service of the country concerned have to face the problem of what to do with the excavated structures. Should they be covered in again, or should they be preserved for posterity, and if preserved, what degree of conservation and restoration is permissible? This is the same kind of problem that arises in connection with the removal of antiquities from their homeland to foreign museums, and there is no generally accepted answer to it. These problems remain to beset archaeology: should Sir Arthur Evans have reconstructed the Palace of Minos at Knossos? Should the art treasures of ancient Greece and Egypt, now in western European museums, be returned? There is no simple, straightforward, overall answer to these difficult questions.

Underwater archaeology. Underwater archaeology is a branch of reconnaissance and excavation that has been developed only during the 20th century. It involves the same techniques of observation, discovery, and recording that are the basis of archaeology on land, but adapted to the special conditions of working underwater. It is obvious that no archaeologist working on submarine sites can get far unless he is trained as a diver. Helmeted sponge divers have made most of the important archaeological discoveries in the Mediterranean. The French scientist Jacques-Yves Cousteau developed the self-contained breathing apparatus known as the scuba, of which the most commonly used type is the aqualung. Cousteau's work at Le Grand Congloué near Marseille was a pioneer underwater excavation, as was the work of the Americans Peter Throckmorton and George Bass off the coast of southern Turkey. In 1958 Throckmorton found a graveyard of ancient ships at Yassı Ada and then discovered the oldest shipwreck ever recorded, at Cape Gelidonya—a Bronze Age shipwreck of the 14th century BC. George Bass of the University of Pennsylvania worked on a Byzantine wreck as Yassı Ada from 1961 onward, developing the mapping of wrecks photogrammetrically with stereophotographs and using a two-man submarine, the "Asherah," launched in 1964. The "Asherah" was the first submarine ever built for archaeological investigation.

INTERPRETATION

Excavation often seems to the general public the main and certainly the most glamorous aspect of archaeology; but fieldwork and excavation represent only a part of the archaeologist's work. The other part is the interpretation in cultural and historical contexts of the facts established—by chance, by fieldwork, and by digging—about the material remains of man's past. This task of interpretation has five main aspects.

Classification and analysis. The first concern is the accurate and exact description of all the artifacts concerned. Classification and description are essential to all archaeological work, and, as in botany and zoology, the first requirement is a good and objective taxonomy. Second, there is a need for interpretive analysis of the material from which artifacts were made. This is something that the archaeologist himself is rarely equipped to do; he has to rely on colleagues specializing in geology, petrology (analysis of rocks), and metallurgy. In the early 1920s, H.H. Thomas of the Geological Survey of Great Britain was able to show that stones used in the construction of Stonehenge (a prehistoric construction on Salisbury Plain in southern England) had come from the Prescelly Mountains of north Pembrokeshire; and he established as a fact of prehistory that over 4,000 years ago these large stones had been transported 200 miles from west Wales to Salisbury Plain. Detailed petrological analysis of the material of Neolithic polished stone axes have enabled archaeologists to establish the location of prehistoric ax factories and trade routes. It is also now possible, entirely on a petrological basis, to study the prehistoric distribution of obsidian (a volcanic glass used to make primitive tools).

Petrological analysis

In the third place, the archaeologist, having dealt with the material of his artifacts by classification and taxonomy, and with its physical nature by petrology and metallurgy, turns to the remaining information he can get from his colleagues in the natural sciences. These tell him the environmental conditions in which the people he is studying lived; he now sees his material remains not as isolated artifacts but in the context of their original environments.

Dating. Having analyzed his discoveries according to their form, material, and biological association, the archaeologist then comes to the all-important problem of dating. Many material remains of man's past have no dating problem: they may be, like coins, or most coins, self-dating, or they may be dated by man-made dates in written records. But the great and difficult part of the archaeologist's work is dating material remains that are not themselves dated. This can be done in one of three ways. Sometimes an object from another culture, the date of which is known (*e.g.*, in the case of pottery, by its style), is found at a previously undated site. Then, using the relative dating principle (see below) the archaeologist reasons that the material found with the imported object is contemporary with it. Conversely, an object from an undated culture may be found at a site whose date is known. Thus nonliterate communities can be dated by their contact with literate ones. This technique is known as cross dating; it was first developed by Sir Flinders Petrie when he dated Palestinian and early Greek (Aegean) sites by reference to Egyptian ones. Much of the prehistoric chronology of Europe in the Neolithic, Bronze, and Early Iron ages is based on cross dating with the ancient Near East.

Aside from cross dating, the archaeologist faced with material in a site having no literate chronological evidence of its own has two other ways of dating his material. The first is relative, the second absolute. Rela-

Relative dating

tive dating merely means the relation of the date of anything found to the date of other things found in its immediate neighbourhood. As has already been described, this method also plays a part in cross dating. Stratigraphy is the essence of relative dating. The archaeologist observes the accumulation of deposits in a gravel pit, a peat bog, in the construction of a barrow, or in accumulated settlements in a tell, and, like the geologists who introduced the principles of stratigraphy in the late 18th and early 19th centuries, he can see the succession of layers in the site and can then establish the chronology of different levels of layers relative to each other. In the excavation of a great tell like Ur or Troy the relative chronology of the various levels of occupation is the first thing to be established. Some archaeologists, even until quite recent times, have mistakenly supposed that depth below ground level is itself an indication of antiquity.

But even in properly observed and recorded stratigraphic levels there is often doubt, and the question arises: are all the artifacts and human remains found in the same level contemporary? Is it possible that there could have been later intrusions that have been difficult to distinguish in the field? The analysis of the fluorine content of bones has been very helpful here. Recognized as a valuable technique by French scientists in the 19th century, it was developed in England by K.P. Oakley in the 1950s. If bones in apparently the same geological or archaeological level have markedly different fluorine content, then it is clear that there must be interference—for example, by a later burial, or by deliberate planting of faked remains, as happened in the case of the Piltdown "man" hoax in England.

Absolute man-made chronology based on king lists and records in Egypt and Mesopotamia goes back only 5,000 years. For a long time archaeologists searched for an absolute chronology that went beyond this and could turn their relative chronologies into absolute dates. Clay-varve counting seemed to provide the first answer to this need for a nonhuman absolute chronology. Called geochronology by Baron Gerard de Geer, its Swedish inventor, this method was based on counting the thin layers of clay left behind by the melting glaciers when the European Ice Age came to an end. This gave a chronology of about 18,000 years—three times as long as the man-made chronology based on Egyptian and Mesopotamian king lists. Thus, absolute dates could be established for artifacts from the Late Paleolithic Period, the whole of the Mesolithic Period, or Middle Stone Age, and much of the Early Neolithic Period.

Dendrochronology, the dating of trees by counting their growth rings, was first developed for archaeological purposes by A.E. Douglass in the United States. The application of this method to archaeology depends, obviously, on the use in antiquity of old datable trees in the construction of houses and buildings. It has been possible by dendrochronology to date prehistoric American sites as far back as the 3rd and 4th centuries BC.

Radioactive carbon dating

The greatest revolution in prehistoric archaeology occurred in 1948, when Willard F. Libby, at the University of Chicago, developed the process of radioactive carbon dating. In this method, the activity of radioactive carbon (carbon-14) present in bones, wood, or ash found in archaeological sites is measured. Because the rate at which this activity decreases in time is known, the approximate age of the material can be determined by comparing it to carbon-14 activity in presently living organic matter. There have been problems and uncertainties about the application of the radioactive carbon method, but, although it is less than perfect, it has given archaeology a new and absolute chronology that goes back 40,000 years.

Following the revolutionary discovery of radioactive carbon dating, other physical techniques of absolute dating were developed, among them potassium–argon dating and dating by thermoluminescence. Potassium–argon dating has made it possible to establish that the earliest remains of man and his artifacts in East Africa go back at least 2,000,000 years, and probably further.

Historical judgments. The last and most important task of the archaeologist is to transmute his interpretation of the material remains he studies into historical judgments. When he is dealing with medieval and modern history he is often doing no more than adding to knowledge already available from documentary sources: but even so his contribution is often of great importance; for example, in relation to the growth and development of towns and the study of deserted medieval villages. When he is dealing with ancient history and prehistory, he is making a contribution of the greatest importance and often one that is more important than that of purely literary and epigraphical sources. For the prehistoric period, which now appears to stretch from 2,000,000 years ago to around 3000 BC, archaeological evidence is the only source of knowledge about human activities. But prehistoric remains have always been the most difficult to interpret, precisely because there are no written records to aid in the task. Now, with exact dating techniques at his disposal, the prehistorian is becoming more like the historical archaeologist and is concerned with the periodization and the historical contexts of his finds.

BIBLIOGRAPHY. Good general introductions to the aims and methods of archaeology are LEONARD WOOLLEY, *Digging Up the Past* (1930); SIR MORTIMER WHEELER, *Archaeology from the Earth* (1954); and GRAHAME CLARK, *Archaeology and Society*, 3rd rev. ed. (1957). For the history of archaeology and its relation to the development of anthropology, see C.W. CERAM, *Götter, Gräber und Gelerte* (1949; Eng. trans., *Gods, Graves and Scholars*, 1951); G. BIBBY, *The Testimony of the Spade* (1956); and GLYN DANIEL, *A Hundred and Fifty Years of Archaeology* (1972).

Anthologies of archaeological writings that relate both to the history of the subject and its present methods are many. The following are recommended: R.F. HEIZER, *The Archaeologist at Work* (1959), and *Man's Discovery of his Past*, 2nd ed. (1970); and JACQUETTA HAWKES, *The World of the Past* (1963). For the development of American archaeology, see G. WILLEY and G. SABLOFF, *The History of American Archaeology* (1972).

Special aspects of the development of archaeology are dealt with in D. BROTHWELL and E. HIGGS, *Science in Archaeology*, 2nd ed. (1969); GEORGE F. BASS, *Archaeology Under Water* (1967); and W.F. LIBBY, *Radiocarbon Dating*, 2nd ed. (1955).

(G.E.D.)

Archery

Shooting arrows with a bow is one of the oldest skills practiced by man. From prehistoric times, and except in Australia, the bow in its many forms was for a time a principal weapon of war and of the hunt throughout the known world. The acquisition, improvement, and retention of proficiency in its use involve many elements that characterize a sport—regular practice to achieve accuracy or increase distance; the development of skill in aiming and loosing (releasing); engaging in competition and the satisfaction of improving performance under pressure.

Though the bow was displaced as a weapon of war in the 15th century, it was retained as a hunting weapon and as a versatile instrument for other kinds of agreeable pastime. Over the intervening years activities with the bow evolved into the modern sport of archery.

The number of archers around the world increased phenomenally after 1930, led by remarkable growth in the United States. By the 1970s the total number of participants in all forms of the sport was probably 10,000,000, of which 8,000,000 were in the United States. These totals encompass a relatively small number who strive for ever greater distances in "flight" shooting; a very much larger number interested in accuracy in shooting at targets of several kinds, at various distances; and a still greater contingent of those who have adopted the bow as their favourite hunting weapon for game.

HISTORY OF THE BOW AND ARROW

The bow was almost certainly the first mechanical device for achieving greater speed in a projectile than could be attained by throwing. It does this by accumulating energy in the bow limbs while drawing, storing it temporarily while holding and aiming, and releasing the

Principle of the bow

stored energy by converting it to energy of flight in the arrow. Its invention, 30 or more millennia ago, may rank in social impact with the invention of the wheel and the art of kindling fire. Very old, if not the oldest known, evidence that the bow was recognized as an effective hunting weapon thousands of years ago are the rock paintings in the Cueva de los Caballos near Albocácer, Castellón, in eastern Spain.

During its long history the bow has assumed many shapes, structures, and dimensions. The variations from their earliest prototype depended on the ingenuity of the artisan bowyer, and the identification of available materials—wood, horn, sinew, gut, and rawhide, for example—especially suited to his purpose of creating a device that would propel the arrow toward its mark at the will and under the control of the archer.

Types of bows: construction and design. Initially, and probably over many years, bows were made of a single material, usually wood (self bows), including those in which two pieces were fastened together to make the equivalent of a single long stave. Later, some bows were made of several materials, such as wood, horn, and sinew, glued together in layers (composite bows).

In Europe, up to about the 11th century, short self bows were used. A weak weapon, because a short self bow cannot withstand a long draw, it was probably not highly esteemed or long used. The longbow, originally also a self bow, appears to have originated in Wales. Its length was about the same as the height of the archer, and the arrow—the famous cloth-yard shaft—about half as long (see below *Equipment and use: Arrows*). The bow was held with outstretched arm and the arrow drawn to the bowman's ear. With the longbow the English yeoman could shoot six aimed shots per minute, with an effective range of about 200 yards. With greater skill in use and with strong bows even greater lengths were possible.

The crossbow

In continental Europe the crossbow became widely used and for centuries held its ascendancy over the longbow. Its use both in warfare and for hunting was especially widespread in Switzerland, parts of Germany, the Low Countries, and France. Usually the bow of the crossbow was short and heavy, made of steel, and, though quite inefficient, could propel its arrow (usually four-sided, hence called a quarrel) with relatively high velocity. Mechanical aids such as levers and winches were needed to draw the string back to the retaining "nut." This was a slow process that limited the weapon to a much lower "firepower" than the longbow. Among many variations in designs, a repeating crossbow was developed in China and an all-wood crossbow was used in West Africa.

In the decades since 1950, a modern version of the crossbow has come into favour as a sports implement. Its bow has benefitted from improvements in the longbow, made since the early 1930s, and mechanical aids for drawing are not required. It now has a place in the annual tournament of the National Archery Association of the United States, which determines championships.

Composite bows, quite different in details of construction from the longbow, are Asiatic in origin and use. The horn, wood, and sinew layers of which the limbs are made all have important functions: the wood strip is the base to which the horn is glued on the inner or compression side, and the layer of sinew fibres on the outer or tension side. The wood bears the principal shearing stress in the bent limb.

Japan's seven-foot bows

The typical Japanese bow, though seeming almost grotesque to those unfamiliar with it, performs surprisingly well in the hands of the skillful Japanese archer. Usually more than seven feet long, as compared with five to six feet for the longbow and less than four feet for the Turkish composite, its upper limb is almost twice the length of its lower. The limbs, with little or no taper, are comprised of thin strips of bamboo glued to both sides of a supporting strip of wood such as mulberry. As an art form, the Japanese also made short bows of other materials. The very long bow of Japan proved itself a potent weapon in battle and remains widely used in sport, although the contemporary American bow is also becoming widely used in Japan.

In other areas—Africa, South America, parts of Asia, Europe, and North America—the various forms into which the bow evolved were influenced, if not determined, by the materials available for their construction and by the nature of the terrain in which they were used. In wooded areas, the shorter bow is preferable to the longer; in open spaces, the longer may have been preferred because of some advantages in smoothness of draw and somewhat better accuracy. Pygmy tribes in their jungle habitat used short weak bows but made their arrows lethal with poison. Some South American Indian tribes who use the bow in fishing use very long bows and arrows. American plains Indians used long bows, made of *bois d'arc* (osage orange) where obtainable; those in timber country tended to keep theirs short. Some southwestern American Indian bows as far back as the 17th century were backed with sinew. In the northern frigid areas where native timber was lacking, available driftwood might be pressed into service by the bowyer, with sinew and baleen (whalebone) likely to be added for strength and sturdiness.

Bowstrings. Bowstrings also were made of available materials—animal and vegetable fibres, hemp, silk, linen, mohair, twisted gut, thin strips cut from hides, and even strips of bamboo and rattan. Various methods of attachment of the string to the bow were used. Stiff bamboo and rattan "strings" had to be secured to the bow tips with flexible, fibrous material that could be looped and tied. Strings of any flexible material might be threaded through a hole near the tip of the limb and firmly lashed, or provided with temporary or permanent loops seated in a groove or "nock" at the end of the limb.

Arrows. Arrows, like bows and strings, have emerged in slow evolution, and in recent years from technological developments, and may differ in shape, dimension, length, stiffness, and fittings, with specific heads intended for specific needs. Stabilizing vanes have been fashioned from tail and pinion feathers of such birds as eagle, crow, goose, and turkey. Pygmy tribes with their small arrows used leaves. Flight arrows such as used by the Turks were stabilized with small sections of pigeon feathers. Arrows of some Indian tribes, used for knocking down small game, were made bulbous and heavy at the forward end, and because this moved the centre of mass so far forward, it made the use of stabilizing vanes superfluous.

The earliest arrows used were probably made of some form of reed or cane, needing little work to make them serviceable. The development of certain tools, flaked stone chips, for example, brought wood from straight-grained trees such as spruce into the picture. These woods combine lightness with strength. Some arrows, instead of being fitted with heads of stone or metal for penetration, were sharpened and charred, and the point rubbed smooth. More commonly, a separate penetrating head of stone, bone, or shell was secured to the foreshaft with lashings or cement or both. Some heads were made detachable, to remain behind when the shaft was drawn from the wound. In South America some tribes favoured very long arrows, the fore-end of which was a short, separable shaft, socketed into the larger one of the arrow proper. When iron and steel became available, a variety of designs of head became possible, each design supposedly suited especially to some particular kind of game, or to warfare. The head of the crossbow quarrel was usually small and solid, of polygonal section, and sharply pointed (see below *Equipment and use*).

HISTORY OF ARCHERY AS A SPORT

The history of archery as a modern, competitive sport with national and international tournaments and champions is the history of the use of the longbow and of its immediate successor, the modern bow.

Role of the longbow

The idea of archery as an organized sport, its organization into amateur archery clubs with national associations as governing bodies to regulate competition, even the competitive events, consisting of shooting groups of arrows from several specified distances at standardized targets with a standardized scoring system, all evolved from practice with the longbow. It was not until the sport

was well established that the longbow itself was discarded (see below).

By the 13th century and extending through most of the 15th, a longbow made of yew was the favoured weapon of the English. During the Hundred Years' War the most notable victories of the English were those won at Crécy, Poitiers, and Agincourt over the French, victories made possible by the use of archers.

Although displaced from battle by firearms during the years following, the longbow held its place as an implement of sport and pastime. In Great Britain and North America, especially, the legend of Robin Hood, served to build and keep alive the reputation of that weapon.

Many British monarchs, up to and including Queen Victoria, practiced archery. Records indicate that from the 13th to the 17th centuries most of them were involved in the practice or promotion of skill with the bow. Many statutes during the reigns through that of Charles II were included in the code of law to assure that there be a plentiful supply of bows, arrows, and related equipment, sufficient to meet any situation, and that the able-bodied among both freemen and servants practice, on specified schedules, to become well trained in shooting. At various times bowls, golf, and other sports were banned lest they interfere with the practice of archery.

With the decline of the bow as an important factor in the defense of the country, there remained those who were fond of shooting. They called themselves toxophilites, those fond of the bow, from the Greek *toxon*, "bow," also the root of "toxic," by way of arrow poison. Among them, the one best known was Roger Ascham through his authorship of *Toxophilus, the Schole of Shooting* (1545), the first book on archery written in English. In the book, dedicated to Henry VIII, he endeavoured to show, regarding archery,

Roger Ascham and the toxophilites

> how honest a pastime for the minde; how holsome an exercise for the body; not vile for great men to use, not costly for poor men to sustayne, not lurking in holes and corners for ill men at their pleasure to misuse it, but abydinge in the open sighte and face of the worlde, for good men if it fault, by their wysdome to correct it.

Archery societies. *The Honourable Artillery Company and the Finsbury Archers.* Henry VIII in 1537 granted a patent to the Fraternity of St. George or Honourable Artillery Company, whose members used "long bowes, cross bowes, and handgunnes." Further patents were granted by James I in 1605 and Charles I in 1633, reinstating the Company's practice grounds, which had been partly fenced in by the owners. The Company established its Artillery Ground at the southern end of its practice grounds, Finsbury Fields, close by Finsbury Square. On the Fields, archers set up many marks for "roving." The roving archer shot from one mark to the next, the object being, as in golf, to use the fewest shots in going around the course. A plan of the Fields published in 1628 showed 164 marks. By 1737 the number had dwindled to 21, but several butts had been provided for target shooting. When the Honourable Artillery Company was discarding bows as war weapons about 1590, the Finsbury Archers, probably consisting of the remaining members of the archery section, began to flourish and held regular competitions until about 1761. The last of the Finsbury Archers joined the Toxophilite Society when it was founded in 1780.

The sport of roving

The Royal Tox and the Royal Company. When the prince of Wales, afterward George IV, became the Toxophilite Society's patron in 1787, it became the Royal Toxophilite Society. He was active in the organization and was responsible for setting "the prince's lengths" of 100 yards (91 metres), 80 yards (73 metres), and 60 yards (55 metres), distances still used in the championship York Round. He was also responsible for the method of scoring hits on the standard target, called the prince's reckoning.

The "Royal Tox," as it is known among archers, was not the first British archery society. It is antedated by the Royal Company of Archers, the queen's bodyguard in Scotland, whose headquarters are at Archer's Hall, Edinburgh. Informally, it had existed since about 1600.

The Grand National British championships. In 1844 the first of the Grand National Archery Meetings—the British championships—was held at York, and the Grand National Archery Society became the governing body of the sport in the United Kingdom.

Archery in North America. The history of archery in North America differed extensively from that in England. In North America, the use of bows and arrows in hunting and in combat was largely confined to the Indians. There is little evidence that the early colonists used weapons other than firearms, and after colonization little time passed until the Indians also had firearms.

The first formal organization in the U.S. devoted to archery was established in 1828 as the United Bowmen of Philadelphia. The club, which was patterned on the English and whose members used imported English bows and arrows, carried on until 1859, after which it remained inactive until its revival in 1932.

During the 1870s a number of archery clubs sprang up in several states. Members from eight of these clubs met in Crawfordsville, Indiana, in January 1879 for the purpose of organizing the National Archery Association of the United States. The first national target archery tournament was held by the association in White Stocking Park, Chicago, the following August. Except in the war years, a tournament has been held annually.

United States archery clubs

Amateur archery clubs in North America generally followed the English example in respect to equipment (use of the longbow), competitive events, and so on. A principal contribution from North America was the revival of hunting with the bow and arrow and the development of the sport of field archery. The popularity of hunting with the bow and arrow was stimulated by the publication of articles by Maurice Thompson, first president of the National Archery Association. Published as a book, *The Witchery of Archery* (1878; 2nd ed. 1879), his efforts kept archery alive until it could gather the momentum needed for substantial growth.

In 1939 a group of California archers organized the National Field Archery Association of the United States, to foster sportsmanship in hunting and promote interest in the "freehand" kind of shooting used in roving or field archery. Like the National Archery Association, it conducts an annual tournament.

International archery. International competition in target archery began with Anglo-French matches at Le Touquet before World War I, and in 1931 the Fédération Internationale de Tir à l'Arc (FITA; also known as the International Archery Federation) was formed to conduct the world championship matches, now held biennially. World championships are based on double FITA Rounds, adopted 1957, shot at metric distances and scored from one to ten. The longest distance for men is 90 metres (98.4 yards) and 70 metres (76.5 yards) for women. Because of growing international interest in the sport, field archery was included in the world championships held in 1969. Under international field archery rules the metric system is used, and the longest shot is from 60 metres (65.6 yards), and the largest of the four target sizes is 60 centimetres (23.6 inches).

Competitive target archery had become a popular amateur sport in Europe in the first third of the 20th century, and from 1931 through 1950 international competition was dominated by European archers, the men's especially by France, Belgium, Sweden, Poland, the United Kingdom, and Czechoslovakia, and the women's by Poland. Since the late 1950s U.S. archers have been predominant.

Archery is one of the sports recognized by the International Olympic Committee as eligible for inclusion in the Olympic Games. Archery events were included in the Games of 1900, 1904, 1908, and 1920. Archery, now under FITA rules and regulations, was again included as one of the official sports in the 1972 Olympic Games. For world records and championships see SPORTING RECORD in the *Ready Reference and Index*. See also *Olympic Games* tables in ATHLETIC GAMES AND CONTESTS.

Archery in the Olympic Games

Archer antiquarians. An important event in the furtherance and preservation of the history of archery was the founding of the Society of Archer Antiquaries in

1956 in London. The antiquity of archery implies, if it does not assure, a wealth of information to be found, pertaining to its use in hunting, in battle, and in sports. There were more than 400 members from many countries in the early 1970s.

Many books have been published about the various aspects of archery, principally the historical, technical, and instructional in combination. Collections of old books, monographs, reprints, and prints have been made by individuals.

EQUIPMENT AND USE

Traditional materials and construction. *The longbow.* The English longbow, famed in legend and history, was generally if not always made of wood from the English yew tree (*Taxus baccata*). Indeed, the word yew became synonymous with bow. It is probable that in the several centuries beginning with the 11th, bows in England were made of native-grown yew wood, which was uniquely suited for the longbow because of its combination of physical properties, a combination found in relatively few other kinds of wood. When the use of the bow became widespread, prior to the advent of firearms, the quantity of yew wood grown in England of suitable quality for bows was inadequate to meet the needs. It became necessary—and statutes were enacted to this end—that staves of yew be imported from Italy, Switzerland, Bavaria, and elsewhere to supply the demand. Such statutes provided that importation of other products, such as wine, from these countries was contingent on the simultaneous importation of specified quantities of yew staves.

During the early decades in which the longbow was used for sport in the U.S., the equipment for archery was imported from England. Eventually—*i.e.*, during the latter part of the 19th century—the sport began to assume independence with the establishment of shops by bowyers and arrow makers.

Abandonment of the longbow in the United States

Whereas in England the art of fashioning bows from yew and arrows from old deal became a craft that was passed along from father to son, the custom did not take root in the United States. That the United States was less bound by tradition explains in part the almost sudden abandonment, during the 1930s, of the longbow for other designs that had proved measurably superior to the traditional longbow. It was known in the United States that some tribes of Indians in the Southwest had good bows made from a native wood, osage orange, as mentioned above. Among the French pioneers it was called *bois d'arc*. The bowwood, bodock, or bowdark, as many of the non-French plains settlers called it, became widely favoured for hunting bows. Both the longbows and hunting bows were self bows, as a rule made entirely of one piece.

Composite bows. In the composite bow, used throughout much of Asia, the wood strip bears the principal shearing stress in the bent limb; the horn bears the compression, and the sinew layer the tension. The strength and resilience of the three components make possible much shorter limbs of reflex curve design than would be needed in a longbow of equal drawing force. Possibly the need for a weapon to be used on horseback was the mother of this ingenious invention. The nomadic hordes of Genghis Khan and his successors were thus armed—evidence that the short, stout composite bow had become a most effective weapon prior to the 12th century.

Turkish craftsmanship in bowmaking

The use of the composite bow became common in Turkey, possibly through the Seljuqs soon after the Mongol invasion. Its qualities led to its adoption in the hunt and other sports, continuing through the first half of the 19th century. Distance records made with the Turkish composite bow stood beyond challenge until the 1960s when modern materials that were more easily worked into superior bows made it possible for archers to match or surpass them (see also below). Turkish bows of the period between the 15th and 19th centuries were models of craftsmanship and art; usually a fine bow was signed and dated by its maker. Many bows are preserved in the Topkapı Museum, Istanbul. It is extraordinary that, in contrast with the large number of old Turkish bows still extant in good condition, not a single bow of yew in anything approaching good condition made in England in the 15th to 17th centuries has been preserved.

From the traditional to the modern. Considering that the English longbow remained unchanged in general design and construction for upward of six centuries prior to about 1935, it lends interest to this account to note the effect of the transition to modern equipment on accuracy. During the early 1930s the shortcomings of the longbow were recognized by a few archers with science and engineering backgrounds who undertook through research and development, as a hobby activity, to improve them to achieve greater efficiency and accuracy.

Shortcomings of the longbow. The wood of which the traditional longbow was made is not homogeneous. One result is that a wood self bow varies in performance with temperature and humidity and almost invariably, after use for a time, changes in strength by acquiring a permanent set. The resulting curve in the limbs is called by archers following the string. To minimize this effect, the longbow is kept braced (strung) only during shooting. Finally, the usual construction of the longbow renders it sensitive to slight changes in the manner in which it is held while it is being drawn and loosed. Usually there is some additional loss of available energy to propel the arrow, indefinite in amount, dependent on the time during which the bow is held at full draw before the release.

To make high scores with the longbow, therefore, the archer must know its peculiarities and allow for them. With wood-shafted arrows, he must also know the flight characteristics of each arrow that is shot and adjust his aim to reduce scatter on the target.

The modern laminated bow. Technological progress made during World War II produced new materials notable for their strength, lightness, and elastic properties. Among them are plastics, some of which are superior to the horn used for many centuries in composite bows, and fibres, particularly of glass, which displace the time-honoured sinew fibres. Since the 1960s the manufacturers of modern bows have exploited the properties of the new materials and have carried forward the research and development begun in the 1930s to produce superbly performing bows and arrows. These embody the engineering features of the improved bows, which resulted from the earlier research and development, as well as some of the desirable features of the Turkish composite bows. They are not subject to the uncertain properties of wood, and their performance is little affected by changes in temperature or humidity.

Use of new materials

The result is that the archer is not diverted from concentration on his shooting technique to make small allowances in aiming each arrow because of faults in both arrow and bow. Assuming archers of equal skill, those with the modern equipment can make scores averaging 20 to 30 percent higher than with the old. Whereas formerly it was notable when an archer shot a round of six bull's-eyes at 60 yards (55 metres), that accomplishment has now become almost commonplace. Indeed, an archer with good equipment is under no handicap in competition with a skillful pistol shot.

The longbow never produced notable results in flight shooting. Its limit was around 300 yards (274 metres). Here also, the adaptation of modern design to the flight bow has steadily increased the distance. In the 1960s the record was more than 850 yards (777 metres), which probably equals or exceeds the performance of the Turkish archers of the 15th century, whose record distances are not precisely known because neither the method they used in measuring distance nor the conversion of their distance units to yards is known. It may be taken for granted that the limit of distance with modern equipment has not been reached. In 1970, in national competition one contestant using a modern flight bow—a "foot" bow with which the archer lies supine, pushes the bow with his feet and draws with both hands—exceeded a mile (1.6 kilometres) by more than 100 yards.

Comparison of designs. The longbow, because of the faults inherent in a bow made solely of wood and of those owing to poor mechanical design, was inefficient in

comparison with its modern counterpart. By efficiency of a bow is meant the percent of the energy in a fully drawn bow that is transferred to the arrow at the instant of loose. One of the results of research and development has been to bring about a doubling of the efficiency, with some slight variation above and below this factor, depending on details of design, materials, and construction.

Doubling of a bow's efficiency

Doubling efficiency means, in practice, that a modern bow, shooting the same arrow as a longbow of the same force at equal draw, will impart to that arrow a velocity about 40 percent higher; *e.g.*, 212 feet (65 metres) per second as against 150 feet (46 metres) per second.

The illustration, top left, shows front and side views of a longbow relaxed, the same bow braced, and at full draw. For comparison, the modern bow is shown below. Differences to be noted are the thin (3/16 to 1/4 inch; .48 to .64 centimetre), wide (1¾ to 2 inches; 4.5 to 5.1 centimetres) limbs of equal length (about 18 inches) of the modern bow and the thick, tapering, narrow (about ¾ inch at grip to 1/4 inch at tip, 1 inch wide at grip) limbs of unequal length (34-inch upper and 32-inch lower) of the longbow; the shaped grip of the new bow, which assures the same location of support when drawn, and the lack of a definite grip in the longbow; the fixed arrow rest of the new bow, and the indefinite arrow rest, namely, the knuckle of the index finger of the bow hand, for the old; and the reverse curve toward the tips of the new bow.

The limbs of the modern bow are composite, laminated. A thin strip of wood serves as the foundation for facing and backing strips secured to it with strong, usually epoxy, glue. These strips are made of strong synthetic resin in which parallel glass fibres are imbedded, so that the composite limb is exceedingly strong in both compression and tension. The modern bow shows marked resemblance in the shape of the relaxed limb to that of the Turkish composite bow, with the difference of a long (approximately 27-inch; 69-centimetre) rigid middle section: this is not used either in the Turkish or the longbow. It provides mass and stability, which contribute to accuracy; and its length compensates for the lack of length of the short limbs, thus contributing to the desirable action that characterizes a long bow. The string, as light and strong as possible, is usually made of man-made fibre, such as dacron.

Addition of rigid middle section

From the instant of loose until the arrow leaves the string, the latter exerts thrust on the arrow, and an equal and opposite thrust against the bow hand. During the few hundredths of a second from the instant of release until the arrow departs from the string, the force exerted on the bow includes one that tends to rotate the bow about an axis parallel to the string, passing approximately through the point where the bow rests in the hand. If rotation in the hand occurs, it can produce a small lateral error. To prevent this, bows may be fitted with one or two "stabilizers," consisting of metal rods weighted near their ends, or provided with adjustable masses, secured to the back of the bow near its middle, and extending outwardly, away from the archer. The stabilizers are intended to increase the moment of inertia about the aforementioned axis of rotation, thereby lessening the effect of the rotational impulse, and thus reducing the lateral dispersion in hits on the target. The drawing force of target bows may be between 30 and 50 pounds (14 and 23 kilograms) for men and between 20 and 40 pounds (9 and 18 kilograms) for women and juvenile archers. Bows for flight shooting may have a drawing force between 50 and 90 pounds (23 and 41 kilograms). A bracer or armguard is often needed to prevent a painful slap of the string on the forearm.

Arrows. The cloth-yard shaft of the longbow was made of well-seasoned planks and beams, usually of Norway pine, salvaged by arrow makers from old buildings. Except in North America where it had been gradually displaced by Port Orford cedar beginning about 1930, seasoned Norway pine continued to be the material preferred for target arrows until the manufacture of metal shafts was perfected after World War II. Hunting shafts often were made of yellow birch.

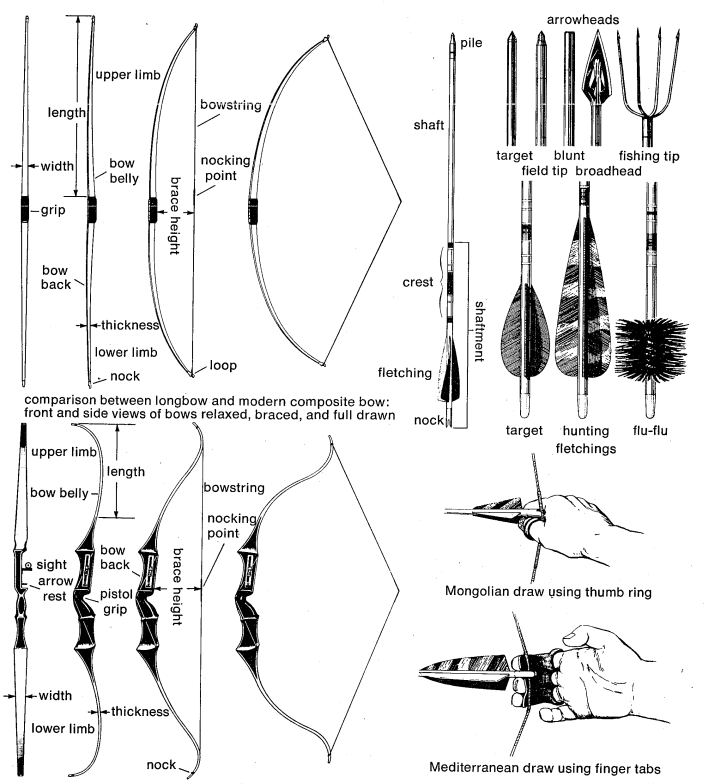
To assure consistent performance of a bow-and-arrow combination, the materials used for the arrows are as important as are those for the bows. Maximum strength, stiffness, and resilience consistent with minimum weight are desirable. The archer's ideal is that every arrow in a set of six or preferably in the dozen lot usually sold have identical flight characteristics. To meet this expectation, the mechanical and elastic properties of the shafts, including straightness and stiffness, must be uniform; the shape and dimensions must be alike in all of them; and the fletching—*i.e.*, the stabilizing vanes, traditionally feathers, and more recently plastic fins—must be the same and attached in the same location and configuration on all the shafts.

Desired properties of arrows

Wood for arrows, insofar as it is still used, must accordingly be straight-grained and meticulously selected to meet the stated requirements. For precision shooting at targets this, though necessary, is not sufficient because of the unavoidable variability in wood. After a quantity of arrows has been finished, each must be subjected to critical inspection to assure equality of weight and dimensions, balance point, straightness, stiffness, and fletching. Out of a large quantity, only a few dozens may pass such inspection for prime quality. Although wood is still favoured by some archers, aluminum-alloy tubing has largely supplanted wood as the most favoured material for arrows.

Depending on the archer's stature, his arrows may be between 25 and 30 inches (64 and 76 centimetres) long. Most modern target arrows are made of strong, metallurgically treated aluminum-alloy tubing, obtainable in diameter and wall thickness with close tolerances. From such tubing arrows can be made so closely alike, especially when the stabilizing vanes are made of plastic fins cemented on in a jig, that they have identical flight characteristics. For precision target shooting, feathers are inferior for fletching because they are easily damaged or put in disarray, causing changes in flight patterns. The parts of an arrow are indicated on the illustration of an arrow for target shooting. Types of arrow points illustrated include the broadhead for hunting. It consists of a hardened steel blade, with a razor edge, and, in one form, a small, auxiliary cutting blade at right angles to the main blade to increase penetration; both are secured to a suitable ferrule for attachment to the shaft.

Target and flight arrows



Types of archery equipment and drawing technique.

Flight arrows are patterned after those used by the Turks, strongly barrelled in the middle with taper to a small point and somewhat less taper toward the rear (the nock that engages the string). They are made of wood with high strength and low density consistent with strength. Effort is made to minimize mass (weight) yet keep the shaft strong enough to withstand the high thrust of the string. The optimum mass of a flight arrow of given dimensions and form, and shot from a given bow, represents a compromise between small mass, to which the bow imparts high velocity, and large mass, which comes off the bow with lower velocity but higher energy. The energy in the arrow is lost in air drag, and since this rate of loss is proportional to the square of the velocity, the lighter arrow loses its supply of energy faster than the heavier and slows down more quickly. So, if one had a series of arrows, all alike except for mass, ranging in mass from light to heavy, and were to shoot each one for maximum distance with the same bow, one of the arrows between the lightest and the heaviest would attain maximum distance. The lightest would not travel farthest because of rapid loss of energy in flight; the heaviest would not travel farthest because the bow could not give it sufficient velocity to achieve distance. To reduce drag, plastic vanes are usually used, and the surface of the arrow is given the smoothest possible finish.

Drawing and loosing. In all the Western nations, the so-called Mediterranean draw is used to draw and loose the arrow. This draw is executed by drawing the string with three fingers, the first being above and the second and third below the nocked arrow. Assuming right-handed shooting, the arrow rests on the knuckle of the index finger of the left hand holding the bow and against the bow. In modern bows an arrow rest is used. The arrow is shot from the left side of the bow.

Mediterranean and Oriental draws

The other basic draw is the Oriental, or Mongolian, in which the thumb is used, protected by a thumb ring of horn, ivory, bone, or perhaps plastic, to draw the string (see illustration). The string rests on the surface of the thumb ring, and the end of the thumb is supported and "locked" in place by the index finger. The arrow is shot from the right side of the bow. Loosing is by releasing the thumb from its locked position. This type of draw was used by Turkish archers and by the Mongols of Genghis Khan and is still widely used in the Far East. Japanese archers use it with a shooting glove, the thumb of which is a modified thumb ring.

A device designed to aid in drawing and loosing, in several variations of design, consists of a piece of strong, light material, shaped to fit the drawing hand, and provided with a small "hook" or ledge on which the string, with arrow nocked, rests. While drawing, the string is held in place by one of the fingers or the thumb of the drawing hand. Releasing results when the restraint is removed to allow the string to slip off the ledge. Such devices have been used in flight shooting since the 1940s, resulting in notably greater distances of flight, and in target shooting since the 1960s, permitting marked improvements in accuracy.

Ballistics. The flight of an arrow is affected by those of its characteristics that produce drag or air resistance, such as roughness of surface, size and smoothness of stabilizing vanes and their method of attachment to the shaft, shape and dimensions of the shaft, and location of balance point. Its mass and initial velocity are also factors, since these, with a given bow, determine the angle at which it must be elevated to travel its intended trajectory. Since the initial velocity is produced by the thrust of the bowstring, the strength of the bow and the manner in which it transfers its energy to the arrow are also important factors.

Interaction of the bow and arrow

In considering the manner in which the bow and arrow interact during the small fraction of a second between the loose and the departure of the arrow from the string, it is simplest to assume the Mediterranean draw with the right hand drawing the bow: at the instant of loose, as the string slides off the three fingers, the nock end of the arrow is given an impulse to the left in addition to the forward thrust. This sets up an oscillation in the arrow around two nodal points, each about one-fifth to one-fourth the length of the arrow from the ends of the shaft. If the time it takes the arrow to pass the bow is the same as the time required for one complete cycle of its oscillation, the arrow passes the bow smoothly; at every instant of its passage the shaft is displaced away from the bow so that the arrow does not touch the bow. Such an arrow is said to be matched to the bow. If the oscillation is faster or slower per cycle than the time it takes the arrow to pass the bow, some part of the arrow strikes the bow in passing, and the arrow is deflected. In this case the arrow is said to have too much or too little spine. The term spine combines ideas of stiffness and mass, since both affect the frequency of its oscillation; its mass also affects the time required to pass the bow.

The size, number, surface texture, and attachment of the vanes affect drag. Their location with respect to the balance point of the arrow determines the amount by which the angle (yaw) of the arrow, in relation to its flight path, deviates from the flight path when it flies in a side or quartering wind. The larger the vanes and the farther forward the balance point, the greater the yaw.

BIBLIOGRAPHY

General: NATIONAL ARCHERY ASSOCIATION OF THE UNITED STATES OF AMERICA (NAA), *The Archer's Handbook* (1966), information on organizing a club, setting up a range, and running a tournament, with simple instructions in shooting, and tournament rules for the NAA and FITA (the International Archery Federation); NATIONAL FIELD ARCHERY ASSOCIATION OF THE UNITED STATES, *Official Handbook of Field Archery*, 12th ed. (1964), information on field archery, its practice and tournaments, and bow hunting; C.J. LONGMAN and H. WALROND, *Archery* (1894, reprinted 1967), target archery as practiced in England and America in the 1890s, with historic background; GORDON GRIMLEY, *Book of the Bow* (1958), an excellent and interesting general treatise on archery; S.T. POPE, *Hunting with the Bow and Arrow* (1923), an interestingly written account of the hunting exploits of Saxton Pope and Arthur Young, which strongly influenced the revival of interest in the use of the bow. See also THEODORA KROEBER, *Ishi in Two Worlds: A Biography of the Last Wild Indian in North America* (1961); H.A. FORD, *Archery: Its Theory and Practice* (1880); and R.P. ELMER, *Archery*, rev. ed. (1933).

Technical aspects: P.E. KLOPSTEG, *Science Looks at Archery* (1934), *Turkish Archery and the Composite Bow*, 2nd ed. (1947), and with F. NAGLER and C.N. HICKMAN, *Archery: The Technical Side* (1947).

Periodicals: *Archery* (monthly), official publication of the National Field Archery Association; *Archery World* (monthly), official publication of the National Archery Association; *The British Archer* (bimonthly); *Society of Archer-Antiquaries* (annual).

(P.E.K.)

Archimedes

A Greek mathematician, scientist, and inventor, Archimedes was the founder of theoretical mechanics. He also pioneered in deducing mathematical laws from experimental observations.

Archimedes was born *c.* 287 BC in the Greek city-state of Syracuse, in Sicily. He was the son of Phidias, an astronomer, and was closely associated with, if not related to, Hieron II, king of Syracuse, and Gelon, his son. He studied at Alexandria in Egypt, which was at that time the centre of learning, after which he returned to his native city. There, he devoted himself to mathematical research.

Archimedes himself set no value on the ingenious mechanical contrivances that made him famous, regarding them as beneath the dignity of pure science and even declining to leave any written record of them except in the case of the sphere he constructed to imitate the motions of heavenly bodies. As these machines impressed the popular imagination, however, they naturally figure largely in the traditions about him. Thus he devised for Hieron engines of war that for three years held the Romans at bay in their siege of Syracuse. There is a story that he constructed an apparatus of convex glass mirrors that reflected the heat of the sun and set the Roman ships on fire when they were within a bowshot of the city's wall.

Archimedes' principle

It is probable that Archimedes had constructed some such burning instrument, though the connection of it with the destruction of the Roman fleet is more than doubtful. More important is the story of Hieron's question to Archimedes of whether a crown made for him and purporting to be of gold did not actually contain a proportion of silver. According to one version, Archimedes was puzzled until one day, as he was stepping into a bath and observed the water running over, it occurred to him that the excess of bulk occasioned by the introduction of alloy could be measured by putting the crown and equal weights of gold and of silver separately into a vessel of water and noting the differences of overflow. If the crown were of pure gold, it would displace an amount of water equal to that displaced by the piece of gold, if alloyed with silver, which weighs less per volume than gold, it would displace a greater amount. He was so overjoyed when this thought struck him that he ran home without his clothes, shouting "eureka, eureka" ("I have found it, I have found it"). From this incident he is also said to have derived the principle of buoyancy, which is called Archimedes' principle; that is, that a body immersed in a fluid loses as much in weight as the weight of the fluid it displaces. His pioneer work in mechanics is illustrated by the story of his having said "give me a place to stand on and I will move the earth." Hieron asked him to give an illustration of his contention that a very great weight could be moved by a very small force. For this purpose he is said to have used a large and fully laden ship and to have employed a mechanical device by which Hieron was enabled to move it by himself; but accounts differ as to the particular mechanical powers put to use. The water screw that he invented is used even now in Egypt for the purpose of irrigating fields. It is a corkscrew-shaped tube inserted in a watertight cylinder. When used for irrigation, one end of the cylinder is under water, the other end rises to the bank where water is discharged as the screw is rotated.

Archimedes died during the capture of Syracuse by the Roman general Marcellus, 212 BC. In the general massacre that followed the fall of the city, Archimedes, while engaged in drawing a mathematical figure on the sand, was stabbed to death by a Roman soldier. No blame attaches to Marcellus, since he had given orders to his men to spare the house and person of the sage, and in the midst of his triumph he lamented the death of so illustrious a person, directed an honourable burial to be given him, and befriended his surviving relatives. In accordance with the expressed desire of the philosopher, his tomb was marked by a sphere inscribed in a cylinder, the discovery of the relation between the surface and volume of a sphere and its circumscribing cylinder being regarded by him as his most valuable achievement.

The range and importance of the scientific labours of Archimedes will be best understood from a brief account of those writings that have survived, and it need only be added that his greatest work was in geometry. The method of exhaustion as originated by Eudoxus of Cnidus (*c.* 390–*c.* 340 BC) and as followed by Euclid is a technique of approximating the measure of an irregular geometric figure by fitting to it measurable regular figures, filling in the spaces with smaller figures, and repeatedly filling in until the irregular figure is approximated to the degree desired. Archimedes so extended the method of exhaustion that in several cases it was actually equivalent to integration. This is demonstrated in his solution for the area of the segments of a parabola (an open geometric curve) and of a spiral, for the surface and volume of a sphere and its segments, and for the volume of any segments of ellipsoids, paraboloids, and hyperboloids (curved geometric solids).

Extant treatises

The extant treatises are as follows:

1. *On the Sphere and Cylinder* is a treatise in two books dealing with the dimensions of spheres, cones, "solid rhombi," and cylinders, all demonstrated in a strictly geometrical method.

2. *Measurement of the Circle* is a short book whose chief proposition shows that the circumference of a circle is less than $3\frac{1}{7}$ and greater than $3\frac{10}{71}$ times its diameter. He thus put the exact value of the ratio of circumference to diameter (known as the Greek pi, π) within narrow limits.

3. *On Conoids and Spheroids* is a treatise on the solids generated by the revolution of the conic sections (circle, ellipse, parabola, and hyperbola) about their axes.

4. *On Spirals* contains propositions dealing with the tangential properties of the curve now known as the spiral of Archimedes and other propositions showing how to express the area included between any portion of the curve and the straight-line distances to its extremities.

5. *On the Equilibrium of Planes or Centres of Gravity of Planes* consists of two books and may be called the foundation of theoretical mechanics, for the previous contributions of Aristotle were comparatively vague and unscientific. Both books demonstrate ways of determining the centres of gravity of various geometric figures.

6. *Quadrature of the Parabola* is a book in 24 propositions, containing two demonstrations that the area of any segment of a parabola is $\frac{4}{3}$ of the triangle that has the same base as the segment and equal height.

7. *On Floating Bodies* is a treatise in two books, the first of which establishes the general principle of hydrostatics; the second discusses with the greatest completeness the positions of rest and stability of a right segment of a paraboloid of revolution (a bullet-shaped solid) floating in a fluid.

8. *The Sand-Reckoner* is a small treatise that deals with the problem of calculating the number of grains of sand that could be contained in a sphere of the size of the "universe" and expounds a new system of expressing large numbers.

9. *The Method* is a treatise of vital interest, since in it Archimedes explains how he first arrived at many of his important results by means of mechanical considerations; namely, by weighing an indefinite number of elements of one geometric figure against similar elements of another.

10. A collection of *Lemmas*, consisting of 15 propositions in plane geometry, has come down through a Latin version of an Arabic manuscript; it cannot, however, have been written by Archimedes in its present form since his name is quoted in it more than once.

Lost works

Of lost works by Archimedes there are references to seven: (1) investigations on polyhedra mentioned by Pappus; (2) *Principles*, a book addressed to Zeuxippus and dealing with the naming of numbers on the system explained in the *Sand-Reckoner*; (3) *On Balances or Levers;* (4) *On Centres of Gravity;* (5) *Catoptrica*, an optical work from which Theon of Alexandria quotes a remark about refraction (bending of light); (6) *On the Calendar*; (7) *On Sphere Making*, in which Archimedes explained the construction of the sphere that he made to imitate the motions of the sun, the moon, and the five planets in the heavens. Cicero actually saw this contrivance and described it.

BIBLIOGRAPHY. Translations of the works of Archimedes include T.L. HEATH (ed.), *The Works* (1897, suppl. 1912, reprinted 1953); and CHARLES MUGLER (ed.), *Archimède*, vol. 1, *De la sphère et du cylindre. La mesure du cercle. Sur les conoïdes et les sphéroïdes* (1970). For discussions of Archimedes' contributions to mathematics and physics, see MARSHALL CLAGETT, "Archimedes," in the *Dictionary of Scientific Biography*, vol. 1 (1970), and *Archimedes in the Middle Ages*, vol. 1, *The Arabo-Latin Tradition* (1964); For a biographical evaluation see A.G. DRACHMANN, "Archimedes and the Science of Physics," *Centaurus*, 12:1–11 (1968); T.L. HEATH, *Archimedes* (1920); and E.J. DIJKSTERHUIS, *Archimedes* (1956).

(B.L.v.d.W./T.L.H.)

Architecture, Art of

Architecture is the art and the technique of building, employed to fulfill the practical and expressive requirements of civilized people. Almost every settled society that possesses the techniques for building produces architecture. It is necessary in all but the simplest cultures; without it, man is confined to a primitive struggle with the elements; with it, he has not only a defense against the natural environment but also the benefits of a human environment.

a prerequisite for and a symbol of the development of civilized institutions.

The characteristics that distinguish a work of architecture from other man-made structures are (1) its suitability to use by human beings in general and its adaptability to particular human activities; (2) the stability and permanence of its construction; and (3) its communication of experience and ideas through form.

All these conditions must be met in architecture. The second is a constant, while the first and third vary in relative importance according to the social function of buildings. If the function is chiefly utilitarian, as in a factory, communication is of less importance. If the function is chiefly expressive, as in a monumental tomb, utility is a minor concern. In some buildings, such as churches and city halls, utility and communication may be of equal importance.

This article is divided into the following sections:

I. Use
- Architectural types
  - Domestic architecture
  - Religious architecture
  - Governmental architecture
  - Recreational architecture
  - Architecture of welfare and education
  - Commercial and industrial architecture
- Architectural planning
  - Planning the environment
  - Planning for use
  - Economic planning

II. Techniques
- Materials
  - Stone
  - Brick
  - Wood
  - Iron and steel
  - Concrete
- Methods
  - Wall
  - Post-and-lintel
  - Arch
  - Vault
  - Dome
  - Truss
  - Framed structures

III. Expression
- Content
  - Symbols of function
  - Expression of technique
- Form
  - Space and mass
  - Composition
  - Scale
  - Light
  - Texture
  - Colour
  - Environment
- Ornament
  - Mimetic ornament
  - Applied ornament
  - Organic ornament

IV. Theory of architecture
- Distinction between the history and theory of architecture
- Distinction between the theory of architecture and the theory of art
- Functionalism
- The art of building
- "Commodity, firmness, and delight": the ultimate synthesis
- Venustas
- Utilitas
- Firmitas

I. Use

The types of architecture are established not by architects but by society, according to the needs of its different institutions. Society sets the goals and assigns to the architect the job of finding the means of achieving them. This section of the article is concerned with architectural typology, with the role of society in determining the kinds of architecture, and with planning—the role of the architect in adapting designs to particular uses and to the general physical needs of human beings.

ARCHITECTURAL TYPES

Architecture is created only to fulfill the specifications of an individual or group. Economic law prevents architects from emulating their fellow artists in producing works for which the demand is nonexistent or only potential. So the types of architecture depend upon social formations and may be classified according to the role of the patron in the community. The types that will be discussed here—domestic, religious, governmental, recreational, welfare and educational, and commercial and industrial—represent the simplest classification; a scientific typology of architecture would require a more detailed analysis.

Domestic architecture. Domestic architecture is produced for the social unit: the individual, family, or clan and their dependents, human and animal. It provides shelter and security for the basic physical functions of life and at times also for commercial, industrial, or agricultural activities that involve the family unit rather than the community. The basic requirements of domestic architecture are simple: a place to sleep, prepare food, eat, and perhaps work; a place that has some light and is protected from the weather. A single room with sturdy walls and roof, a door, a window, and a hearth are the necessities; all else is luxury (see FARM BUILDINGS).

The essential components of a domestic building

"Vernacular" architecture. In much of the world today, even where institutions have been in a continuous process of change, dwelling types of ancient or prehistoric origin are in use. In the industrialized United States, for instance, barns are being built according to a design employed in Europe in the 1st millennium BC. The forces that produce a dynamic evolution of architectural style in communal building are usually inactive in the home and farm. The life of the average person may be unaltered by the most fundamental changes in his institutions. He can be successively a slave, the subject of a monarchy, and a voting citizen, without having the means or the desire to change his customs, techniques, or surroundings. Economic pressure is the major source of his conservatism; lacking the resources of the group, the average individual is forced to restrict his demands to a level far below that which the technology of his time is capable of maintaining. Frequently he builds new structures with old techniques because experiment and innovation are more costly than repetition. But in wealthy cultures economy permits and customs encourage architecture to provide conveniences such as sanitation, lighting, and heating, as well as separate areas for distinct functions, and these may come to be regarded as necessities. The same causes tend to replace the conservatism of the home with the aspirations of institutional architecture and to emphasize the expressive as well as the utilitarian function.

"Power" architecture. As wealth and expressive functions increase, a special type of domestic building can be distinguished that may be called power architecture. In almost every civilization the pattern of society gives to a few of its members the power to utilize the resources of the community in the construction of their homes, palaces, villas, gardens, and places of recreation. These few, whose advantages usually arise from economic, religious, or class distinctions, are able to enjoy an infinite variety of domestic activities connected with the mores of their position. These can include even communal functions: the palace of the Flavian emperors in ancient Rome incorporated the activities of the state and the judicial system; the palace of Versailles, a whole city in itself, provided the necessities and luxuries of life for several thousand persons of all classes and was the centre of government for the empire of Louis XIV. Power architecture may have a complex expressive function, too, since the symbolizing of power by elegance or display is a responsibility or a necessity (and often a fault) of the powerful. Since this function usually is sought not so much to delight the patron as to demonstrate his social position to others, power architecture becomes communal as well as domestic. In democracies such as ancient Greece and in the 20th-century Western world, this show of power may be more reserved, but it is still distinguishable.

Group housing. A third type of domestic architecture accommodates the group rather than the unit and is



Palace of Versailles, France, built chiefly by Louis Le Vau and Jules Hardouin-Mansart during the last half of the 17th century.
Aerofilms, Ltd., London

Cultures in which group housing has been produced

therefore public as well as private. It is familiar through the widespread development of mass housing in the modern world, in which individuals or families find living space either in multiple dwellings or in single units produced in quantity. Group housing is produced by many kinds of cultures: by communal states to equalize living standards, by tyrants to assure a docile labour force, and by feudal or caste systems to bring together members of a class. The apartment house was developed independently by the imperial Romans of antiquity to suit urban conditions and by the American Indians to suit agricultural conditions. Group architecture may be power architecture as well, particularly when land values are too high to permit even the wealthy to build privately, as in the 17th-century Place des Vosges in Paris, where aristocratic mansions were designed uniformly around a square, or in the 18th-century flats in English towns and spas. Although most domestic architecture of the 20th century employs the style and techniques of the past, the exceptions are more numerous and more important for the development of architecture than ever before. This is because the distribution of wealth and power is widespread in parts of the world where architecture is vital and because the modern state has assumed responsibility for a large amount of high-quality housing at low cost.

Religious architecture. The history of architecture is concerned more with religious buildings than with any other type, because in most past cultures the universal and exalted appeal of religion made the church or temple the most expressive, the most permanent, and the most influential building in any community.

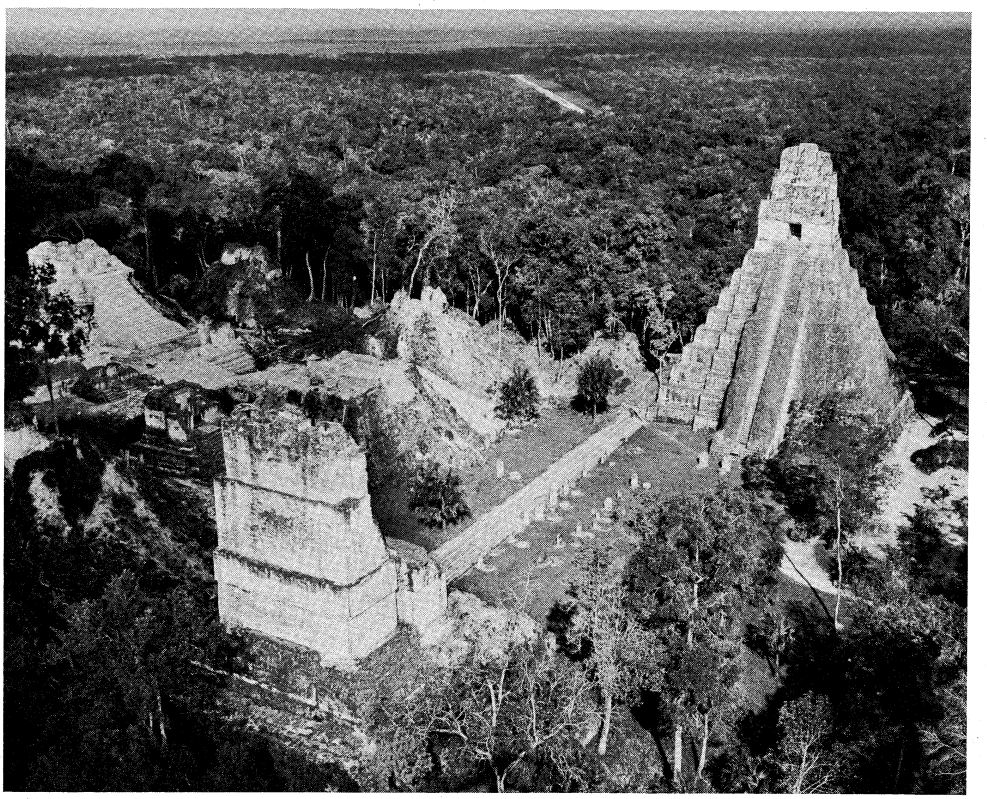
The typology of religious architecture is complex, because no basic requirements such as those that characterize domestic architecture are common to all religions and because the functions of any one religion involve many different kinds of activity, all of which change with the evolution of cultural patterns.

The temple or church. Temples or churches serve as places of worship and as shelters for the images, relics, and holy areas of the cult. In the older religions, the temple was not always designed for communal use. In ancient Egypt and India it was considered the residence of the deity, and entrance into the sanctum was prohibited or reserved for priests; in ancient Greece it contained an accessible cult image, but services were held outside the main facade; and in the ancient Near East and in the Mayan and Aztec architecture of ancient Mexico, where the temple was erected at the summit of pyramidal mounds, only privileged members of the community were allowed to approach. Few existing religions are so exclusive. Beliefs as dissimilar as Christianity, Buddhism, Judaism, and Islām are based on communal participation in rites held inside the temple or church. The buildings have even evolved into similar plans, because of a common requirement that the maximum number of worshippers be able to face the focal point of the service (the mosque's "point" is the wall facing the direction of Mecca, the city of Muḥammad's birth and therefore the most sacred of all Islāmic religious sites). Consequently, the Muslims were able to adopt the Byzantine church tradition, modern synagogues are often scarcely distinguishable from churches, and early Protestantism absorbed Catholic architecture with only minor revision (elimination of subsidiary chapels and altars, repositories of relics, and some symbolic decoration).

Design related to the demands of the ritual

Shelter is not always required for worship. Primitive rites are often practiced outdoors with some monument as a focus, while the altar of Pergamum and the Ara Pacis Augustae (Augustan Altar of Peace) in Rome are evidences of the open-air religious observances of the classical world. The atrium of early Christian architecture and the cloister were isolated areas for prayer.

The complex programs of later religions made the place of worship the focus for varied activities demanding architectural solutions; for example, the baptistery, bell towers, and chapter houses of Christian architecture, the



Mayan ruins at Tikal, with Temple of the Giant Jaguar (right); central Petén department, Guatemala, *c.* AD 300–900.
George Holton—Photo Researchers, Inc.

minarets of the Muslims, and the holy gates of Buddhism. Most modern sects demand space for religious education adjoining the community church or temple. Catholicism and the religions of Asia have produced monasteries, convents, and abbeys—connected to places of worship—that accommodate the organized practice of religion, adding domestic and often industrial, agricultural, and scholarly functions to the religious.

Shrines and memoria. Shrines consecrate a holy place for its miraculous character or for its association with the life of the founder, gods, or saints of a cult. Since the importance of such structures is usually proportionate to the antiquity of their tradition and associations with cult origins, they have had little importance in later architectural history. The major commemorative buildings of Christianity are those connected with the life of Jesus Christ (Church of the Nativity in Bethlehem) and the apostles or early Church Fathers (St. Peter's in Rome) or with the medieval cult of relics (Santiago de Compostela in Spain). No single formal design characterizes this type, but the theme of the domed or central-plan structure (round, square, polygon, Greek cross, etc.) connects the *memoria* of Asia (the Indian stupa, Chinese pagoda), pagan antiquity (the Pantheon in Rome), and Christianity (the Church of the Holy Sepulchre in Jerusalem). The significance of the form is discussed below under *Content.*

Funerary art. Expressing man's relationship to the afterlife, funerary art is not always architectural, since it may be purely symbolic and therefore suitable to sculptural treatment, as in the classic Greek, medieval, and modern tomb. Funerary architecture is produced by societies whose belief in the afterlife is materialistic and by individuals who want to perpetuate and symbolize their temporal importance. Monumental tombs have been produced in ancient Egypt (pyramids), Hellenistic Greece (tomb of Mausolus at Halicarnassus, which is the source of the word mausoleum), ancient Rome (tomb of Hadrian), Renaissance Europe (Michelangelo's Medici Chapel, Florence), and Asia (Tāj Mahal, Āgra, Uttar Pradesh, India). Modern tomb design has lost vitality, though it remains as elaborate (Monument to Victor Emmanuel II, Rome) or as meaningful in terms of power (Lenin Mausoleum, Moscow) as before. The exceptional examples are partly sculptural in character (*e.g.*, Louis Sullivan's Wainwright Tomb, St. Louis, Missouri; Walter Gropius' war memorial, Weimar, Germany).

Since the 18th century much of religious architecture has lost individuality and importance through the weakening of liturgical traditions. But today, as in the past, outstanding architects have met new demands of use and expression with superior solutions.

Governmental architecture. The basic functions of government, to an even greater extent than those of religion, are similar in all societies: administration, legislation, and the dispensing of justice. But the architectural needs differ according to the nature of the relationship between the governing and the governed. Where governmental functions are centralized in the hands of a single individual, they are simple and may be exercised in the ruler's residence; where the functions are shared by many and established as specialized activities, they become complex and demand distinct structures. There are, however, no basic formal solutions for governmental architecture, since the practical needs of government may be met in any sheltered area that has convenient space for deliberation and administration. A distinct type is created rather by expressive functions arising from the ideology of the different systems of political organization (monarchy, theocracy, democracy, etc.) and from the traditions of the various offices of government (law courts, assembly houses, city halls, etc.). Governments that exercise power by force rather than by consent tend to employ the expressive functions of architecture to emphasize their power; they tend to produce buildings of a monumentality disproportionate to their service to the community. Those in which the ruler is given divine attributes bring religious symbolism into architecture. Democratic governments have the responsibility of expressing in their architecture the aims of the community itself, a difficult task in the modern world, when the community may be neither small enough to express itself easily nor homogeneous enough to agree on how to do so.

Relation of political ideology to expressive function

The simple democratic processes of the Greek city-

states and the medieval free towns produced governmental architecture on a domestic scale, while the Roman Empire and later monarchies seldom made important distinctions between the palace and the seat of state functions. The widespread growth of representative government and the increase in the size and functions of the state in the 19th century created a great variety of buildings, some for entirely new uses. Some examples are: first, capitols, courthouses, parliament buildings, printing offices, and mints and, later, post offices, embassies, archives, secretariats, and even laboratories, when the work, the increased personnel, and the complexity of mechanical aids demanded specialized architectural solutions. Bureaucracy, for better or for worse, has made governmental architecture more important than at any time in history.

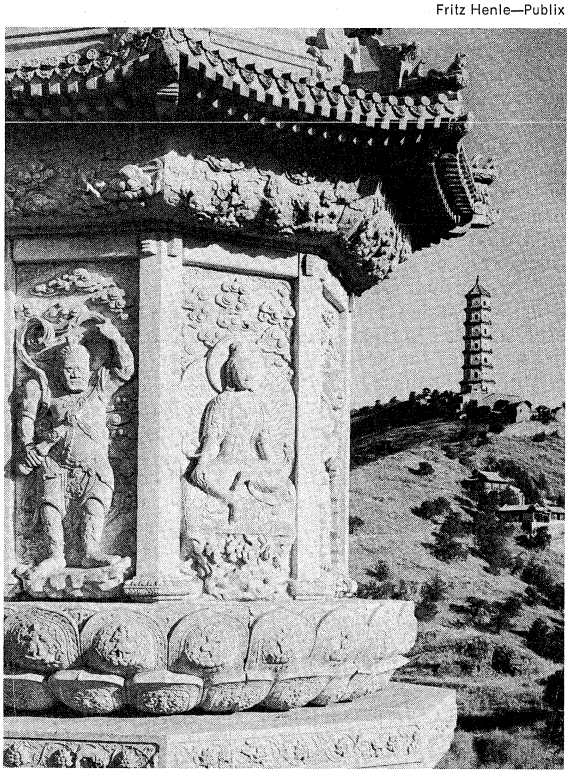
In the first rapid expansion from about 1780 to 1840, Neoclassical architects found impressive solutions to the new problems, but afterward governmental architecture lapsed into a century of conservatism, following at a safe distance behind private building. After World War II, governmental architecture showed new vitality. Outstanding are Le Corbusier's work at Chandīgarh, Punjab, India, the United Nations Educational, Scientific and Cultural Organization headquarters in Paris, and the program of the U.S. Department of State for building American embassies.

Military architecture is closer to the governmental type than to others, but its expressive function is so much subordinated to the practical that it is usually regarded as a class of engineering (see ENGINEERING, MILITARY; FORTIFICATIONS).

Recreational architecture. Few recreations require architecture until they become institutionalized and must provide for both active and passive participation (athletic events, dramatic, musical performances, etc.) or for communal participation in essentially private luxuries (baths, museums, libraries). Throughout history, recreational architecture has been the most consistent in form of any type. Diversions may change, but, as in domestic architecture, the physical makeup of the human being provides consistency. If his participation is passive he must be able to hear and to see in comfort. If his participation is active, he must be given spaces suited to the chosen activity. In most cultures, recreational institutions have their origins in religious rites, but they easily gain independence, and religious expression is reduced or eliminated in their architecture.

Historical consistency of building types

Fritz Henle—Publix



Base of the Marble Pagoda with the Jade Pagoda in the background, Jade Fountain Hill, west of Peking, 18th century (Ch'ing dynasty).

Theatres. Theatres originated in ancient Greece with the rites of the god Dionysus, first as temporary installations and later as outdoor architecture using the natural slope and curves of hillsides to bring the spectator close to the stage and to avoid the need for substructures. The Greek theatre was monumentalized and modified by the Romans, whose arches and vaults allowed construction of sloping seats from level foundations. In the Middle Ages churches and temporary structures were used for dramatic purposes, and in the Renaissance the form of the Roman theatre was occasionally revived (Andrea Palladio's Teatro Olimpico in Vicenza, Italy). The 17th-century development of opera, drama, and ballet in Europe brought about a revival of theatre building but in a new form conceived to satisfy class and economic distinctions (*e.g.*, the Teatro Farnese in Parma, Italy; Residenztheater, in Munich). A flat or inclined pit accommodated standing patrons, tiers of boxes rose vertically above in a horseshoe plan, and permanent covering (for both acoustics and comfort) made artificial lighting an important feature in theatrical performances. While the modern theatre has been greatly improved in efficiency by new acoustical methods and materials, it also has kept much of the Baroque form. However, it provides seating throughout and usually substitutes sloping galleries (into which the unprivileged have been moved) for boxes. The motion picture has had little effect on theatre design (see THEATRES AND STAGES):

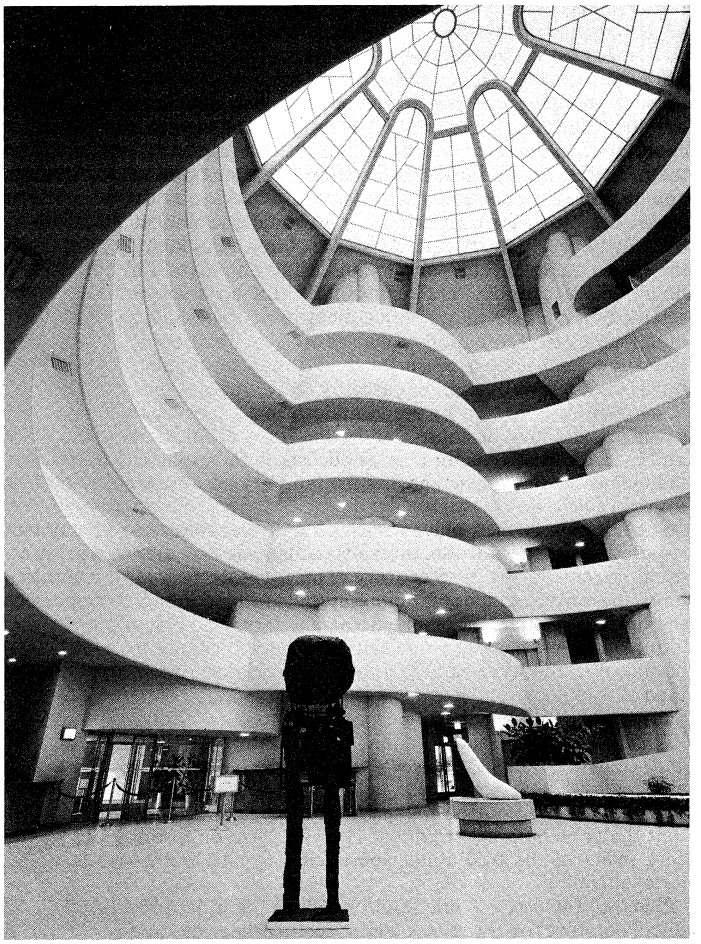
Auditoriums. The auditorium is distinguished by the absence of stage machinery and by its greater size. The development of large symphony orchestras and choirs and of the institution of lectures and mass meetings combined with growing urban populations to produce this modification of the theatre.

Athletic facilities. Sport arenas, racetracks, and public swimming pools of the present day owe their origin to the ancient Romans (though certain precedents can be found in Crete and Greece). Although the classical tradition of sports was broken from the early Middle Ages to the 19th century, even the design of arenas and tracks has been scarcely altered from the Colosseum and Circus Maximus, though the construction of large grandstands has inspired magnificent designs in reinforced concrete (stadiums at Florence, Helsinki, and the Universidad Nacional Autónoma de México). Sports that have no precedents in antiquity, such as baseball, have required modifications in design but have not been important for architecture (see STADIUM).

Importance of Roman prototypes

Museums and libraries. Museum and library architecture was also an innovation of classical antiquity (library architecture appears independently in ancient China and Japan). Early examples are found on the acropolis of Hellenistic Pergamum and in Roman Ephesus. Museums were not cultivated in the Middle Ages, and libraries were incorporated into monasteries. In the Renaissance and Baroque periods, library construction like Johann Bernhard Fischer von Erlach's Hofbibliothek in the Hofburg, Vienna, was rare, but important civic buildings were designed within religious institutions (Michelangelo's Biblioteca Laurenziana in Florence) and universities (Sir Christopher Wren's Trinity College Library, Cambridge; James Gibbs' Radcliffe Camera, Oxford). This type of architecture became truly communal for the first time in the 19th century, when the size of library collections and the number of visitors inspired some of the finest architecture of the modern period (Michael Gottlieb Bindesbøll's Thorvaldsen Museum, Copenhagen; Sir Robert Smirke's British Museum in London; Henri Labrouste's Bibliothèque Sainte-Geneviève in Paris; Alvar Aalto's library in Viipuri, Finland; Frank Lloyd Wright's Solomon R. Guggenheim Museum in New York City). (See MUSEUM; LIBRARY.)

Architecture of welfare and education. The principal institutions of public welfare are those that provide facil-



Interior of the Solomon R. Guggenheim Museum, New York, by Frank Lloyd Wright; opened 1959.
By courtesy of the Solomon R. Guggenheim Museum, New York

ities for education, health, public security, and utilities. Some of these functions are performed by the church and the state, but, since their character is not essentially religious or political, they may require independent architectural solutions, particularly in urban environments. A consistent typology of this architecture, however, cannot be established throughout history, because the acceptance of responsibility for the welfare of the community differs in degree in every social system.

Buildings for the specific purposes of public welfare were seldom considered necessary in antiquity, in most of Eastern architecture, or in the early Middle Ages. But in ancient Greece health facilities were included in precincts of Asclepius, the god of healing, and in the East within Buddhist precincts. The Romans produced a highly developed system of water supply and sewerage, of which their monumental aqueducts are an impressive survival (see AQUEDUCT).

In the later Middle Ages consistent forms began to emerge. With the separation of the university from a purely religious context, a concept of planning developed (particularly at Oxford, Cambridge, and Paris) that still influences educational architecture. Hospitals designed as large halls were established as adjuncts to churches, convents, and monasteries (Hôtel-Dieu, Beaune, France) and gained architectural independence in the Renaissance (Ospedale degli Innocenti, Florence). Ancient and medieval prisons and guardhouses were occasionally isolated from military architecture (*e.g.*, Tower of London; Bargello in Florence), but the prison did not become an important architectural type until the late 18th and 19th centuries (*e.g.*, George Dance's Newgate Prison, London; Henry Hobson Richardson's Allegheny County Jail, Pittsburgh).

The expansion of education and health facilities beginning in the 19th century created a widespread and consistently growing need for specialized architectural solutions. Schools, from the nursery to the university, now demand not only particular solutions at all levels but structures for a variety of purposes within each level; advanced education demands buildings for scientific research, training for trades and professions, recreation, health, housing, religious institutions, and other purposes. Most of the countries of the Western world have produced educational architecture of the highest quality; this architectural type is more important than in any past age.

Specialized solutions demanded since the 19th century

Commercial and industrial architecture. Buildings for exchange, transportation, communication, manufacturing, and power production meet the principal needs of commerce and industry. In the past these needs were mostly unspecialized. They were met either within domestic architecture or in buildings distinguished from domestic types chiefly by their size. Stores, banks, hostelries, guildhalls, and factories required only space for more persons and things than houses could accommodate. Bridges, warehouses, and other structures not used for sheltering people were, of course, specialized from the beginning and survived the Industrial Revolution without basic changes. The Industrial Revolution profoundly affected the typology as well as the techniques of architecture. Through the introduction of the machine and mass production, economic life moved out of the domestic environment into an area dominated by devices and processes rather than by individuals, creating the need for buildings more specialized and more numerous than the total accumulation of types throughout history. All the types cannot be discussed here, but a categorical listing into which they can be fitted will illustrate their importance for architecture: exchange (office buildings, stores, markets, banks, exchanges, warehouses, exhibition halls); transportation (roads, bridges, tunnels; stations for rail, sea, and air transport and the dispensing of fuel; garages, hangars, and other storage facilities; hotels); communication (structures for the transmission and reception of telephone, telegraph, radio, television, and radar communication; for the printing and distribution of newspapers, magazines, books, and other reading matter; for motion-picture production; and for advertising functions); production (mines, factories, laboratories, food-processing plants); power (dams, generating plants; fuel storage, processing, and distribution installations). (See BRIDGES, CONSTRUCTION AND HISTORY OF; DAM; HOTEL AND MOTEL INDUSTRY; LIGHTHOUSE; RESTAURANT; ROADS AND HIGHWAYS.)

Effect of the Industrial Revolution

Each of these functions demands its own architectural solution, but in general they may be divided into two classes according to whether the plan must give greater attention to the size and movement of machinery or of persons. Wherever human activity is the chief concern, there has been less departure from traditional expression; banks in the form of Roman temples are an obvious example. The demands of machines have no tradition and have encouraged a search for greater, simpler, and more flexible spaces; but frequently the practical function has entirely eliminated the expressive, so that with some distinguished exceptions (*e.g.*, Frank Lloyd Wright's S.C. Johnson & Sons, Inc., building, Racine, Wisconsin; Eero Saarinen's General Motors Technical Center, Warren, Michigan), most modern factories are not architecture. Where both men and machines must be given equal attention, as in railroad stations, architecture of the 19th and 20th centuries has vacillated between creating new forms and grasping for irrelevant traditions.

Difference between man- or machine-centred buildings

ARCHITECTURAL PLANNING

The architect usually begins to work when the site and the type and cost of a building have been determined.

The site involves the varying behaviour of the natural environment that must be adjusted to the unvarying physical needs of human beings; the type is the generalized form established by society that must be adjusted to the special use for which the building is required; the cost implies the economics of land, labour, and materials that must be adjusted to suit a particular sum.

Thus, planning is the process of particularizing and, ultimately, of harmonizing the demands of environment, use, and economy. This process has a cultural as well as a utilitarian value, for in creating a plan for any social activity the architect inevitably influences the way in which that activity is performed.

Planning the environment. The natural environment is at once a hindrance and a help, and the architect seeks both to invite its aid and to repel its attacks. To make buildings habitable and comfortable, he must control the effects of heat, cold, light, air, moisture, and dryness and foresee destructive potentialities such as fire, earthquake, flood, and disease.

The methods of controlling the environment considered here are only the practical aspects of planning. They are treated by the architect within the context of the expressive aspects. The placement and form of buildings in relation to their sites, the distribution of spaces within buildings, and other planning devices discussed below are fundamental elements in the aesthetics of architecture.

Importance of determining a building's axis

Orientation. The arrangement of the axes of buildings and their parts is a device for controlling the effects of sun, wind, and rainfall. The sun is regular in its course; it favours the southern and neglects the northern exposures of buildings in the Northern Hemisphere, so that it may be captured for heat or evaded for coolness by turning the axis of a plan toward or away from it. Within buildings, the axis and placement of each space determines the amount of sun it receives. Orientation may control air for circulation and reduce the disadvantages of wind, rain, and snow, since in most climates the prevailing currents can be foreseen. The characteristics of the immediate environment also influence orientation: trees, land formations, and other buildings create shade and reduce or intensify wind, while bodies of water produce moisture and reflect the sun.

Architectural forms. Planning may control the environment by the design of architectural forms that may modify the effects of natural forces. For example, overhanging eaves, moldings, projections, courts, and porches give shade and protection from rain. Roofs are designed to shed snow and to drain or preserve water. Walls control the amount of heat lost to the exterior or retained in the interior by their thickness and by the structural and insulating materials used in making them. Walls, when properly sealed and protected, are the chief defense against wind and moisture. Windows are the principal means of controlling natural light; its amount, distribution, intensity, direction, and quality are conditioned by their number, size, shape, and placement and by the characteristics of translucent materials (*e.g.*, thickness, transparency, texture, colour). But the planning of fenestration is influenced by other factors, such as ventilation and heating. Since most translucent materials conduct heat more readily than the average wall, windows are used sparingly in extreme climates. Finally, since transparent windows are the medium of visual contact between the interior and exterior, their design is conditioned by aesthetic and practical demands.

The role of fenestration

Colour. Colour has a practical planning function as well as an expressive quality because of the range of its reflection and its absorption of solar rays. Since light colours reflect heat and dark colours absorb it, the choice of materials and pigments is an effective tool of environmental control.

Materials and techniques. The choice of materials is conditioned by their own ability to withstand the environment as well as by properties that make them useful to human beings. One of the architect's jobs is to find a successful solution to both conditions; to balance the physical and economic advantages of wood against the possibility of fire, termites, and mold, the weather resistance of glass and light metals against their high thermal conductivity, and many similar conflicts. The more violent natural manifestations, such as heavy snow loads, earthquakes, high winds, and tornadoes, are controlled by special technical devices in regions where they are prevalent.

Any number of these controls may be out of reach of the planner for various reasons. The urban environment, for example, restricts freedom of orientation and design of architectural forms and creates new control problems of its own: smoke, dirt, noise, and odours.

Interior control. The control of the environment through the design of the plan and the outer shell of a building cannot be complete, since extremes of heat and cold, light, and sounds penetrate into the interior, where they can be further modified by the planning of spaces and by special conditioning devices.

Temperature, light and sound are all subject to control by the size and shape of interior spaces, the way in which the spaces are connected, and the materials employed for floors, walls, ceilings, and furnishings. Hot air may be retained or released by the adjustment of ceiling heights and sources of ventilation. Light reflects in relation to the colour and texture of surfaces and may be reduced by dark, rough walls and increased by light, smooth ones. Sounds are transmitted by some materials and absorbed by others and may be controlled by the form of interiors and by the use of structural or applied materials that by their density, thickness, and texture amplify or restrict sound waves.

Early temperature conditioning devices

Conditioning devices played only a small part in architecture before the introduction of mechanical and electrical systems in the 19th century. The fireplace was almost the only method of temperature control (though the ancient Romans anticipated the modern water system for radiant heating); fuel lamps and candles had to be movable and were rather in the sphere of furnishings than of architecture; the same is true of the tapestries and hangings used for acoustical purposes and to block drafts.

Today, heating, insulation, air conditioning, lighting, and acoustical methods have become basic parts of the architectural program. These defenses and comforts of industrialization control the environment so efficiently that the contemporary architect is free to use or to discard many of the traditional approaches to site and interior planning.

Planning for use. While environmental planning produces comfort for the senses (sight, feeling, hearing) and reflexes (respiration), planning for use or function is concerned with convenience of movement and rest. All activities that demand architectural attention require unique planning solutions to facilitate them. These solutions are found by differentiating spaces for distinct functions, by providing circulation among these spaces, and by designing them to facilitate the actions of the human body.

Differentiation. The number of functions requiring distinct kinds of space within a building depends not only upon the type of building but also upon the requirements of the culture and the habits and activities of the individual patrons. A primitive house has a single room with a hearth area, and a modern one has separate areas for cooking, eating, sleeping, washing, storage, and recreation. A meetinghouse with a single hall is sufficient for Quaker religious services, while a Roman Catholic cathedral may require a nave, aisles, choir, apse, chapels, crypt, sacristy, and ambulatory.

The planning of differentiated spaces involves as a guide to their design (placement, size, shape, environmental conditions, sequence, etc.) the analysis of use (number of uses and character, duration, time of day, frequency, variability, etc., of each), users (number, behaviour, age, sex, physical condition, etc.), and furniture or equipment required.

Circulation. Communication among differentiated spaces and between the exterior and the interior may be achieved by openings alone in the simplest plans, but most buildings require distinct spaces allotted to horizontal and vertical circulation (corridors, lobbies, stairs, ramps, elevators, etc.). These are designed by the procedure of analysis employed for differentiating uses. Since their function is usually limited to simplifying the movement of persons and things toward a particular goal, their efficiency depends on making the goal evident and the movement direct and easy to execute.

Facilitation. The convenience of movement, like the comfort of environment, can be increased both by planning and by devices. Planning methods are based on

analysis of the body measurements, movements, and muscular power of human beings of different ages and sexes, which results in the establishment of standards for the measurements of ceilings, doorways, windows, storage shelves, working surfaces, steps, and the like and for the weight of architectural elements that must be moved, such as doors, gates, and windows. These standards also include allowances for the movement of whatever furnishings, equipment, or machinery are required for the use of any building. Devices for facilitating movement within buildings replace or simplify the labours of daily life: the traditional pumps, plumbing, and sewerage systems and the innumerable modern machines for circulation, food preparation and preservation, industrial processing, and other purposes.

Economic planning. Major expenses in building are for land, materials, and labour. In each case they are high when the commodity is scarce and low when it is abundant, and they influence planning more directly when they become restrictive.

Thomas Airviews



Rockefeller Center, New York, by Reinhard & Hofmeister, Corbett, Harrison & MacMurray, and Hood & Fouilhoux, 1929–40; an example of urban planning using skyscrapers.

Cost of land, materials, and labour as planning factors

The effect of high land values is to limit the amount of space occupied by any building as well as the amount of expenditure that can be reserved for construction. When land coverage is limited, it is usually necessary to design in height the space that otherwise would be planned in breadth and depth, as in the ancient Roman *insula* (apartment houses) or the modern skyscraper. When the choice of materials is influenced by cost, all phases of architectural design are affected, since the planning procedure, the technique, and the form of buildings are dependent on materials. High labour costs influence the choice of techniques and, consequently, of materials. They encourage simplification in construction and the replacement of craftsmanship by standardization. The development in the 19th and 20th centuries of light wood-frame construction and methods of prefabrication were largely the result of the rising cost of labour.

Planning involves not only the control of cost in each area but also the proportioning of expenditures among land, materials, and labour in order to produce the most effective solution to an architectural problem.

II. Techniques

The techniques of architecture in the sense that they will be considered here are simply the methods by which structures are formed from particular materials. These methods are influenced not only by the availability and character of materials but also by the total technological development of society, for architecture depends on an organized labour force and upon the existence of the tools and skills necessary to secure, manufacture, transport, and work durable materials.

The influence of economic and expressive forces in the development of technique

The evolution of techniques is conditioned by two forces. One is economic—the search for a maximum of stability and durability in building with a minimum of materials and labour. The other is expressive—the desire to produce meaningful form. Techniques evolve rapidly when economic requirements suggest new expressive forms or when the conception of new forms demands new procedures. But they remain static when architects avoid the risk of pioneering with untried and possibly unsuccessful methods and depend instead on proved procedures or when the need for the observance of tradition, for the communication of ideas, or for elegance and display is best fulfilled by familiar forms.

The ultimate purpose of building techniques is to create a stable structure. In mechanical terms, structures are stable when all their parts are in a state of equilibrium, or rest. Walls and roofs can buckle, crack, or collapse if they are not properly designed. These movements are caused by forces that tend to push or pull bodies in a given direction. Forces acting on any member (part) of a building are, first, its own weight and, second, the loads it carries, principally from other members but also from persons, furnishings, wind, etc. Their action encounters a reaction in opposing forces that hold the member in place by resisting at its joints. These forces may be active in all directions, and they must be balanced for stability. They tend to crush, pull apart, and bend the member—in other words, to change its size and shape.

Within the member itself there are forces, too, that tend to resist any deformation. They are called stresses, and they vary according to the strength of materials and the form of the member. The kinds of stress under consideration are compression, which resists crushing; tension, which resists pulling apart; and bending, which occurs when one part of a member is in compression and the other is in tension. A column is put into compression by the loads it carries; in a trussed roof the piece that forms the base of the triangle is put into tension by the outward-pushing forces in the sides; and a lintel or beam (the member that spans a space) is put into bending by loads and forces that push down on its top and encounter a reacting force at its ends. Some materials are strong only in compression (*e.g.*, stone, brick, cast iron, concrete) and others in tension as well (*e.g.*, wood, steel, reinforced concrete), so the latter are more efficient in resisting bending forces.

Finally, the stability of the total structure whose single members are all in equilibrium is achieved by diverting the loads from all of them downward so that they may be resisted by the upward-supporting forces of the ground.

Techniques will be discussed in terms of the characteristics of building materials and the methods by which they are used in architecture (see BUILDING CONSTRUCTION; CARPENTRY).

MATERIALS

Stone. In most areas where stone is available, it has been favoured over other materials for the construction of monumental architecture. Its advantages are durability, adaptability to sculptural treatment, and the fact that it can be used in modest structures in its natural state. But it is difficult to quarry, transport, and cut, and its weakness in tension limits its use for beams, lintels, and floor supports.

The simplest and cheapest stonework is rubble; *i.e.*, roughly broken stones of any shape bounded in mortar. The strongest and most suitable stonework for monumental architecture is ashlar masonry, which consists of regularly cut blocks (usually rectangular). Because of its weight and the precision with which it can be shaped, stone masonry (in contrast with brick) does not depend on strong bonding for stability where it supports only direct downward loads. The entablatures (the upper sections of a classical order that rest on the capital of a column) of an ancient Greek temple, for example, were bonded by small bronze dowels. But the weight creates problems of stability when loads push at an angle; stone vaults and arches require more support and buttressing than equivalent forms in other materials.

Durability of stone

The best stone (and brick) bonding is that in which blocks are placed so that the vertical joints in one course are not above the joints in the courses above and below, since the stone resists deformation better than any bonding material. Many stones are strong enough to provide monolithic supports (columns and piers) and beams (lintels); and in some styles stone slabs are employed even for roofing (ancient Egyptian temples, early Christian basilicas in Syria). But this roofing requires so many columns that unvaulted masonry buildings are almost always combined with floors and covering in wood. Stone has been consistently used for building since the Stone Age, as exemplified by Stonehenge, in England. Although it has generally been replaced as a structural material by cheaper and more efficient manufactured products, it is still widely used as a surface veneer for its practical and expressive qualities.

Brick. Brick compares favourably with stone as a structural material for its fire- and weather-resisting qualities and for the ease of production, transportation, and laying. The size of bricks is limited by the need for efficient drying, firing, and handling, but shapes, along with the techniques of bricklaying, have varied widely throughout history. Special shapes can be produced by molding to meet particular structural or expressive requirements (for example, wedge-shaped bricks are sometimes employed in arch construction and bricks with rounded faces in columns). Bricks may be used in construction only in conjunction with mortar, since the unit is too small, too light, and too irregular to be stabilized by weight. Each course must be laid on an ample mortar bed with mortar filling the vertical joints. The commonest ancient Roman bricks were cut into triangles and laid with the base out and the apex set into a concrete filling that provided additional strength. Rectangular bricks are bonded either as headers (short side out) or stretchers (long side out). Standard modern types provide a ratio of width to length of slightly less than 1:2 to permit a wide variety of bonding patterns within a consistent module, or standard of measurement. Brick, which has been used since the 4th millennium BC, was the chief building material in the ancient Near East. The versatility of the medium was expanded in ancient Rome by improvements in the manufacture of both bricks and mortar and by new techniques of laying and bonding. Employed throughout the Middle Ages, brick gained greater popularity from the 16th century on, particularly in northern Europe. It is widely used in the 20th century, often for nonbearing walls in steel frame construction (see MASONRY).

Wood. Wood is easier to acquire, transport, and work than other natural materials. All parts of a building can be efficiently constructed of wood except foundations; its disadvantage is susceptibility to fire, mold, and termites.

Organic nature of wood structures

The strength of wood in both tension and compression arises from its organic nature, which gives it an internal structure of longitudinal and radial fibres that is not impaired by cutting or long exposure. But like all organisms it contains moisture and is not uniformly strong, so it must be carefully selected and seasoned to prevent warping, splitting, and failure under loads. Wood is used in building both solid and skeletal structures. The principal solid system, called log construction, is employed when only primitive cutting tools are available. Four walls must be built up together in horizontal layers of single hewn or uncut logs and jointed at the corners. The stability of the log building depends entirely on the mutual support of the walls, and the method is suitable only for simple structures of limited size. The skeletal system requires precise cutting and shaping of lumber. It provides a rigid framework of jointed or nailed members independent of the walls, which are attached to the exterior and interior surfaces after completion.

Almost all masonry buildings of the past had wood floors and coverings, since wood is the lightest, the most practical, and the most inexpensive material for spanning spaces.

The monumental architecture of the West has typically employed materials rarer than wood for expressive purposes, but the history of wood construction can be traced consistently in China, Korea, and Japan and in the domestic architecture of northern Europe and North America. Wood continues to be used in a growing number of techniques and products: heavy framing systems with compound beams and girders, interior and exterior facing with plywood and other composite panels, and arch and truss systems with laminated members that can be designed to meet particular structural demands (see WOOD AND WOOD PRODUCTS).

Iron and steel. The development of construction methods in iron and steel was the most important innovation in architecture since ancient times. These methods provide far stronger and taller structures with less expenditure of material than stone, brick, or wood and can produce greater unsupported spans over openings and interior or exterior spaces. The evolution of steel frame construction in the 20th century entirely changed the concept of the wall and the support.

In architecture before 1800, metals played an auxiliary role. They were used for bonding masonry (dowels and clamps), for tension members (chains strengthening domes, tie rods across arches to reinforce the vaults), and for roofing, doors, windows, and decoration. Cast iron, the first metal that could be substituted for traditional structural materials, was used in bridge building as early as 1779. Its ability to bear loads and to be produced in an endless variety of forms, in addition to its resistance to fire and corrosion, quickly encouraged architectural adaptations, first as columns and arches and afterward in skeletal structures. Because cast iron has much more compressive than tensile strength (for example, it works better as a small column than as a beam), it was largely replaced in the late 19th century by steel, which is more uniformly strong, elastic, and workable, and its high resistance in all stresses can be closely calculated.

Advantages of steel construction

Steel structural members are rolled in a variety of shapes, the commonest of which are plates, angles, I beams, and U-shaped channels. These members may be joined by steel bolts or rivets, and the development of welding in the 20th century made it possible to produce fused joints with less labour and materials. The result is a rigid, continuous structure in which the joint is as firm as the member and which distributes stresses between beams and columns. This is a fundamental change in architectural technique, the effect of which cannot yet be estimated.

Normally, steel must be protected against corrosion by surface coverings, but alloys such as stainless steel have been developed for exposed surfaces. Aluminum and other light metal alloys have come to be favoured for exterior construction because of their weather resistance.

Concrete. Concrete is a manufactured mixture of cement and water, with aggregates of sand and stones, which hardens rapidly by chemical combination to a stonelike, water- and fire-resisting solid of great compressive (but low tensile) strength. Because it can be poured into forms while liquid to produce a great variety of structural elements, it provides an economical substitute for traditional materials, and it has the advantages of continuity (absence of joints) and of fusing with other materials.

Concrete was employed in ancient Egypt and was highly developed by the ancient Romans, whose concrete made with volcanic-ash cement (pozzolana) permitted a great expansion of architectural methods, particularly the development of domes and vaults (often reinforced by brick ribbing) to cover large areas, of foundations, and of structures such as bridges and sewerage systems where waterproofing was essential. The technique of manufacture declined in the Middle Ages and was regained in the 18th century, but concrete had only a limited importance for architecture until the invention of reinforced concrete in the 1860s.

Development of reinforced concrete

Reinforced concrete was developed to add the tensile strength of steel to the compressive strength of mass concrete. The metal is embedded by being set as a mesh into the forms before pouring, and in the hardened material the two act uniformly. The combination is much more versatile than either product; it serves not only for constructing rigid frames but also for foundations, columns, walls, floors, and a limitless variety of coverings, and it does not require the addition of other structural materials. Although the making of forms is a slow and costly process, the technique competes economically with steel frame construction because the mesh, composed of thin, bendable metal rods or metal fabric, employs far less steel, and concrete is itself inexpensive.

The steel reinforcement is employed to take full advantage of the plastic, or sculptural, character of concrete. It can be jointed or bent to unify supporting members with the floors and the coverings they carry. Furthermore, stresses produced in floors, domes, and vaults may be distributed within the slabs themselves to reduce load, and the diminished load may be concentrated at desired points so that the number and size of supports is greatly reduced.

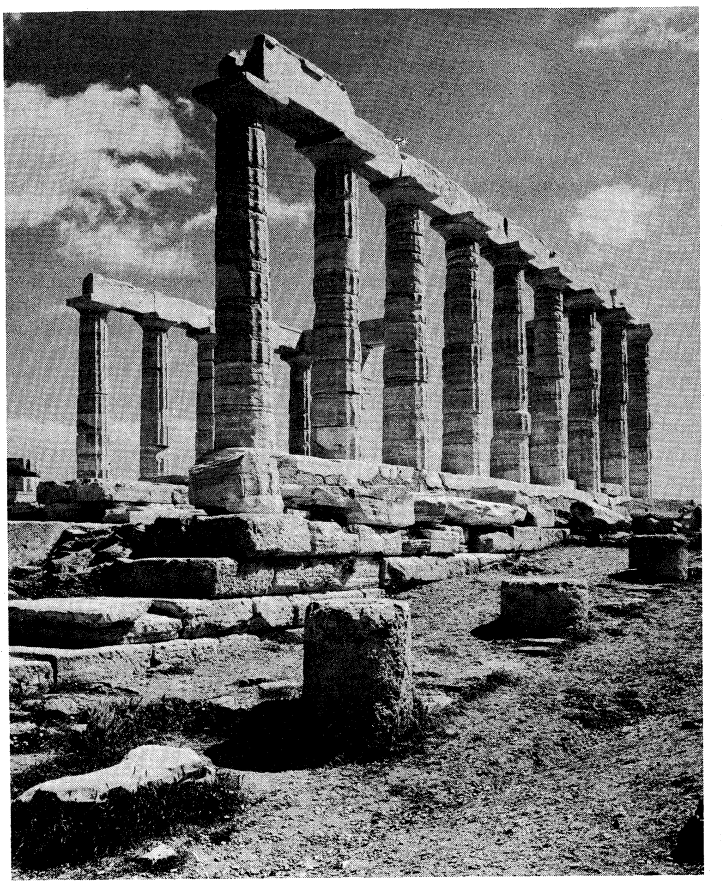
Concrete shells, precast concrete, and prestressed concrete

Three 20th-century developments in production are destined to have a radical effect on architecture. The first, concrete-shell construction, permits the erection of vast vaults and domes with a concrete and steel content so reduced that the thickness is comparatively less than that of an eggshell. The second development, precast-concrete construction, employs bricks, slabs, and supports made under optimal factory conditions to increase waterproofing and solidity, to decrease time and cost in erection, and to reduce expansion and contraction. Finally, prestressed concrete provides bearing members into which reinforcement is set under tension to produce a live force to resist a particular load. Since the member acts like a spring, it can carry a greater load than an unstressed member of the same size (see CEMENT; CONCRETE).

METHODS

Wall. The two types of wall are load bearing, which supports the weight of floors and roofs, and nonbearing, which at most supports its own weight.

Load-bearing wall. The load-bearing wall of masonry is thickened in proportion to the forces it has to resist: its own load, the load of floors, roofs, persons, etc., and the lateral forces of arches, vaults, wind, etc., that may cause it to crack or buckle. Its thickness often can be reduced at the top, because loads accumulate toward the base; in high buildings this is done by interior or exterior setbacks at the floor level of upper stories. Walls that must resist lateral forces are thickened either along the whole length or at particular points where the force is concentrated. The latter method is called buttressing. Doors and windows weaken the resistance of the wall and divert the forces above them to the parts on either side, which must be thickened in proportion to the width of the opening. In multistory buildings, windows—unless they are very small—must be placed one above the other so as to leave uninterrupted vertical masses of wall between them to transfer loads directly to the ground. The number of openings that can be used depends on the strength of the masonry and the stresses in the wall. Walls in light, wood-framed structures and in reinforced-concrete construction may have a bearing function also. But the nature of the material admits other means of resisting forces than the increase of mass.



Post-and-lintel system of support, temple of Poseidon, Cape Soúnion, southeast of Athens, *c.* 430 BC.
Toni Schneiders

The placement of walls is determined by the type of support for floors and roofs. The commonest support is the beam, which must be jointed to walls at both ends; consequently, its maximum permissible length establishes the distance between bearing walls. All floors and coverings are most easily supported on straight, parallel walls except the dome (see below *Dome*).

Nonbearing wall. Excluding the independent garden variety, the nonbearing wall appears only where loads are carried by other members, as in heavy timber and other skeletal structures. Modern steel and reinforced-concrete frames require exterior walls only for shelter and sometimes dispense with them on the ground floor to permit easier access. Since the wall rests or hangs upon members of the frame, it becomes a curtain or screen and admits treatment in any durable, weather-resisting material. Traditional materials are often used, but light walls of glass, plastic, metal alloys, wood products, etc., can be equally efficient. This freedom of choice extends also to the form of walls and offers greatly expanded opportunities for creative expression.

Curtain walls

Post-and-lintel. The simplest illustration of load and support in construction is the post-and-lintel system, in which two upright members (posts, columns, piers) hold up a third member (lintel, beam, girder, rafter) laid horizontally across their top surfaces. This is the basis for the evolution of all openings. But, in its pure form, the post-and-lintel is seen only in colonnades and in framed structures, since the posts of doors, windows, ceilings, and roofs are part of the wall.

The job of the lintel is to bear the loads that rest on it (and its own load) without deforming or breaking. Failure occurs only when the material is too weak or the lintel is too long. Lintels composed of materials that are weak in bending, such as stone, must be short, while lintels in materials that are strong in bending, such as steel, may span far greater openings. Masonry lintels are inefficient because they must depend on the cohesiveness

Lintel materials

of mortar, which is weaker than the blocks it bonds; so, in masonry construction, lintels of monolithic (single-slab) stone, wood, and stronger materials are employed.

The job of the post is to support the lintel and its loads without crushing or buckling. Failure occurs, as in lintels, from excessive weakness or length, but the difference is that the material must be especially strong in compression. Stone, which has this property, is more versatile as a post than as a lintel; under heavy loads it is superior to wood but not to iron, steel, or reinforced concrete. Masonry posts, including those of brick, may be highly efficient, since the loads compress the joints and add to their cohesiveness. Although monolithic stone columns are used, they are extravagant to produce for large structures, and columns are usually built up of a series of cylindrical blocks called drums.

From prehistoric times to the Roman Empire, the post-and-lintel system was the root of architectural design. The interiors of Egyptian temples and the exteriors of Greek temples are delineated by columns covered by stone lintels. The Greeks opened their interior spaces by substituting wooden beams for stone, since the wood required fewer supports. The development of the arch and vault challenged the system but could not diminish its importance either in masonry construction or in wood framing, by its nature dependent on posts and beams.

Ancient uses of the post-and-lintel were refined but not fundamentally altered until the production of cast-iron columns, which, offering greater strength and smaller circumference, greatly reduced the mass and weight of buildings. Much construction in modern materials is based on the post-and-lintel system of the past. Steel and concrete skeletons restore to modern architecture the formal simplicity of the oldest structures known. But, because they are rigid frames, they abandon the fundamental concept of the duality of post-and-lintel by fusing them into a unit throughout which stresses are distributed. The "mushroom" column is a further departure, since the unit can be extended into a covering slab and becomes a ceiling as well as a support.

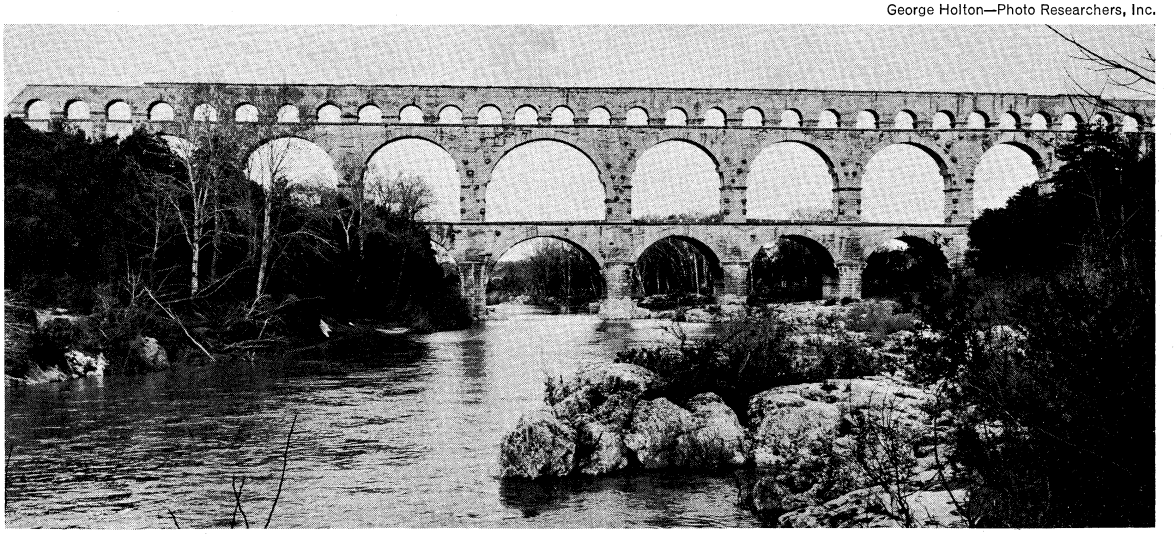


"Mushroom" column with fountain, supporting a cantilever, Museo Nacional de Antropología, Mexico City, by Pedro Ramirez Vaquez, 1964.
Victor Englebert—De Wys, Inc.

Arch. The arch can be called a curved lintel. Early masonry builders could span only narrow openings because of the necessary shortness and weight of monolithic stone lintels. With the invention of the arch, two problems were solved: (1) wide openings could be spanned with small, light blocks, in brick as well as stone, which were easy to transport and to handle; and (2) the arch was bent upward to resist and to conduct into its supports the loads that tended to bend the lintel downward. Because the arch is curved, the upper edge has a greater circumference than the lower, so that each of its blocks must be cut in wedge shapes that press firmly against the whole surface of neighbouring blocks and conduct loads uniformly. This form creates problems of equilibrium that do not exist in lintels. The stresses in the arch tend to squeeze the blocks outward radially, and loads divert these outward forces downward to exert a resultant diagonal force, called thrust, which will cause the arch to collapse if it is not properly buttressed. So an arch cannot replace a lintel on two free-standing posts unless the posts are massive enough to buttress the thrust and to conduct it into the foundations (as in ancient Roman triumphal arches). Arches may rest on light supports, however, where they occur in a row, because the thrust of one arch counteracts the thrust of its neighbours, and the system will remain stable as long as the arches at either end of the row are buttressed by walls, piers, or earth.

Problems of equilibrium in the arch

George Holton—Photo Researchers, Inc.



Pont du Gard, Roman aqueduct, Nîmes, France, by Marcus Vipsanius Agrippa, during the Augustan period.



Two types of vaulting used in the Medieval period.
(Left) Barrel-vaulted nave of Saint-Savin-sur-Gartempe, France, early 12th century.
(Right) Rib-vaulted nave of the Bourges Cathedral, France, begun *c.* 1195.
Left) Jean Roubier; (right) Archives Photographiques

The size of arches is limited only by economy; large arches exert large thrusts, and they are hard to buttress and to build. The form may be varied to meet specific problems; the most efficient forms in masonry are semicircular, segmental (segment of a circle), and pointed (two intersecting arcs of a circle), but noncircular curves can be used successfully.

Arches were known in Egypt and Greece but were considered unsuitable for monumental architecture. In Roman times the arch was fully exploited in bridges, aqueducts, and large-scale architecture. New forms and uses were found in medieval and particularly Gothic architecture (flying buttress, pointed arch), and Baroque architects developed a vocabulary of noncircular forms for expressive reasons. Steel, concrete. and laminated-wood arches of the 20th century have changed the concept and the mechanics of arches. Their components are completely different from wedge-shaped blocks; they may be made entirely rigid so as to require only vertical support; they may be of hinged intersections that work independently, or they may be thin slabs or members (in reinforced concrete) in which stresses are so distributed that they add the advantages of lintels to those of arches, requiring only light supports. These innovations provide a great freedom of design and a means of covering great spans without a massive substructure.

Vault. The evolution of the vault begins with the discovery of the arch, because the basic "barrel" form, which appeared first in ancient Egypt and the Near East, is nothing more than an arch deep enough to cover a three-dimensional space. Since the barrel vault exerts thrust as the arch does, it must be buttressed along its entire length by heavy walls in which openings must be limited in size and number. This is a disadvantage, since it inhibits light and circulation.

Vaults in antiquity and the Middle Ages

But Roman builders discovered that openings could be made by building two barrel vaults that intersected at right angles to form the groin vault, which is square in plan and may be repeated in series to span rectangular areas of unlimited length. This vault has the additional advantage that its thrusts are concentrated at the four corners, so that the supporting walls need not be uniformly massive but may be buttressed where they support the vault.

Two disadvantages of the groin vault encouraged Gothic builders to develop a modification known as the rib vault. First, to build a groin vault, a form must be made to pour or lay the entire vault, and this requires complex scaffolding from the ground up; second, the groin vault must be more or less square, and a single vault cannot span extended rectangular areas. The rib vault provided a skeleton of arches or ribs along the sides of the area and crossing it diagonally; on these the masonry of the vault could be laid; a simple centring sufficed for the ribs. To cover the rectangular areas, the medieval mason used pointed arches, which, unlike round arches, can be raised as high over a short span as over a long one. Thus, the vault could be composed of the intersection of two vaults of different widths but the same height.

To reduce further the thickness of the wall (to the point of substituting large areas of glass for masonry), Gothic builders developed the flying buttress, which counteracts vault thrust not by continuous wall mass and weight but by counterthrust created by exterior half-arches placed at the height of the vaults at the points of greatest stress. These buttresses conduct stresses to heavier wall buttresses below the window level.

The next important development in vaults, as in arches, came with 19th-century materials. Great iron skeleton vaults were constructed as a framework for light mate-

Modern methods of vaulting

rials such as glass (Crystal Palace, London). The elimination of weight and excessive thrust, the freedom in the use of materials, and the absence of centring problems favoured the simple barrel vault and made more complex types obsolete. But in many of the modern frame systems the vault itself loses its structural function and becomes a thin skin laid over a series of arches.

While the arch is supplanting the vault in one area of technique, the vault has abandoned the arch principle in another. The reinforced-concrete shell vault, based on the principle of the bent or molded slab, is one of the most important innovations in the history of architecture. It has all the advantages of load distribution of the concrete floor slab, plus the resistance to bending provided by its curved form. The shell is reinforced in such a way that it exerts no lateral thrust and may be supported as if it were a beam or truss; hence, the form no longer necessitates the conducting of loads into the wall, and the vault may be designed with great freedom.

Dome. Domes appeared first on round huts and tombs in the ancient Near East, India, and the Mediterranean region but only as solid mounds or in techniques adaptable only to the smallest buildings. They became technically significant with the introduction of the large-scale masonry hemispheres by the Romans. Domes, like vaults, evolved from the arch, for in their simplest form they may be thought of as a continuous series of arches, with the same centre. Therefore, the dome exerts thrusts all around its perimeter, and the earliest monumental examples required heavy walls. Since the walls permitted few openings and had to be round or polygonal to give continuous support, early domes were difficult to incorporate into complex structures, especially when adjacent spaces were vaulted.

Development of pendentives

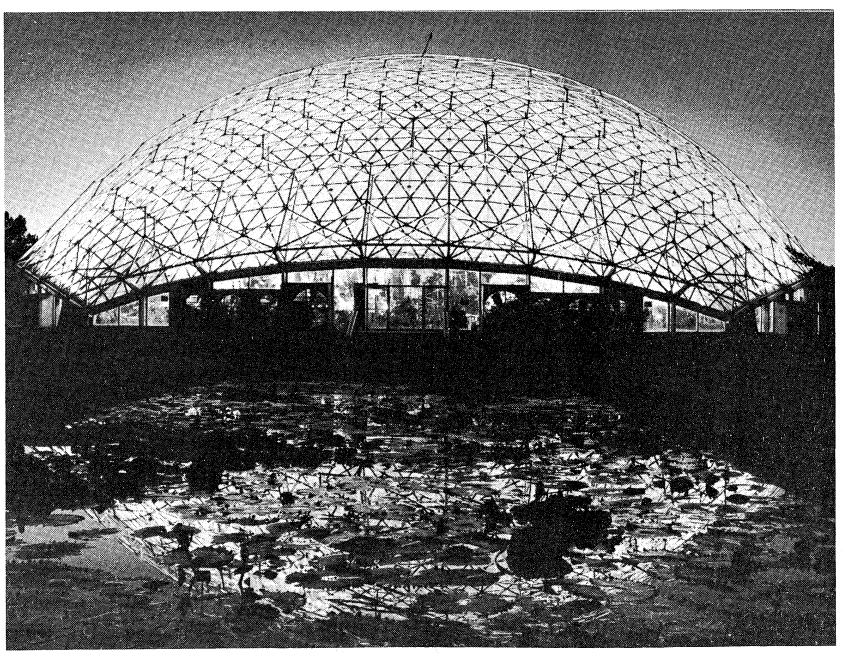
Byzantine architects perfected a way of raising domes on piers instead of walls (like groin vaults), which permitted lighting and communication from four directions. The transition from a cubic plan to the hemisphere was achieved by four inverted spherical triangles called pendentives—masses of masonry curved both horizontally and vertically. Their apexes rested on the four piers, to which they conducted the forces of the dome; their sides joined to form arches over openings in four faces of the cube; and their bases met in a complete circle to form the dome foundation. The pendentive dome could rest directly on this foundation or upon a cylindrical wall, called a drum, inserted between the two to increase height.

The dome was unsuited to the lightness and verticality of late-medieval styles but was widely used in the Renaissance and Baroque periods. Renaissance builders adapted the Gothic rib system to dome construction and found new means to reduce loads and thrust (concentric chains, etc.) that permitted high drums and variations in the curvature of the dome. The awkward, tunnellike effect produced on the interior by high domes was often hidden by an internal shell built on the same foundations (as at Florence Cathedral and St. Paul's Cathedral, London).



Interior showing dome on pendentives, Hagia Sophia, Istanbul, by Anthemius of Tralles and Isidore of Miletus, completed 537.
Hirmer Fotoarchiv, Munich

Buckminster Fuller

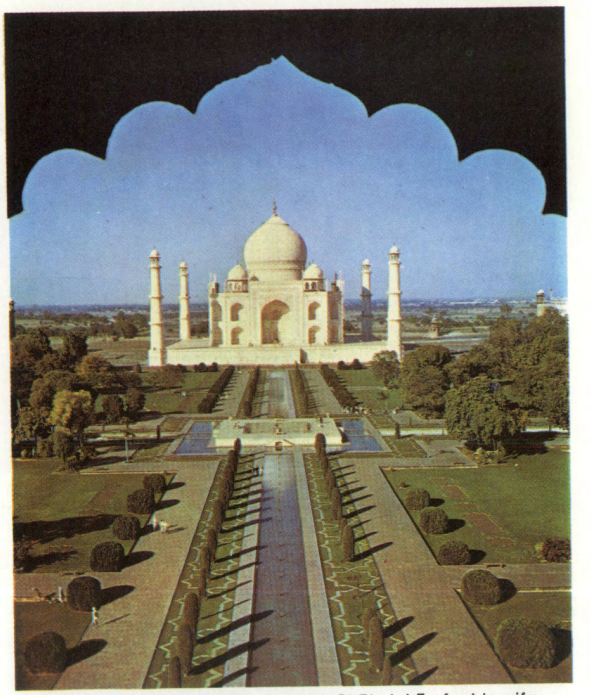


Climatron, geodesic dome, Missouri Botanical Garden, St. Louis, by R. Buckminster Fuller, opened 1960.

Variety of architectural types



Säynätsalo town hall group, Finland, designed by Alvar Aalto, 1950–52: governmental architecture in which indigenous building traditions and materials are combined with modern design and building technology.



The Tāj Mahal at Agra, India, built by Shāh Jahān for his wife, Mumtāz Mahal, 1632–49: marble mausoleum, with formal gardens.



Place des Vosges, Paris, built by the French king, Henry IV, 1605–12: innovative aristocratic group housing with a residential square and an arcaded walking area.

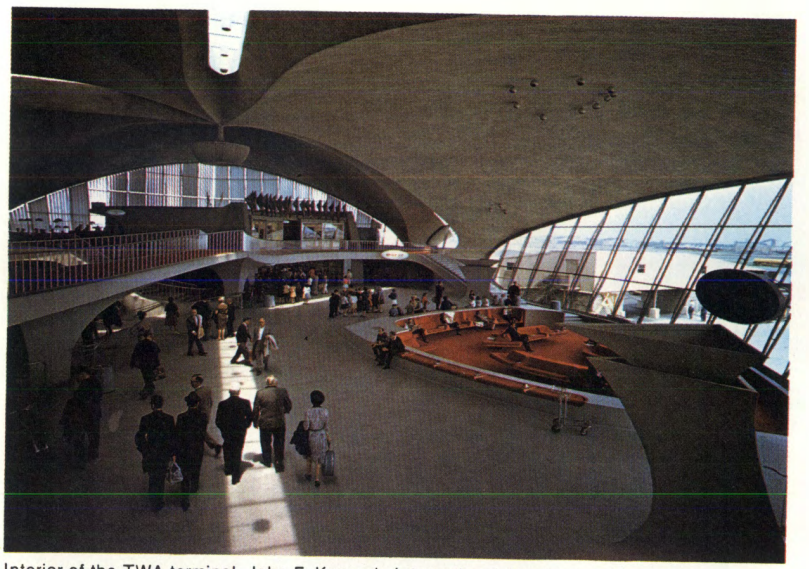


Alcázar at Segovia, Spain, built by Henry IV of Castile, 15th century: combination of military and residential architecture that uses siting functionally, as well as for psychological and aesthetic effect.



Auditorium of the Auditorium Building, Chicago, by Dankmar Adler and Louis Sullivan, 1889, renovated 1967–68: horseshoe-planned theatre with boxes and sloping galleries engineered for acoustical perfection and advantageous sight lines.

Plate 1: (Top left) GEKS, (top right) C.A. Peterson—Publix, (centre left) © Interphototheque Doc. Francaise. SODEL. Mr. Brigaud, (bottom left) Erwitt—Magnum, (bottom right) Gerald Clyde—FPG



Interior of the TWA terminal, John F. Kennedy International Airport, New York, by Eero Saarinen, 1956–62: imaginative sculptural use of reinforced concrete.



Borgund Stave church, Norway, *c.* 1163: rich visual texture and colour created by wood.

Diverse kinds of building materials and architectural surface effects



The Seagram Building, New York City, by Ludwig Mies van der Rohe and Philip Johnson, 1956–58: functionalism and elegance of design expressed in steel frame and curtain wall construction.



Contemporary vernacular architecture in Zaria, Nigeria: clay houses decorated with low relief ornament and vibrantly coloured designs.



"Fallingwater," Mill Run, Pennsylvania, residence by Frank Lloyd Wright, 1936–38: sensitivity to the environment expressed by site orientation and use of natural as well as man-made materials.

Plate 2: By courtesy of (bottom right) the Western Pennsylvania Conservancy, Pittsburgh; photographs, (top left) Marvin B. Winter—Photo Researchers, Inc., (top right) C.A. Peterson—Publix, (centre right) Frank Willett, (bottom left) Photo Media, Ltd.



Nave with view toward the apse and crypt, showing roof trusses, S. Miniato al Monte, Florence, 11th century.
Alinari

The effort and ingenuity devoted to doming rectangular buildings can be explained principally by the symbolic character of the form, since vaulting is a simpler alternative. So it was chiefly the desire to observe tradition that preserved the dome in the early era of iron and steel construction, and, with rare exceptions (Halle aux Blés, Paris; the Coal Exchange, London), 19th-century examples retained masonry forms without exploiting the advantages of metal.

Modern techniques of vault construction

Newer techniques, however, have added practically to the expressive advantages of domes. The reinforced-concrete slab used in vaulting can be curved in length as well as width (like an inflated handkerchief or a parachute). And in this development the distinction between vaults and domes loses significance, being based on nothing but the type of curvature in the slab. Geodesic domes, developed in the 20th century by R. Buckminster Fuller (*q.v.*), are spherical forms in which triangular or polygonal facets composed of light skeletal struts or flat planes replace the arch principle and distribute stresses within the structure itself, as in a truss. Geodesic domes can be supported by light walls and are the only large domes that can be set directly on the ground as complete structures.

Truss. By far the commonest covering throughout history is the trussed roof, constructed upon a frame composed of triangular sections spaced crosswise at intervals and made rigid in length by beams. Trusses formerly were principally of wood and were used to cover masonry as well as framed structures, even when these were vaulted. The variety of trusses is so great that only the general principle of the form can be given here.

Principle of the truss system

The truss is based on the geometric law that a triangle is the only figure that cannot be changed in shape without a change in the length of its sides; thus, a triangular frame of strong pieces firmly fastened at the angles cannot be deformed by its own load or by external forces such as wind pressure. These forces, which in a vault thrust outward against the walls, are contained within the truss itself, because the piece (chord) at the base of the triangle resists by tension the tendency of the two sides to behave like a vault. With its forces in equilibrium, the truss exerts only a direct downward pressure on the walls, so that they need not be thickened or buttressed. This explains why most roofs are triangular in cross section.

In trusses that are too large to be constructed of three members of moderate size, a complex system of small triangles within the frame replaces the simple triangle.

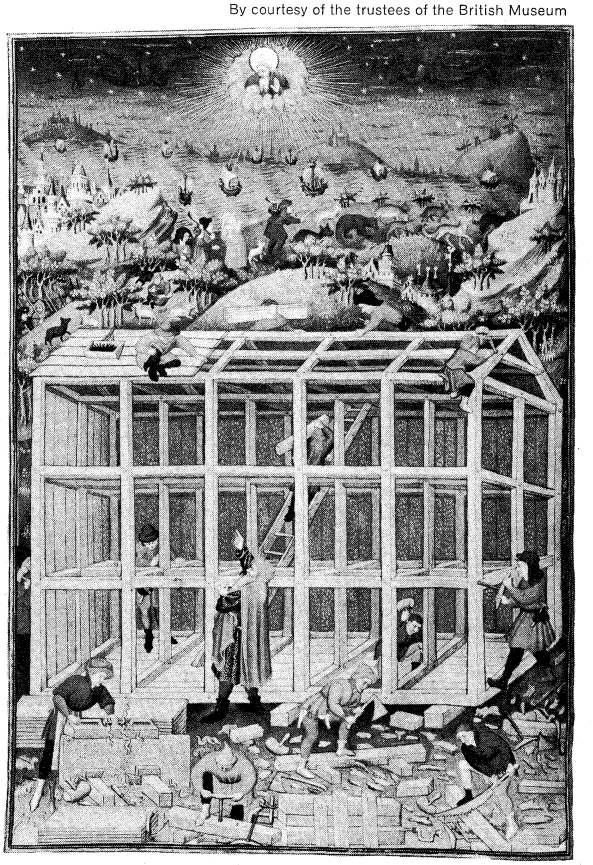
Not all peaked roofs are trusses, for in primitive building, in ancient Greece, and in much Chinese and Japanese wood architecture the chord is omitted and the sides exert thrust. Nor are all trusses triangular, since the principle may be modified (as in modern steel and heavy timber construction) to apply to arches and vaults if chords of sufficient strength can be found.

Framed structures. A framed structure in any material is one that is made stable by a skeleton that is able to stand by itself as a rigid structure without depending on floors or walls to resist deformation. Materials such as wood, steel, and reinforced concrete, which are strong in both tension and compression, make the best members for framing. Masonry skeletons, which cannot be made rigid without walls, are not frames. The heavy timber frame, in which large posts, spaced relatively far apart, support thick floor and roof beams, was the commonest type of construction in eastern Asia and northern Europe from prehistoric times to the mid-19th century. It was supplanted by the American light wood frame (ballon frame), composed of many small and closely spaced members that could be handled easily and assembled quickly by nailing instead of by the slow joinery and dowelling of the past. Construction is similar in the two systems, since they are both based on the post-and-lintel principle. Posts must rest on a level, waterproof foundation, usually composed of masonry or concrete, on which the sill (base member) is attached. Each upper story is laid on crossbeams that are supported on the exterior wall by horizontal members. Interior walls give additional beam support.

The American ballon frame

In the heavy-timber system, the beams are strong enough to allow the upper story and roof to project be-

By courtesy of the trustees of the British Museum



Miniature showing heavy timber frame construction, from the "Bedford Book of Hours," British Museum, Add. 18850, Folio 15 v., French, 15th century.

yond the plane of the ground-floor posts, increasing the space and weather protection. The members are usually exposed on the exterior. In China, Korea, and Japan, spaces between are enclosed by light screen walls and in northern Europe partly by thinner bracing members and partly by boards, panels, or (in half-timbered construction) bricks or earth.

The light frame, however, is sheathed with vertical or horizontal boarding or shingling, which is jointed or overlapped for weather protection. Sheathing helps to brace as well as to protect the frame, so the frame is not structurally independent as in steel frame construction. The light-frame system has not been significantly improved since its introduction, and it lags behind other modern techniques. Prefabricated panels designed to reduce the growing cost of construction have not been widely adopted. Modern heavy-timber and laminated-wood techniques, however, provide means of building up compound members for trusses and arches that challenge steel construction for certain large-scale projects in areas where wood is plentiful.

Steel-frame structures

Steel framing is based on the same principles but is much simplified by the far greater strength of the material, which provides more rigidity with fewer members. The load-bearing capacity of steel is adequate for buildings many times higher than those made of other materials. Because the column and beam are fused by riveting or welding, stresses are distributed between them, and both can be longer and lighter than in structures in which they work independently as post-and-lintel. Thus, large cubic spaces can be spanned by four columns and four beams, and buildings of almost any size can be produced by joining cubes in height and width. Since structural steel must be protected from corrosion, the skeleton is either covered by curtain walls or surfaced in concrete or, more rarely, painted. The steel frame is used also in single-story buildings where large spans are required. The simple cube then can be abandoned for covering systems employing arches, trusses, and other elements in a limitless variety of forms in order to suit the functions of the building.

Differences between reinforced-concrete and steel framing are discussed in the section on materials. The greater rigidity and continuity of concrete frames give them more versatility, but steel is favoured for very tall structures for reasons of economy in construction and space. An example is the system called box frame construction, in which each unit is composed of two walls bearing a slab (the other two walls enclosing the unit are nonbearing curtain walls); this type of construction extends the post-and-lintel principle into three dimensions. Here, again, concrete crosses the barriers that separated traditional methods of construction.

III. Expression

Expression in architecture is the communication of quality and meaning. The functions and the techniques of building are interpreted and transformed by expression into art, as sounds are made into music and words into literature.

The nature of expression varies with the character of culture in different places and in different times, forming distinct modes or languages of expression that are called styles. Style communicates the outlook of a culture and the concepts of its architects. The boundaries of a style may be national and geographical (*e.g.*, Japanese, Mayan) or religious (*e.g.*, Islāmic) and intellectual (*e.g.*, Renaissance), embracing distinct linguistic, racial, and national units, and different expressions within each of these boundaries are produced by the particular style of regions, towns, groups, architects, or craftsmen. The lifespan of styles may be long (ancient Egyptian, over 3,000 years) or short (Baroque, less than 200 years) according to the changeability of cultural patterns. The principal forces in the creation of a style are tradition, the experience of earlier architecture; influence, the contribution of contemporary expressions outside the immediate cultural environment; and innovation, the creative contribution of the culture and the architect. These forces operate to produce an evolution within every style and ultimately to generate new styles that tend to supplant their predecessors.

The components of expression, which communicate the particular values of style, are content and form. Since content can be communicated only through form, the two are organically united, but here they will be discussed separately in order to distinguish the specific and concrete meaning (content) from the abstract expression of qualities (form).

CONTENT

Content is the subject matter of architecture, the element in architectural expression that communicates specific meanings that interpret to society the functions and techniques of buildings.

Symbols of function. Society requires that architecture not only communicate the aspirations of its institutions but also fulfill their practical needs. Differences in expression, apart from differences in planning, distinguish the forms of architectural types (the house from the church, etc.), the kinds of use (the Catholic from the Protestant church), and the traditions and customs of users (the English from the Swiss Protestant church). When architectural forms become the vehicles of content—in plan, elevation, and decoration—they are symbolic. Their symbolism can be understood consciously or unconsciously, by association (*e.g.*, spire = church) to a building one has seen before and by the fact that it suggests certain universal experiences (*e.g.*, vertical forms "rise"; low roofs "envelop"). One comprehends the meaning of symbols that are new, as well as those that are known, by association, because the laws of statics restrain builders from putting them into forms so completely unfamiliar that they do not suggest some tradition, just as the structure of language permits endless new meanings but retains a fairly constant vocabulary. The meaning of architectural symbols—or of words—may even change, but the process must be both logical and gradual, for, if the change is irrational, the purpose—communication—is lost.

Expressive identity through symbolic form

The architectural plan, when used symbolically, communicates through its shape. From prehistoric times and in many cultures, the circle, with its suggestion of the planets and other manifestations of nature, gained a symbolic, mystical significance and was used in the plans of houses, tombs, and religious structures. By slow processes it came to be employed for *memoria* and shrines and for hero cults in both the East and the West. When building techniques permitted, its symbolism often merged with that of the dome. In Hindu temples, the square (and the cross plans developed from it) expressed celestial harmony. The central-plan Christian church (circle, polygon, Greek cross, ellipse) fascinated the architects of the Renaissance by its symbolic and traditional values, and it is found in their drawings and treatises to the virtual exclusion of the more practical longitudinal basilicas that the architects were often commissioned to build.

Plan symbolism remained almost exclusively in the sphere of religion after antiquity, and its traditions gradually disappeared in the course of the 19th century. The modern plan is determined by problems of form (space–mass relationships, etc.) and by the practical demands of use rather than by symbolic communication.

Symbolic forms in elevation

In elevation the most consistent symbolic forms have been the dome, the tower, the stairway, the portal, and the colonnade. Domes imply the meanings of the circle and more, since a dome is a covering. Long before masonry domes could be built, the hemisphere was associated with the heavens as a "cosmic canopy," and throughout history domes have been decorated with stars and astrological symbols.

In ancient Rome and among Christians and Indian Buddhists the dome came to mean universal power. During the Renaissance it spread from religious structures to palaces and government buildings, retaining some of its implications of power. In the United States the national capitol is domed and there are few state capitols without



West facade with dome of cast-iron construction, U.S. Capitol building, Washington, D.C., dome begun by Thomas Ustick Walter, 1851.
Wayne Andrews

domes; the symbol has survived the loss of its original meanings.

The tower, with origins in primitive nature rites, has consistently symbolized power. The Chinese pagoda extends central-plan symbolism into towers; many towers and spires rose from the northern European Gothic cathedral; and the medieval Italian city was a forest of towers erected by nobles in constant competition to express their supremacy. This meaning survives in modern skyscrapers; their height is more frequently boasted of than their efficiency or beauty.

Erich Muller



Baroque staircase, Residenz (1719–44), Würzburg, West Germany, by Balthasar Neumann.

Functional architectural elements as symbols

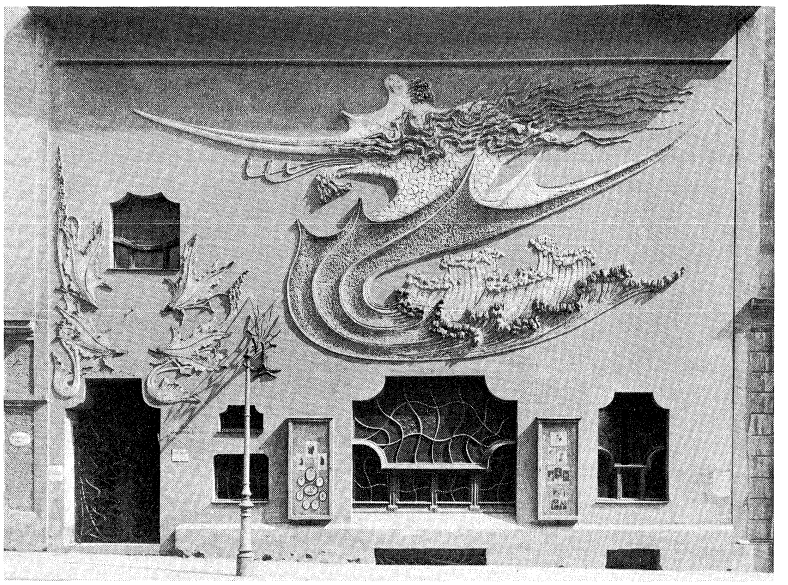
Architectural elements conceived to facilitate the use of buildings may also take on symbolic significance. The stairway, employed in the past to give "monumentality" to important buildings, frequently became more expressive than convenient, especially in Baroque palaces. Portals, from the time of ancient Egyptian temple pylons and Babylonian city gates, became monuments in themselves, used to communicate a heightened significance to what lay behind them. In the Gothic cathedral they became the richest element of the facade—a translation of biblical doctrine into stone. Since the classic development of the ancient Greek temple, the colonnade on the exterior of buildings has borne similar implications.

Such symbols have become archaic in modern culture and appear as a sign of resistance to new forms. This resistance is especially evident in the popular symbolism of domestic architecture, where the atmosphere of the home is often expressed by cottage-like roofs, shutters, trellises, mullioned windows, grilles, and other associations with a more peaceful past.

Decoration, the most easily recognized medium of content, communicates meaning either through architectural elements or through the figural arts (sculpture, painting, mosaic, stained glass, etc.). The architectural elements used decoratively, such as the classical orders, usually originate in technique and in time lose their structural significance to become symbols. In Rome and from the Renaissance to the 20th century, the formal Grecian orders were applied to buildings of many different techniques as expressions of the continuing influence of Greek institutions. Similarly, the new vocabulary of Gothic architecture, developed with new building techniques (the pointed arch, the flying buttress, etc.), became in later periods a source for religious and romantic symbolism. The Art Nouveau of the turn of the 20th century, a system of ornament based on floral and other organic forms, survived for only two decades, perhaps because its symbols were neither drawn from a tradition nor derived from a structural system. (The role of architectural ornament will be further treated below).

The role of the figural arts

The function of the figural arts in conveying content is a subject outside the scope of this article, but its importance for architecture must be mentioned. The figural arts not only offer the means of expressing more specific ideas



The Studio Elvira, 1897–98, Munich, by August Endell; an example of Art Nouveau ornament.
Foto Marburg

than any architectural symbols, but in many architectural styles they define the character of mass and space. The sculptures of the Hindu temple, the mosaics of the Byzantine church, and the stuccoes of Moorish palaces are not ornamental applications; they determine the form of the building itself.

The virtual absence of traditional symbols in modern architecture is evidence of the failure of these symbols to express the cultural patterns of the 20th century. In these times, architecture, like painting, sculpture, and other arts, has tended to be abstract, to emphasize qualities of form rather than the communication of familiar ideas through symbols.

Expression of technique. The second aspect of content is the communication of the structural significance of materials and methods. Its purpose is to interpret the way in which architecture is put together. The characteristics of materials that are imporant in expressing design techniques are the properties of their composition (*e.g.*, structure, weight, durability) and the way they are used in structure. Their properties may be expressed and interpreted by the treatment of the surface, and their use may be expressed by emphasis on the dimensions and joining of the building units into which they are formed.

The hardness, weight, and crystalline composition of stone masonry traditionally have been emphasized by devices not necessarily connected with structural methods: rustication (finishing in rough, uneven surfaces), drafting (more refined, linear cutting), and polishing. Niches and other indentations, projecting courses, or frames around openings suggest massiveness. In nonbearing walls, a smooth, unbroken surface implies thinness. The use of stone or brick masonry in construction is emphasized by clarifying the limits of each block and by the amount of mortar used and by distinguishing lintels, arches, and other specific members from the construction of the wall. The properties of wood are suggested by revealing and emphasizing its texture in load-bearing members and by treating the sheathing of light wood frames in patterns (of shingling or boarding) that communicate thinness. The plasticity of concrete is shown by freedom in modelling and its use in construction by emphasizing the impressions of the wooden forms in which it is cast. The sections of light metal curtain walls are frequently stamped into geometric patterns to illustrate their nonbearing character. Materials that must be covered for protection, such as unfired brick and the steel used in framing, are not adaptable to this type of communication.

At times building methods are demonstrated simply by exposing the structure, as in the heavy timber frame, but in many styles the functions of structural systems have been interpreted by designing their members in forms that often are more explanatory than efficient. The Greek column, which is narrower at its summit than at its base, is diminished by a curve beginning slightly below the midpoint, giving it an effect of an almost muscular power to resist loads. The expression is more explicit in the caryatid, a human figure that replaces the column, and in the burdened animals and dwarfs that support the columns of Romanesque portals. Many elements in the Gothic cathedral serve as diagrams of structure: the supporting piers are clusters of shafts, each of which extends upward without interruption to become the rib of the vault, and the ribs themselves are an elucidation of technique; the flying buttress and the window tracery are elegant interpretations of their functions. In the modern steel-frame building, the hidden forms of the skeleton are often repeated on the facade to enable one to "see through" to the technique, but the system also permits the alternative of expressing the lightness and independence of the curtain wall by sheer surfaces of glass and other materials. The work of the concrete slab is made explicit by projecting indications of the placement of reinforcement or of the distribution of stresses.

Interpretive expression of structural method

The expression of technique is characteristic not of all architectural styles but only of those such as the Gothic and modern, in which new techniques excite a search for the interpretive design of their materials and methods. More often than not, both materials and methods have been disguised by decorative forms or surfacing such as veneers, stucco, or paint, because of emphasis on the expression of content or of form. Most early stone architecture in Egypt, Greece, and India retained as decoration the forms developed in wooden forerunners. The precious marble of Greek temples was disguised under painted stucco; Roman brickwork was hidden by slabs of coloured marble; and 19th-century cast-iron columns were molded into classic or Gothic forms. The history of domes is filled with examples of the successful disguising of method, of giving the ponderous mass the effect of rising from the exterior and of floating from within.

Disguised methods and materials

Technical content has been one of the foundation stones of 20th-century architectural theory, particularly in its early phases, and has represented a reaction against 19th-century symbolic content. It is essential for the understanding of modern architecture that the expression of technique be seen as an art—a creative interpretation that heightens awareness of the nature of architecture.

FORM

In the sphere of function and technique, the architect is responsible to the patterns of his culture on one hand and

to the patterns of technology on the other; but, in the expression of form, he is free to communicate his own personality and concepts. Not every architect has the gift to exercise this prerogative to the fullest. As in other arts and sciences, a few individuals generate new styles, and others follow, interpreting these styles in original and personal ways. But the majority accepts styles as given and perpetuates them without leaving its mark. The architect's principal responsibility in the formation of style is to create meaningful form. When form is spoken of in the arts, not only the physical shape, size, and mass of a work are meant but also all the elements that contribute to the work's aesthetic structure and composition. Many of these may be without a fixed form of their own—a rest in music, a line in painting, a space in architecture—and gain significance only as they are organized into the finished product. The basic formal elements of architecture in this sense are space and mass. The process of organizing these elements into an ordered form is called composition, and the principal means by which they are given expressive quality are scale, light, texture, and colour.

Space and mass. Space, that immaterial essence that the painter suggests and the sculptor fills, the architect envelops, creating a wholly human and finite environment within the infinite environment of nature. The concept that space can have a quality other than emptiness is difficult to grasp. When a building is entered, floor, supports, walls, and a ceiling are seen, all of which can be studied and perhaps enjoyed, while the space, in the sense that one is accustomed to think of it, is void: the absence of mass, filled by air.

Experiencing space

But spatial experiences that express something are common to everyone, though they are not always consciously grasped. One feels insecure in a low cave or a narrow defile, exhilarated and powerful on a hilltop; these are psychological and motor reactions that result from measuring one's potential for movement against the surrounding spaces, and the same reactions take root even in language ("confining" circumstances and "elevating" experiences are spoken of). An infinite variety of such reactions may be summoned by the architect, because he controls the limits above, below, and on all sides of the observer. As a person enters the architect's space he measures it in terms of the degree and the quality of his potential for movement. The concept of potentiality is important, first, because the observer can anticipate where he may move merely looking about him and, second, because he can conceive movements that he cannot execute. Thus, in the nave of a Gothic cathedral the high walls closely confining the observer on two sides restrict his possible movements, suggesting advance along the free space of the nave toward the altar; or their compression forces him to look upward to the vaults and the light far overhead, there to feel a sense of physical release, though he is earthbound. The experience of Gothic space is called uplifting because it urges one to rise.

Renaissance space, on the other hand, attempts to balance its suggestion of movement, to draw the observer to a focal point at which he can sense an equilibrium of movement in all directions, a resolution of the conflict of compression and release. At this point one feels physically at rest, at the opposite extreme from the elevating sensation of the cathedral.

Of course, one does not use his eyes alone to feel spatial quality, because only the simplest spaces—a cubic room, for example—can be wholly experienced from one standpoint. In a complex of spaces, such as that of the cathedral, the observer walks about, gaining new sensations, seeing new potentials for movement at every step. Most modern architecture, in its free organization of space sequences, demands mobility; its techniques have made it possible to remove the heavy walls and supports of the past, reducing the sense of compression. Walls become membranes to be arranged at will for spatial experience, and some are transparent and so extend one's potential for movement into the limitless out-of-doors.

Spatial experience is not restricted to the interiors of buildings. The sensations one has in nature's open spaces may be recreated by art. City squares and streets, even gardens, achieve a variety of expression comparable with that of interiors. The Baroque piazza of St. Peter's in Rome, which directs the observer along its great embracing arcs toward the entrance, is at least as moving as the church interior.



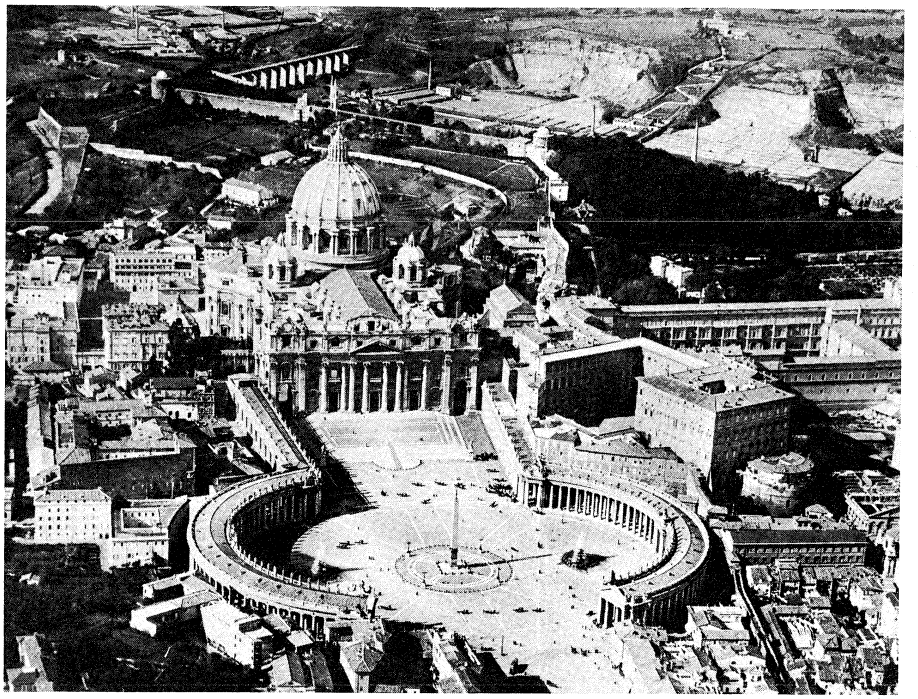
Nave toward the apse, S. Giorgio Maggiore, Venice, designed by Andrea Palladio, 1566; an example of Renaissance space.
Osvaldo Bohm—EB Inc.

Appreciating mass

The exterior of a single building, particularly one that is isolated from other architecture, does not create a space. It occupies the space of nature. Thus, it may be experienced as sculpture, in terms of the play of masses in a void. The aesthetics of masses, like that of spaces, is rooted in one's psychology. When a tall tree or a mountain is called majestic and a rocky cliff menacing, human attributes are being projected. Man inevitably humanizes inert matter and so gives the architect the opportunity to arouse predictable patterns of experience.

The appreciation of mass, like that of space, depends on movement, but this movement must be physical. It cannot be experienced in anticipation, because, no matter where one stands to observe even the simplest building, part of it is out of sight. The mass of a complex building is differently composed from every point of view. The 20th-century art critic Sigfried Giedion, emphasizing the need for movement in experiencing modern architecture, suggested that architecture may be four-dimensional, since time (for movement) is as meaningful as the spatial dimensions.

Some architecture depends much more on mass expression than on space expression. The Egyptian pyramid, the Indian stupa, and the dagoba of Sri Lanka have no meaningful interior spaces; they are architectural in function and technique, sculptural in expression. The interior of a Greek temple is of little interest compared with the wonderful play of forms on its colonnaded exterior, while early Christian and Byzantine architecture reverse the emphasis, making the simple exterior a shell for a splendid and mystical space. Gothic architecture



St. Peter's, Vatican City, colonnade and piazza designed by Gian Lorenzo Bernini, begun 1656.
Alinari .

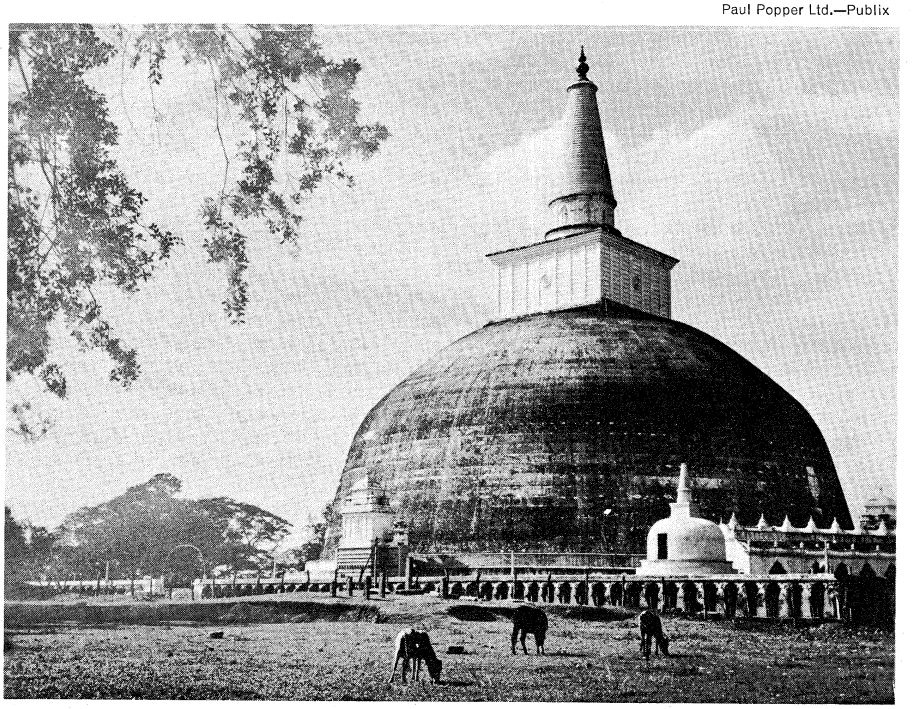
balances the two, partly in order to express a dual content: earthly power over the world outside, spiritual power inside. Modern techniques permit a reduction of the contrast between space and mass expressions by reducing the mass of walls and the size and number of supports and by allowing the interpenetration of interior and exterior space.

Composition. Space and mass are the raw materials of architectural form; from them the architect creates an ordered expression through the process of composition. Composition is the organization of the whole out of its parts—the conception of single elements, the interrelating of these elements, and the relating of them to the total form.

The simplest architectural element is a plane, the flat, two-dimensional surface that limits masses and spaces. The simplest plane is a rectangular one without openings or decoration—the wall of a room, for example. This wall is given quality solely by the proportion of its width to its height. Now a door is put into the wall; the door itself has a certain proportion, and a third element is injected, the relation of two proportions. A window is added, and the composition becomes more complex; then a row of windows, and sequence becomes a factor in addition to the elements of proportion and relation. Sequence again involves the concept of motion; the row of windows is said "to run along" the facade or is "rhythmically" designed.

Finally, this wall may gain rich subtleties of composition within its proportions and rhythms. It can be mod-

Paul Popper Ltd.—Publix



The Ruanveli dagoba at Anurādhapura, Sri Lanka, 2nd century.

elled—into a complex of planes or irregular or curved surfaces—to provide the dimension of depth to its proportions; or symbols of use or of technique can become part of its expressive form.

No architectural planes stand alone, of course; they always intersect other planes. The room wall meets two other walls, the floor, and the ceiling, and a facade wall meets the ground, the roof, and two other walls. So the total composition of a wall must be harmonized with the composition of other planes in a three-dimensional whole.

Systems of proportions

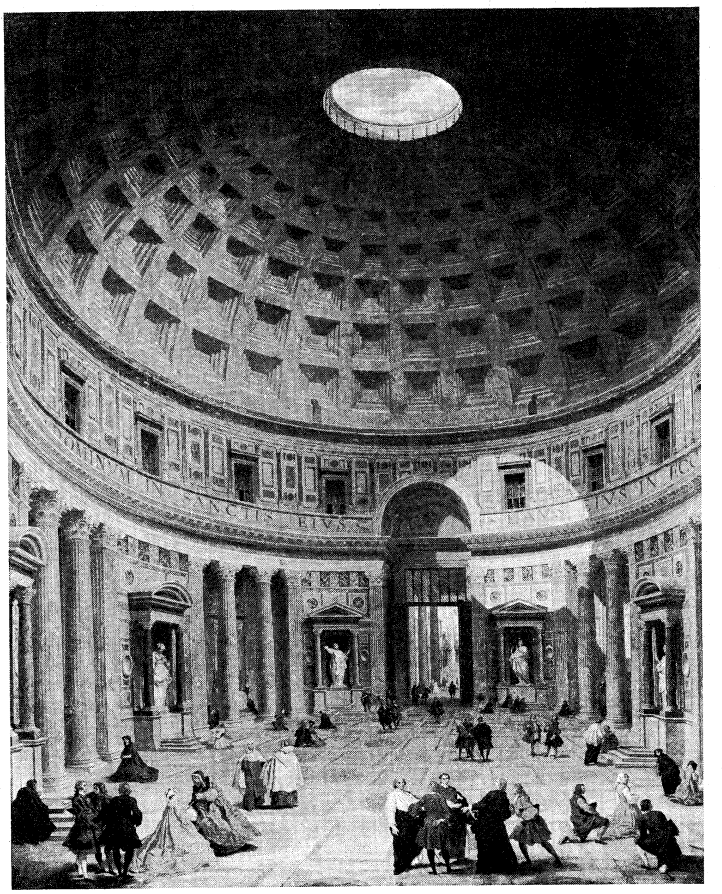
The means of achieving this harmony differ in every style. Greek architects developed a system of proportions based on the lower diameter of the temple column, from which spatial intervals and the measurement of masses were derived by multiplication and division. Medieval architects first used arithmetical modules based on the measurements of areas in the cathedral plan and, in the Gothic period, changed to a geometric system that employed chiefly the equilateral triangle and the square, figures that had symbolic and mystical values. In Renaissance theory, proportions and harmonies were developed from systems of musical composition, since architects believed that relationships in all the arts depended on an all-pervading celestial harmony. Several modular and proportional systems have been evolved by modern architects (*e.g.*, Le Corbusier's "Modulor"), but none has been widely adopted.

Behind these changing theoretical methods, however, there seems to be a constant human reaction to spatial relationships that distinguishes harmony from cacophony, that makes one bored with a perfectly cubic room or prefer certain rectangular forms to others. This psychological response to form probably is connected to one's mechanisms of balance, movements, and stature—in short, to one's own composition—but the scientific analysis of the process is still at an early stage of development.

Some buildings have only a single, simple interior space (the Pantheon in Rome) or exterior mass (the pyramids of Egypt) and are not less expressive on this account. But composition carries on into a richer dimension as soon as two or more spaces or masses are organized into the whole. Such a complex composition must give a coordinated form to connecting spaces and masses, each of them in itself a unique harmony. The observer must be made to feel, in moving through the spaces and around the masses, not only that each is related to the one that precedes and follows it but that each one is contributing to a concept of the whole: a form that is greater than the sum of its parts. In the Gothic cathedral, the nave, the aisles alongside it, the transept that crosses it toward one end, the choir, and apse beyond may each be experienced separately for its own quality. But the experience gains its full meaning only when the form of the total expression is realized: the low aisles giving grandeur to the high nave, the three together leading to the confluence of the two transept arms at the crossing in a vast climax that prepares for the resolution or finale at the altar. In the same way, the significance of a total mass composition unfolds as one moves about its separate parts. At St. Peter's in Rome, the three projecting apses are gathered into a unity by the undulating walls; they prepare for the cylindrical drum, the drum for the dome, the form of which leads to the culminating lantern, which is harmonized with the drum. Toward the facade, two little domes frame and prophesy the great one, as the cathedral aisles do the nave. While these particular examples from the past illustrate symmetrical compositions with a climax, other buildings that are of equal quality might be chosen to show irregular unity that is no less expressive (*e.g.*, the Erechtheum in Athens; the abbey of Mont Saint-Michel in France).

Perception of both the whole and its parts

In modern architecture, as in modern painting, Renaissance laws of composition, which emphasized the symmetry and balance of semi-independent units, have been supplanted by principles that imply the continuity of the whole and remove distinctions between parts. The biological term organic is sometimes used to describe a process of composition that seeks to develop interdependent spaces and masses that function expressively as members of an organism.



"The Interior of the Pantheon," oil painting by Gian Paolo Pannini (1691/92–1765). In the National Gallery of Art, Washington, D.C.

By courtesy of the National Gallery of Art, Washington, D.C., Samuel H. Kress Collection

If composition were merely a matter of organizing a certain number of relationships, the process would be mechanical, not creative, and all architecture would be equally good or, more likely, bad. The purpose of composition is to express particular concepts and experiences, and it is successful only when these are fully communicated to the observer.

Scale. When the proportions of architectural composition are applied to a particular building, the two-termed relationship of the parts to the whole must be harmonized with a third term—the observer. He not only sees the proportions of a door and their relationship to those of a wall (as he would in a drawing of the building), but he measures them against his own dimensions. This three-termed relationship is called scale.

A well-scaled building such as a Greek temple will serve for illustration. If it were to be magnified to the size of St. Peter's in Rome, with its proportions remaining unchanged in their own relationships, the temple would be out of scale, and the result would appear monstrous. If the columns were to be doubled in width while the temple remained the same size, they would be out of scale and out of proportion with the whole. The proportions of the temple are satisfactory as they are because they are based on certain aesthetic principles established by the Greeks, principles that are partly rooted in human psychological makeup and partly accepted by custom (*e.g.*, as are musical consonances). It is difficult to understand, however, why the scale of these temples is so successful within a certain range of size, for neither the ancient Greeks nor anyone else established laws to relate scale to size. They found their solution by experiment and subjective judgment.

It may be that the success of scale depends upon man's

Comprehension of proportions

ability to comprehend proportions in relation to some unit or module that is roughly human sized and close enough to a person in a building to permit him to measure it against himself. The Greeks, in employing the base of the column as a module for all the proportions of a temple, found a unit of a size that can be grasped easily and one that is close to eye level as a person approaches the temple. This module is a key to relationships among elements too far away to measure. This can be done in much larger buildings, too, where the elements close to the observer are too massive to be measured easily. Roman and Renaissance architecture retained the ancient Greek orders as decoration partly for this reason, using them to break up huge masses into more comprehensible parts. In entirely different styles of architecture, such as the Gothic, where the expressive function requires immeasurable proportions, there is still a measurable module given in the base of the pier. But piers and columns are not always a source of the module. In masonry construction, the single block can serve the same purpose. In frame construction, the bay (distance between floors or columns) or doors and windows may make a better key. The most successful modern skyscrapers retain a comprehensible scale, in spite of their size, by the repetition of some such module, and this is one reason why the skeleton is so often expressed on the exterior even when it is hidden behind walls.

Light. Light is a necessity for sight and, in architecture, a utility. But light is also a powerful, though ephemeral, vehicle of expression. Because it moves, changes character, and comes and goes with its source, light has the power to give to the inert mass of architecture the living quality of nature. The architect, though he does not quite control it, can predict its behaviour well enough to catch its movements meaningfully. He channels it through openings into his spaces and molds it on the surfaces of his masses by changes of plane, making it enliven his forms by contrast with shadow (see LIGHTING AND LIGHTING DEVICES).

Attempts to control light for expressive effect

The sunlight that falls on the exteriors of buildings cannot be directed or changed in quality, but it can be reflected or absorbed in a wide range of modulation by the relief and texture of surfaces. The planes and decoration of a facade, therefore, are not just the lines the architect makes on his working drawings but receptacles of light and shadow that change in character, even in form, as the Earth moves about the Sun.

Because of this link between nature and art, an important part in the formation of local architectural styles is played by the variation in the quality and intensity of light in different climatic regions.

The architect controls interior light better than exterior light, since he can select the position, size, and shape of its source. With glass and other transparent materials he transforms even its colour and intensity and so gives light a meaning independent of that which it imparts to the structure. One realizes this most powerfully in the Gothic cathedral, where the stained-glass windows transform the rays of the sun into a mystical diffusion that descends from above like a supernatural vision.

Furthermore, light may be illusory, dissolving rather than clarifying form. When it comes out of darkness in great intensity it seems to spread outward from its architectural channel. This illusion may be employed to express meanings, as at Hagia Sophia in Istanbul, where the light from the base of the dome hides the supports, giving the impression that the canopy floats on air.

Texture. Texture plays a dual role in architecture: it expresses something of the quality of materials, and it gives a particular quality to light. Although one absorbs both qualities simultaneously by eye, the first has tactile, the second visual associations.

Specific tactile textures are peculiar to every material by virtue of its manufacture or natural composition, but they may be altered to produce a variety of expressive qualities. Any stone may be used in its natural, irregular state, or it may be chiselled in a rough or smooth texture or highly polished to convey a range of meanings from vigour to refinement.

Visual textures are produced by the patterns given to the lighting of the surface both through the way the materials are worked (*e.g.*, vertical or horizontal chiselling of stone) and through the way they are employed in building (*e.g.*, vertical or horizontal boarding, projection and recession of courses of brick). Like all patterns, visual textures create associations of movement, giving rhythm to the surface.

A single texture is rarely employed in building. The variety of materials and treatments typically produces a complex of textures that must be composed and harmonized like the forms and spaces of architecture into a consistent expressive whole.

Colour. Since colour is a characteristic of all building materials, it is a constant feature of architecture. But building materials are selected primarily for their structural value, and their colours are not always suited to expressive requirements; thus, other materials chosen for their colour are frequently added to the surface. These include pigments, which usually preserve the texture of the original surface, and veneers of stone, wood, and a variety of manufactured products that entirely alter the surface character.

But colour, regardless of how it is produced, is the most impermanent element in architecture. It changes with the weathering and staining of materials (the white Gothic cathedrals are now deep gray), or, if it is superficial, it can easily be altered or removed (as the coloured stucco veneers of ancient Greek temples or the bright marble facing on Roman brickwork).

The values that are associated with colour (yellow and red, for instance, are called "gay," black and deep blue "sombre") are independent of materials and forms, and they give the architect a range of expression not provided by other means at his disposal. A different expressive device is provided by the great range of light reflection in the colour scale. Colours that reflect light brilliantly appear to advance toward the viewer, and those that absorb light appear to recede; the degree of projection and recession of architectural forms may be altered, emphasized, or subdued by the colours of their surfaces.

Environment. Architecture, unlike most of the other arts, is not often conceived independently of particular surroundings. The problems of design extend beyond the organizing of space and mass complexes to include the relating of the total form to its natural and architectural environment.

In site planning, a primary function of architectural design, the architect aims to create harmonies with preexisting elements in the landscape and "townscape."

But the province of the architect is not limited to the conception of single structures in harmony with a given setting. Throughout history, architects have been employed in giving a new form to the environment itself: planning the natural surroundings by the design of parks, roadways, waterways, etc.; designing complexes of related buildings; and organizing the urban environment into areas of residence, recreation, assembly, commerce, etc., both to increase their utility and to give them unique expressive qualities through the interrelationship of groups of buildings to the open areas about them.

(J.S.Ac.)

ORNAMENT

Although it would be difficult to cover in any single definition all conceptions, past and present, of what constitutes ornament in architecture, three basic and fairly distinct categories may be recognized: mimetic, or imitative, ornament, the forms of which have certain definite meanings or symbolic significance; applied ornament, intended to add beauty to a structure but extrinsic to it; and organic ornament, inherent in the building's function or materials.

Mimetic ornament. Although it is still found in the 20th century, mimetic ornament is by far the commonest type of architectural ornament in primitive cultures, in Asian civilizations, and generally throughout antiquity. It grows out of what seems to be a universal human reaction to technological change: the tendency to repro-

duce in new materials and techniques shapes and qualities familiar from past usage, regardless of appropriateness. This tendency may be called the principle of mimesis. Most common building types in antiquity, both East and West (*e.g.*, tombs, pyramids, temples, towers), began as imitations of primeval house and shrine forms. An obvious example is the dome, which developed as a permanent wooden or stone reproduction of a revered form originally built of pliable materials. In the mature stages of early civilizations, building types tended to evolve beyond primitive prototypes; their ornament, however, usually remained based on such models. Decorative motifs derived from earlier structural and symbolic forms are innumerable and universal. In developed Indian and Chinese architecture, domical and other originally structural forms occur often and lavishly as ornament. In ancient Egypt, architectural details continued to preserve faithfully the appearance of bundled papyrus shafts and similar early building forms. In ancient Mesopotamia, brick walls long imitated the effect of primitive mud-and-reed construction. In the carved-stone details of the Greco-Roman orders (*e.g.*, capitals, entablatures, moldings), the precedent of archaic construction in wood was always clearly discernible.

Traditional nature of mimetic ornament

The prevalence of mimetic ornament in architecture may be explained in two ways. Some (perhaps most in primitive cultures) is religious in origin. Certain forms and shapes, through long association with religious rites, became sacred and were preserved and reproduced for their symbolic value. These forms continued to be understood even though they were often stylized into abstract or geometric patterns, unrecognizably removed from their naturalistic models. Much mimetic ornament, however, even in early times, can be ascribed simply to inertia or conservatism. People have generally tended to resist change; they find it reassuring to be surrounded by known and familiar forms. Reproducing them as ornament on newly introduced forms is a common reaction to the vague feeling of uneasiness that rapid social and technological change induces; it provides a satisfying sense of continuity between the past and the present. This resistance was a factor in the 19th- and early-20th-century practice of disguising new techniques of construction in metal and glass by an overload of ornament imitating earlier styles.

Applied ornament. Architectural ornament in the 19th century exemplified the common tendency for mimetic ornament, in all times and places, to turn into mere applied decoration, lacking either symbolic meaning or reference to the structure on which it is placed. By the 5th century BC in Greece, the details of the orders had largely lost whatever conscious symbolic or structural significance they may have had; they became simply decorative elements extrinsic to the structure. The Doric frieze is a good case: its origin (*i.e.*, an imitation of the effect of alternating beam ends and shuttered openings in archaic wood construction) remained evident, but it came to be treated as a decorative sheath without reference to the actual structural forms behind. In losing their mimetic character, the details of the Greek orders acquired a new function; they served to articulate or unify the building visually, organizing it into a series of coordinated visual units that could be comprehended as an integrated whole, rather than as a collection of isolated units. This concept of applied decoration was passed on through the Greco-Roman period. The triumphal arch of Rome, with its system of decorative columns and entablature articulating what is essentially one massive shape, is a particularly good illustration; the Colosseum is another. Most of the great architecture of the Renaissance and Baroque periods depends on it; to a large extent, the difference between these styles is the difference in decoration. The characteristic serenity and balance of Filippo Brunelleschi's architecture in the 15th century, for example, is very largely effected by his treatment of pilasters (rectangular ornamental columns with bases and capitals) and entablatures applied to them, whereas, in 16th-century wall-surface designs such as Michelangelo's Medici chapel or the dome of St. Peter's, the same elements are used in different combinations to create a quite opposite effect of tension and release.

Function as an element to achieve organization

Judicious and intelligent use of applied ornament remained characteristic of most Western architecture until the 19th century, when the rationale of applied ornament frequently broke down, and an often indiscriminate and inappropriate use of decoration became characteristic. The reasons for this development are complex. In part it was a reaction to an overly rapid pace of social change during the period; partly, also, it was a logical outgrowth of the increasingly lavish decoration of late Baroque and Rococo architecture in the 18th century. Also, there was an overemphasis on the purely literary and associative values attached to the ornament characteristic of historical architectural styles. But compounding all these factors was the development of machinery, such as multiple lathes and jigs, which provided builders with cheap prefabricated ornament to give their often shoddy and ill-proportioned structures an illusion of elegance. Architectural ornament and architectural forms proper tended to part company and to be designed quite independently of each other.

Organic ornament. By the early 20th century a preoccupation with the proper function of architectural ornament was characteristic of all advanced architectural thinkers; by the mid-20th century a concept of architectural ornament had been formulated that has been called organic ornament. This concept, however, is by no means peculiar to the 20th century. Its essential principle is that ornament in architecture should derive directly from and be a function of the nature of the building and the materials used. This principle is characteristic of both Christian and Islāmic religious architecture of the medieval period. In the architectural ornament of Muslim India or Persia, as in early Christian and Byzantine work, there is a strong mimetic element. The proscription of representational forms in the Qur'ān and the tendency of both Muslim and early Christian artists to borrow and adapt their formal vocabulary from preceding cultures led inevitably to their transforming what had been meaningful forms into systems of abstract ornament. But basically this ornament was neither mimetic nor applied. Throughout the Middle Ages, church buildings were conceived primarily as tangible symbols of heaven. Their architectural ornament, no matter how various or lavish, was consistently designed to promote this symbolism; whether by gilt, intricacy, or multiplicity, it all contributed to an overall effect of glory and so was integral to the architectural form.

Relation of ornament to nature, function, and structure

Twentieth-century concepts of the function of architectural ornament, generally speaking, began with an understanding of this medieval usage that grew out of the 19th-century writings of the English art critic John Ruskin and the French Gothic Revival architect Viollet-le-Duc, as well as through the interpretations and applications of the British designer William Morris. The immediate influence of these men proved rather unfortunate. The first result of Viollet-le-Duc's disciplined and scholarly investigations into the principles of medieval architecture was a school of slick archaeological architects, capable of decorating all manner of collegiate, civic, and domestic buildings with frigidly correct reproductions of the details of medieval cathedrals and châteaus. Out of Ruskin's demonstration of the origins of medieval decoration in natural forms there grew the so-called Art Nouveau movement toward exaggerated floral and curvilinear ornament; and out of Morris' insistence on handicrafts, inspired by infatuation with the medieval guild system, developed the Arts and Crafts movement.

As early as the 1870s the U.S. architect H.H. Richardson adopted the Romanesque style, less for its historical associations than for the opportunities it afforded him to express the nature and texture of stone. In mature examples of his architecture from the mid-1880s, ornament in the older, applied sense had virtually disappeared, and his buildings depend for their aesthetic effect mainly on the inherent qualities of their materials. The generation following Richardson saw a further international development of this principle.

In Great Britain Sir Edwin Lutyens and Charles Rennie Mackintosh, in The Netherlands Hendrick Petrus Berlage, and in the United States Louis Sullivan were among many architects who contributed to the new ornamental expression. It was largely based on intrinsic texture and pattern but with interspersed bands and patches of naturalistic ornament, applied with studied discipline. With the general reaction against 19th-century eclectic principles of ornamentation after World War I, however, leading designers rejected even this kind of applied ornament and relied for ornamental effect on building materials alone. The so-called International Style, in which the German architect Walter Gropius and the Swiss-French architect Le Corbusier were the chief figures, dominated advanced design during the late 1920s and 1930s. The barrenness that resulted from their reliance on such materials as concrete and glass, however, along with other factors, resulted in a reaction in the 1940s in favour of the neglected precedent set by the U.S. architect Frank Lloyd Wright in his early-20th-century work, which emphasized more visually interesting materials, intricate textural patterns, and natural settings as the proper basis of architectural ornament. This trend continued into the 1960s and 1970s; the New Brutalism was related to it. (Al.Go.)

20th-century usage of ornament

IV. Theory of architecture

The term theory of architecture was originally simply the accepted translation of the Latin term *ratiocinatio* as used by Vitruvius, a Roman architect-engineer of the 1st century AD, to differentiate intellectual from practical knowledge in architectural education; but it has come to signify the total basis for judging the merits of buildings or building projects. Such reasoned judgments are an essential part of the architectural creative process. A building can be designed only by a continuous creative, intellectual dialectic between imagination and reason in the mind of each creator.

A variety of interpretations has been given to the term architectural theory by those who have written or spoken on the topic in the past. Before 1750 every comprehensive treatise or published lecture course on architecture could appropriately be described as a textbook on architectural theory. But, after the changes associated with the Industrial Revolution, the amount of architectural knowledge that could be acquired only by academic study increased to the point where a complete synthesis became virtually impossible in a single volume.

The historical evolution of architectural theory is assessable mainly from manuscripts and published treatises, from critical essays and commentaries, and from the surviving buildings of every epoch. It is thus in no way a type of historical study that can reflect accurately the spirit of each age and in this respect is similar to the history of philosophy itself. Some architectural treatises were intended to publicize novel concepts rather than to state widely accepted ideals. The most idiosyncratic theories could (and often did) exert a wide and sometimes beneficial influence; but the value of these influences is not necessarily related to the extent of this acceptance.

Analysis of historical buildings

The analysis of surviving buildings provides guidance that requires great caution, since, apart from the impossibility of determining whether or not any particular group of buildings (intact or in ruins) constitutes a reliable sample of the era, any such analyses will usually depend on preliminary evaluations of merit and will be useless unless the extent to which the function, the structure, and the detailing envisaged by the original builders can be correctly re-established. Many erudite studies of antique theories are misleading because they rest on the assumption that the original character and appearance of fragmentary ancient Greek and Hellenistic architectural environments can be adequately deduced from verbal or graphic "reconstructions." Even when buildings constructed before 1500 remain intact, the many textbooks dealing with antique and medieval theories of architecture seldom make qualitative distinctions and generally imply that all surviving antique and medieval buildings were good, if not absolutely perfect.

Nevertheless, the study of the history of architectural philosophy, like that of the history of general philosophy, not only teaches what past generations thought but can help the individual decide how he himself should act and judge. For those desirous of establishing a viable theory of architecture for their own era, it is generally agreed that great stimulus can be found in studying historical evidence and in speculating on the ideals and achievements of those who created this evidence.

DISTINCTION BETWEEN THE HISTORY AND THEORY OF ARCHITECTURE

The distinction between the history and theory of architecture did not emerge until the mid-18th century; indeed, the establishment of two separate academic disciplines was not even nominal until 1818, when separate professorships with these titles were established at the École des Beaux-Arts in Paris. Even then, however, the distinction was seldom scrupulously maintained by either specialist. It is impossible to discuss meaningfully the buildings of the immediate past without discussing the ideals of those who built them, just as it is impossible to discuss the ideals of bygone architects without reference to the structures they designed. Nevertheless, since any two disciplines that are inseparably complementary can at the same time be logically distinguishable, it may be asserted that this particular distinction first became manifest in *Les Ruines des plus beaux monuments de la Grèce* ("The Ruins of the Most Beautiful Monuments of Greece"), written in 1758 by a French architecture student, Julien-David LeRoy. Faced with the problem of discussing Athenian buildings constructed in the time of Vitruvius, he decided to discuss them twice, by treating them separately under two different headings. Before this date, "history" was of architectural importance only as a means of justifying, by reference to classical mythology, the use of certain otherwise irrational elements, such as caryatids. Even Jacques-François Blondel, who in 1750 was probably the first architectural teacher to devote a separate section of his lecture courses to "history," envisaged the subject mainly as an account of the literary references to architecture found in antique manuscripts —an attitude already developed by the 15th-century Renaissance architect Leon Battista Alberti.

Emergence of two disciplines in the 18th century

The modern concept of architectural history was in fact simply part of a larger trend stimulated by the leading writers of the French Enlightenment, an 18th-century intellectual movement that developed from interrelated conceptions of reason, nature, and man. As a result of discussing constitutional law in terms of its evolution, every branch of knowledge (especially the natural and social sciences) was eventually seen as a historical sequence. In the philosophy of architecture, as in all other kinds of philosophy, the introduction of the historical method not only facilitated the teaching of these subjects but also militated against the elaboration of theoretical speculation. Just as those charged with the responsibility of lecturing on ethics found it very much easier to lecture on the history of ethics, rather than to discuss how a person should or should not act in specific contemporary circumstances, so those who lectured on architectural theory found it easier to recite detailed accounts of what had been done in the past, rather than to recommend practical methods of dealing with current problems.

The modern concept of architectural history

Moreover, the system of the Paris École des Beaux-Arts (which provided virtually the only organized system of architectural education at the beginning of the 19th century) was radically different from that of the prerevolutionary Académie Royale d'Architecture. Quatremère de Quincy, an Italophile archaeologist who had been trained as a sculptor, united the school of architecture with that of painting and sculpture to form a single organization, so that, although architectural students were ultimately given their own professor of theory, the whole theoretical background of their studies was assimilated to the other two fine arts by lecture courses and textbooks such as Hippolyte Taine's *Philosophie de l'art*, Charles Blanc's *Grammaire des arts du dessin*, and Eugène Guillaume's *Essais sur la théorie du dessin*.

Similarly, whereas before 1750 the uniformity of doctrine (the basic premises of which were ostensibly unchanged since the Renaissance) allowed the professor of architecture to discuss antique and 16th-century buildings as examples of architectural theory and to ignore medieval buildings completely, the mid-19th-century controversy between "medievalists" and "classicists" (the "Battle of the Styles") and the ensuing faith in Eclecticism turned the studies of architectural history into courses on archaeology.

Importance of visual analysis in the 20th century

Thus, the attitudes of those scholars who, during the 19th and early 20th centuries, wished to expound a theory of architecture that was neither a philosophy of art nor a history of architecture tended to become highly personal, if not idiosyncratic. By 1950 most theoretical writings concentrated almost exclusively on visual aspects of architecture, thereby identifying the theory of architecture with what, before 1750, would have been regarded as simply that aspect that Vitruvius called *venustas* (*i.e.*, "beauty"). This approach did not necessarily invalidate the conclusions reached; but many valuable ideas then put forward as theories of architecture were only partial theories, in which it was taken for granted that theoretical concepts concerning construction and planning were dealt with in other texts.

DISTINCTIONS BETWEEN THE THEORY OF ARCHITECTURE AND THE THEORY OF ART

Before embarking on any discussion as to the nature of the philosophy of architecture, it is essential to distinguish between two mutually exclusive theories that affect the whole course of any such speculation. The first theory regards the philosophy of architecture as the application of a general philosophy of art to a particular type of art. The second, on the contrary, regards the philosophy of architecture as a separate study that, though it may well have many characteristics common to the theories of other arts, is generically distinct.

The first notion (*i.e.*, that there exists a generic theory of art of which the theory of architecture is a specific extension) has been widely held since the mid-16th century, when the artist and writer Giorgio Vasari published in his *Le vite de' più eccellenti pittori, scultori ed architettori italiani . . .* (*The Lives of the Most Eminent Italian Painters, Sculptors and Architects . . .*) his assertion that painting, sculpture, and architecture are all of common ancestry in that all depend on the ability to draw. This idea became particularly prevalent among English-speaking theorists, since the word design is used to translate both *disegno* ("a drawing") and *concetto* ("a mental plan"). But its main influence on Western thought was due to Italophile Frenchmen, after Louis XIV had been induced to establish in Rome a French Academy modelled on Italian art academies.

As a result of the widespread influence of French culture in the 17th and 18th centuries, the concept of the *beaux arts* (literally "beautiful arts" but usually translated into English as "fine arts") was accepted by Anglo-Saxon theorists as denoting a philosophical entity, to the point where it was generally forgotten that in France itself the architectural profession remained totally aloof from the Académie Royale de Peinture et de Sculpture until they were forced to amalgamate after the French Revolution.

This theory of fine art might not have been so widely adopted but for the development of aesthetics, elaborated after 1750. Thus, when academies of fine art were being established successively in Denmark, Russia, and England on the model of the French Academy in Rome, German philosophers were gradually asserting (1) that it was possible to elaborate a theory of beauty without reference to function (*Zweck*); (2) that any theory of beauty should be applicable to all sensory perceptions, whether visual or auditory; and (3) that the notion of beauty was only one aspect of a much larger concept of life-enhancing sensory stimuli.

The alternative theory (*i.e.*, that a philosophy of architecture is unique and can therefore be evolved only by specific reference to the art of building) will be dealt with below with reference to the traditional triad usually cited in the formula coined, by the English theorist Sir Henry Wotton, in his book *The Elements of Architecture*, namely "commodity, firmness, and delight."

Theorists' reluctance to use specific buildings as examples

Generally speaking, writers on aesthetics have been noticeably reluctant to use architectural examples in support of speculations as to the nature of their general theories; but references to buildings have been used in most "philosophies of art" ever since the German philosophers Immanuel Kant and G.W.F. Hegel first popularized the philosophical discipline. Kant, in his *Kritik der Urteilskraft* (1790; Eng. trans., *Critique of Judgment*, 1951), distinguished between what he termed free beauty (*pulchritudo vaga*) and dependent beauty (*pulchritudo adhaerens*). He classified architecture as dependent beauty, saying that in a thing that is possible only by means of design (*Absicht*)—a building or even an animal—the regularity consisting in symmetry must express the unity of the intuition that accompanies the concept of purpose (*Zweck*), and this regularity belongs to cognition. Nevertheless, he claimed that a flower should be classified as free beauty (where the judgment of taste is "pure") "because hardly anyone but a botanist knows what sort of thing a flower ought to be; and even he, though recognizing in the flower the reproductive organ of the plant, pays no regard to this natural purpose if he is passing judgment on the flower by taste." What Kant's reaction would have been to a modern plastic imitation flower is impossible to guess; but it will readily be perceived (1) why those who, in the 19th century, accepted the notion that beauty in architecture is *pulchritudo adhaerens* felt such antipathy toward "shams," (2) how the distinction between "pure art" and "functional art" (*Zweckkunst*) became confused, and (3) why there arose a tendency to pursue definitions of "pure beauty" or "pure art" without specifically referring to the function and structure of any particular class of beautiful or artistic objects, such as buildings.

This latter tendency was reinforced when the French philosopher Victor Cousin, in the first half of the 19th century, classified the history of philosophy under three distinct headings: the true, the beautiful, and the good. The ensuing acceptance of the idea that beauty was to be studied independently of truth and goodness produced a natural tendency not merely to regard beauty as something added to a building (rather than conceptually inseparable from the truth and goodness of its structure and function) but to regard beauty as limited to visual and emotional qualities.

Influence of aesthetics especially influential in supporting 20th-century theory

In the first half of the 20th century, philosophers grew less dogmatic about aesthetics. But its influence on theories of architecture became stronger because of the popular view that sculpture was essentially nonrepresentational. Thus, although the assertion that "aesthetically, architecture is the creation of sculpture big enough to walk about inside" is meaningful in the 20th century, it would have seemed nonsensical to any architectural theorist living before 1900, when sculpture was invariably thought of either as representational or as a carved refinement of load-bearing wood or stone.

FUNCTIONALISM

The notion of functional art, most actively promoted by German writers and termed by them *Zweckkunst*, is most appropriately related to architectural theory under three headings, namely (1) the idea that no building is beautiful unless it properly fulfills its function, (2) the idea that if a building fulfills its function it is ipso facto beautiful, and (3) the idea that, since the form of any artifact relates to its function, all artifacts, including buildings, are a species of industrial, or applied, art (known in German as *Kunstgewerbe*).

The first proposition will be dealt with later under the heading *utilitas*. The second proposition, though widely popularized through the publication of the French architect Jean-Nicholas-Louis Durand's lectures delivered during the economic depression of the beginning of the 19th century, has had little influence except during similar periods of economic depression. The third proposition has,

however, had a wide influence, since, unlike the second proposition, it is closely akin to (rather than antagonistic toward) the theory of aesthetics, in that it regards all the visual arts as generically related.

This last theory seems to have been popularized, if not originated, by Gottfried Semper, an architect from Dresden who, after finding political asylum in England (where he then helped to organize the Great Exhibition of 1851), published a book in German on arts and crafts that seems to have been influential not only in Germany but also in areas of the United States heavily populated by German-speaking immigrants, such as Chicago. Later, in 20th-century Germany, the Bauhaus (officially Hochschule für Bau Gestaltung; Academy for Form Giving) was ostensibly intended to train students in separate creative disciplines, but its didactic method was based on the assumption—implied by the general introductory courses—that, if one could design anything, one could design everything. In the explanatory words of its founder, the architect Walter Gropius, "The approach to any kind of design—a chair, a building, a whole town or a regional plan—should be essentially identical."

Bauhaus functionalism

THE ART OF BUILDING

The notion that architecture is the art of building was implied by Alberti in the first published treatise on the theory of architecture, *De re aedificatoria* (1485; Eng. trans., *Ten Books on Architecture*, 1955); for, although he was a layman writing for other lay scholars, he rejected, by his title, the idea that architecture was simply applied mathematics, as had been claimed by Vitruvius. The specific denotation of architecture as "the art of building," however, seems to be a French tradition, deriving perhaps from the medieval status of master masons, as understood by the 16th-century architect Philibert Delorme. This definition occurs in most French treatises published before 1750; and, although the humanistic and antiquarian aspects of fine building were rarely questioned after the Renaissance, the distinction between "architecture" and "building" never had any appreciable significance before Renaissance ideas succumbed to the combined assault of "aesthetics" and the Gothic Revival movement.

Before the 18th century it was generally accepted that the theory of architecture was concerned mainly with important private or civic buildings such as palaces, mansions, churches, and monasteries. Buildings such as these required the superior skill that only book learning could provide, and so relatively little attention was given, in theoretical writings, to simple and straightforward buildings that could be competently built in accordance with local traditions by unlettered craftsmen. But, with the expansion of the architectural profession, with the perversion of the idea that social prestige was symbolized by ornamentation, with the wider distribution of wealth, and with the growing urge toward individualism in an increasingly egalitarian society, the real distinction between these two kinds of buildings was obscured, and in its place was substituted an antithesis. Henceforth, "building" was associated with the notion of cheapness, whereas "architecture" was associated with what John Ruskin would have called "sacrifice" (but which his antagonists would have called conspicuous waste). A distinction was made between the respective attitudes of "art architects" and practical-minded civil engineers. This distinction persisted because of the different methods of training candidates for the two professions. Whereas a fledgling engineering student is seldom asked to design a whole structure (such as a bridge), architectural students begin by designing whole structures and proceed with structures of increasing size and complexity, either graphically or by means of small-scale models.

Differentiation between "building" and "architecture"

It was doubtless the difference in educational methods that prompted Le Corbusier to state:

> The engineer, inspired by the law of economy and led by mathematical calculation, puts us in accord with the laws of the universe. He achieves harmony. The architect, by his arrangement of forms, achieves an order which is a pure creation of his spirit . . . it is then that we experience beauty.

Yet some 80 years previously the English critic James Fergusson had felt obliged to qualify, with a comparable distinction, his enthusiasm for the new architecture of the Crystal Palace, by observing that "it has not a sufficient amount of decoration about its parts to take it entirely out of the category of first-class engineering and to make it entirely an object of fine art." The distinction between architecture and "mere building" was stated by Nikolaus Pevsner in the opening paragraph of his *Outline of European Architecture* (1942): "a bicycle shed is a building; Lincoln Cathedral is a piece of architecture . . . the term architecture applies only to buildings designed with a view to aesthetic appeal." Whatever the justification for such assertions, it must nevertheless be recognized that neither of these authors suggests that aesthetic appeal or art are synonyms for superfluity. Although adjustment in proportions or refinement of profiles may increase the thickness of short-span structural members beyond the structural analytical minima, this does not necessarily imply any radical decrease in real economy but simply indicates a concept of economy that takes into account the assembly and amenity of spatial enclosures and admits that there is value in environmental harmony. It is thus as misleading to imply (as Fergusson implied) that architecture is civil engineering plus ornament as it is to imply (as Le Corbusier did) that the status of the two professions is to be distinguished by the relative superiority of beauty over harmony.

It is important to insist that the theory of architecture is concerned primarily with the attainment of certain environmental ideals rather than with their cost; for these two problems are philosophically distinct, as is clear if one considers such a concept as, for example, that of standardization. The financial saving made by standardizing rolled-steel sections or by casting concrete in reusable formwork is so obvious that it requires no elaboration with respect to Vitruvius' demand for *oeconomia*. But such standardization also fulfills Vitruvius' concurrent demand for order, arrangement, eurythmy, symmetry, and propriety.

Standardization as a concept

The Place Vendôme in Paris is adorned with over 100 identical pilasters and half columns, all carved with the same Corinthian capitals under the supervision of a member of the Académie Royale de Peinture et de Sculpture. Whether or not the resultant uniformity was or still is both pleasing and desirable is certainly open to discussion; but it will be perceived that any argument about architectural standardization must primarily be a question of value, rather than of cost, and it is with values that architectural theory has always been predominantly concerned.

"COMMODITY, FIRMNESS, AND DELIGHT": THE ULTIMATE SYNTHESIS

It has been generally assumed that a complete theory of architecture is always concerned essentially in some way or another with these three interrelated terms, which, in Vitruvius' Latin text, are given as *firmitas*, *utilitas*, and *venustas* (*i.e.*, structural stability, appropriate spatial accommodation, and attractive appearance). Nevertheless, a number of influential theorists after 1750 sought to make modifications to this traditional triad (1) by giving its components a radically different equilibrium (such as the primacy given by the 18th-century French architect Étienne-Louis Boullée to the effects of geometric forms in light or the claim made by J.N.L. Durand that the fulfillment of function was the sole essence of architectural beauty), (2) by adding ethical values (such as Ruskin's "sacrifice" and "obedience"), or (3) by introducing new scientific concepts (such as Giedion's "space-time").

Furthermore, it has been argued that the traditional concept of *firmitas*, *utilitas*, and *venustas* ceased to have any real value after 1800, when engineers began creating structures that seemed so ostentatiously to defy the stonemasons' laws of gravity, when scientific studies were creating more and more doubts as to the economical, sociological, psychological, acoustical, thermal, or optical determinants of appropriate spatial accommodation and when beauty was "altogether in the eye of the beholder."

Clearly, one must be wary of attributing too much importance to the sequence, since a slight variation occurs in the writings of even the most traditional theorists. Vitruvius gives these terms in the sequence *firmitas*, *utilitas*, *venustas*, whereas both Alberti and, following him, the 16th-century Venetian architect and theorist Andrea Palladio reverse the order of the first two. Thus, Sir Henry Wotton's sequence (which is normally used in English-language texts) does not, as so often stated, derive directly from the Latin text of Vitruvius but from the Italian text of Palladio's *I quattro libri dell' architettura* (*i.e.*, *comodità*, *perpetuità*, *bellezza*). But it does seem worth noting that *venustas* generally comes last, implying that *firmitas* and *utilitas* are to be regarded as essential logical prerequisites of architectural beauty.

On the other hand, the practical advantages, in academic treatises, of giving priority to *venustas* are evident. Jacques-François Blondel, in his nine-volume *Cours d'architecture* (1771–77) used this sequence because he observed that considerations of "decoration" are almost entirely within the domain of the theory of architecture, whereas neither distribution (*utilitas*) nor construction (*firmitas*) can be explained properly without practical experience. The growing emphasis on aesthetics, combined with developments in psychology and the influence of art-historical methods, added weight to this argument, while the corresponding independence of scientific techniques of structural and spatial analysis led many teachers of architecture to consider *utilitas* and *firmitas* as totally seperate academic disciplines. Important exceptions can be found to this generalization. At the end of the 19th century, Julien Guadet, in reaction against the creation of a chair of aesthetics at the Paris École des Beaux-Arts, considered it his duty, as professor of architectural theory, to devote his lectures to the study of architectural planning, and this method, which achieved prestige as a result of his keen mind and wide historical knowledge, was pursued by many later scholars. But Guadet's approach became unfashionable, and since the 1960s the predominant methods of teaching architectural theory have ranged from a return to the synthesis of structural, spatial, and formal values espoused by Robert Venturi to the exploration of the architectural implications of general theories of linguistics advanced by Christian Norberg-Schulz.

Modern concepts of teaching architectural theory

Venustas. This Latin term for "beauty" (literally, the salient qualities possessed by the goddess Venus) clearly implied a visual quality in architecture that would arouse the emotion of love; but it is of interest to note that one of the crucial aspects of this problem was already anticipated by Alberti in the 15th century, as is made clear by his substitution of the word *amoenitas* ("pleasure") for Vitruvius' more anthropomorphic term *venustas*. Alberti not only avoids the erotic implications of the term *venustas* but, by subdividing *amoenitas* into *pulchritudo* and *ornamentum*, gives far more precise indications as to the type of visual satisfaction that architecture should provide. *Pulchritudo*, he asserts, is derived from harmonious proportions that are comparable to those that exist in music and are the essence of the pleasure created by architecture. *Ornamentum*, he claims, is only an "auxiliary brightness," the quality and extent of which will depend essentially on what is appropriate and seemly. Both *pulchritudo* and *ornamentum* were thus related to function and environment in that, ideally, they were governed by a sense of decorum; and, since the etymological roots of both "decoration" and "decorum" are the same, it will be understood why, before 1750, the term decoration had in both English and French a far less superficial architectural implication than it often does today.

After the German philosopher Alexander Gottlieb Baumgarten had introduced the neologism aesthetics in 1750, the visual merits of all artifacts tended to be assessed more subjectively than objectively; and, in the criticism of all those sensory stimuli that, for want of a better term, critics somewhat indiscriminately lumped together as the fine arts, the visual criteria were extended to include not only beauty but also sublimity, picturesqueness, and even ugliness. Now it is clear that, once ugliness is equated with beauty, both terms (being contradictory) become virtually meaningless. But ugliness, after the mid-19th century, was not only the most important theme of many dramas and novels; it was also often considered the most appropriate architectural expression for all sorts of virtues such as manliness, sincerity, and so on.

Before 1750, architects had expressed these qualities more subtly (*e.g.*, by slight modifications of proportions or by unobtrusive ornament). In later years, when the value of proportion and ornament became highly controversial, architectural theorists tended to avoid committing themselves to any criteria that might be subsumed under the heading *venustas*. In the last resort, however, some concept of beauty must be essential to any theory of architecture; and, whether one considers Le Corbusier's buildings beautiful or not, his most stabilizing contribution toward the theory of modern architecture was undoubtedly his constant reiteration of this term and his insistence on the traditional view that beauty in architecture is essentially based on harmonious proportions, mathematically conceived.

20th-century response to Alberti's theory

In the 20th century the main obstacle to an acceptance of Alberti's notions of *pulchritudo* and *ornamentum* resulted from the influence of nonrepresentational sculpture after 1918, whereby ornament was no longer conceived as an enrichment of proportioned structure but as an integral, all-pervading part of each building's totality. This ideal of the fusion between good proportions and "auxiliary brightness" was expressed by Walter Gropius in *The New Architecture and the Bauhaus* when he wrote in 1935:

> Our ultimate goal, therefore, was the composite but inseparable work of art, the great building, in which the old dividing-line between monumental and decorative elements would have disappeared for ever.

The idea was accepted in most schools of architecture by the mid-20th century; but one may question whether it fully justified the expectations of its protagonists, once it had been exemplified and proliferated in so many urban environments. It is by no means certain that Gropius' concept of the fundamental interdependence of architectural proportion and architectural ornament was irrevocably established by the Bauhaus theorists or that future architectural theorists need only concentrate on such minor modifications to the concept as may be required by sociological and technological developments.

Utilitas. The notion that a building is defective unless the spaces provided are adequate and appropriate for their intended usage would seem obvious. Yet the statement itself has been a source of controversy since the 1960s. The main reasons for the controversy are: first, whereas there are seldom exact statistical means of computing spatial adequacy or appropriateness, there are many building types or building elements for which one cannot even establish the optimum forms and dimensions with any confidence that they will be generally accepted. Second, edifices are frequently used for purposes other than those for which they were originally planned. Furthermore, there is some doubt as to whether "form follows function" or "function follows form," since, although, in general, it can reasonably be assumed that an architect's task is to construct specific spaces for the fulfillment of predetermined functions, there is plenty of historical evidence to suggest that many important social institutions have resulted from spaces already built. No better example could be found than the evolution of parliamentary systems. The British system, based on the concept of legislatures in which the sovereign's government and the sovereign's opposition confront each other, originated in the fact that the earliest parliaments met in the medieval palace chapel. The French system, created concurrently with the Greek and Roman revivals, was based on the concept of legislatures addressed by orators, and its environment was that of an antique theatre. In the former system the seating was designed in accordance with the liturgical requirements of a Christian church; in the latter, with the evolution of Greek drama. Neither had anything to do with preconceived notions regarding

The issue of a building fulfilling its function

the most effective environment for parliamentary debate, yet both have had divergent influences on constitutional procedures, thereby deeply affecting the whole theory of government.

Third, the exact significance of what is meant by "adequate appropriate spaces" becomes far more complex in buildings requiring a large number of interrelated spaces than it is in single-cell buildings. The emotional effect of transitions from spacious to constricted volumes and vice versa transcends in architectural importance the statistical evaluation of floor areas; a fact which explains the attractiveness of theories that have tacitly adopted places of worship as spatial paradigms and bolstered their arguments by historical reference to temples and churches. This bias is perceptible not only in the most influential theories enunciated before 1900 (when the prototypes were either primeval, antique, or medieval) but also in the most influential ideas promulgated by such great architectural leaders of the 20th century as Frank Lloyd Wright and Ludwig Mies van der Rohe.

The idealization of monumental single-cell spaces is sometimes justified, but the difficulty of evolving theories of planning by the use of historical prototypes should be emphasized. It is in this branch of architectural theory that the influences of historicism have been most insidious, precisely because they are less obvious here than in systems of construction, of proportions, and of ornamentation. Such influences persist mainly because of art-historical indifference to the essential distinction between building types, since such distinction conflicts with the chronological sequence of particular architects' stylistic evolution; but it is for this reason that Julien Guadet's greatest contribution to the theory of architecture may well have been his decision to evolve a history of architecture in which all buildings were classified solely in accordance with their function.

Firmitas. Two plausible reasons can be given for according logical primacy in the Vitruvian triad to *firmitas*. The first is the notion that architecture is essentially the "art of building." The second is that, since the uses or functions of a building tend to change, the structures serving such functions may be considered as taking logical precedence over them. This idea was expressed with characteristic lapidary vigour by the 20th-century French architect Auguste Perret when he asserted that

> architecture is the art of organizing space; but it is by construction that it expresses itself . . . Functions, customs, and building regulations and fashions impose conditions which are only transitory.

Some later architectural theorists have become so concerned with the rapid obsolescence of modern buildings that they have envisaged edifices that express the temporary nature of these transitory qualities and are therefore built in such a way as to enable the structures themselves to be discarded completely after a few years. On the other hand (since the economic feasibility of this technique is questionable), there are still many architects who believe in the inevitability of permanent buildings and who therefore hold views more compatible with this belief.

The issue of revealing the structural system to the observer

From the time of the Renaissance to the mid-18th century—as also before the decline of the ancient Roman Empire on which the culture of this era was modelled—little concern seems to have been given to the idea that there was any virtue in manifesting the actual structural system of a building. Alberti recommended a distinctive articulation of the skeleton frame in conformity with the antique concept of trabeation, or the post-and-lintel system (and hence the independence of the "infilling" elements, such as arches or solid walling); but the more commonly accepted notion seems to have been that, provided a trabeated system was expressed externally, the relationship of this visual expression to the actual system of construction was relatively unimportant. Theoretical pronouncements on this matter depended of course on the architectural traditions of each country. In Italy (where the traditional technique of building had, even during the Middle Ages, assumed that structure was independent of appearance and where it was common to complete a building in brick before adding its marble facades) the idea that there could be any theoretical dilemma regarding the unison between these two elements was virtually inconceivable. Palladio and his generation seem to have generally accepted the idea that, in regions where masonry was scarce, the use of stuccoed, painted, or veneered brickwork, with plastered timber beams, was architecturally as "genuine" as the use of stone, provided it was all of one colour. But in the Île-de-France region around Paris, on the contrary, the medieval traditions of French masonry construction, combined with the abundance of good freestone, caused theorists from the Renaissance to the time of the French Revolution to favour a less tenuous relationship between the external appearance of a building and the system by which it was constructed. Nevertheless, it is probably fair to say that in all European countries before the end of the 18th century, as well as in their American colonies, the only problem concerned with *firmitas* (other than technical problems) was the problem of the relationship between "real and apparent stability"; and, when theorists pronounced on this problem, it was usually to assert that a building should not only be structurally stable but should also appear to be so.

A violent assault upon this point of view was launched by the Gothic Revivalists, who, in the mid-19th century, contended that the breathtaking counterpoise of a cathedral's flying buttresses was far more dramatically expressive of *firmitas* than the ponderous massiveness of its sturdy western towers. It was in this era that the term daring (which Ruskin had frequently used with reference to the paintings of the English Romantic artist Joseph Mallord William Turner) became popular as a laudatory epithet, thereby indicating an ideal of structural expression that was to be increasingly exploited when steel and reinforced concrete permitted higher buildings with fewer and more slender supports.

The concept of truthful architecture

But the most controversial issue concerning *firmitas* in the 19th century—which also arose through the influence of the Gothic Revival movement—concerned the extent to which a building should manifest its structural system and the materials used. The attraction of this particular interpretation of the concept of truthful architecture was probably due to the popularity of new attitudes toward experimental science and to the disrepute into which mythology had been cast by the philosophers of the Enlightenment. Presumably, truth was no less prized in the 17th or 18th centuries than in the 19th century (though shams may have been less rife), while hypocrisy was regarded with as much contempt. Moreover, although the 19th century was a period of growing realism in literature, it was also a period of growing expressiveness in painting and music. Whatever the reason for this change of attitude, the 19th century saw a general acceptance of the notion that buildings were "true" only insofar as their structural form and appearance corresponded to the structural systems and materials employed, and this dogma was developed by means of many elaborate biological and mechanical analogies.

This particular doctrine had a highly beneficial influence on architectural evolution during the 20th century, since it helped to demonstrate why the radical changes in building technology rendered earlier concepts of architectural form (based on load-bearing masonry construction) theoretically untenable. For, while it may readily be admitted that a building can express many other things besides its function and structure, failure to express the latter in some manner, however remote, must always lead to arbitrariness. This would not only be harmful to the evolution of architectural form but would inevitably result in a somewhat cynical concept of building as "pure form"—a concept that only those who regard architecture as nothing more than large-scale packaging or abstract sculpture could accept. (Pe.C.)

BIBLIOGRAPHY

General: JOHN FLEMING, HUGH HONOUR, and NIKOLAUS PEVSNER, *The Penguin Dictionary of Architecture* (1966), a general reference work of architectural terminology and biography; PAUL FRANKL, *Die Entwicklungsphasen der neueren Baukunst* (1914: Eng. trans., *Principles of Architectural His-*

tory, 1914, reprinted 1968), a classic analysis of architectural form, 1400–1850; SIGFRIED GIEDION, *Space, Time and Architecture*, 5th ed. (1967), a stimulating survey and justification of modern architecture and its antecedents; STEEN E. RASMUSSEN, *Experiencing Architecture*, 2nd ed. (1962), a beginner's guide to architectural appreciation; JULIUS SCHLOSSER, *Die Kunstliteratur* (1924; Italian trans., *La letteratura artistica*, 2nd ed., 1956), a bibliography of theoretical writing up to 1800; selected titles in the "Pelican History of Art Series," general surveys of significant periods.

Use: JEFFREY ARONIN, *Climate and Architecture* (1953), on the influence of physical environment on planning; SIGFRIED GIEDION, *Mechanization Takes Command* (1948, reprinted 1969), on the impact of machinery on 19th- and 20th-century building; LEWIS MUMFORD, *The Culture of Cities* (1938, reprinted 1970), the growth of modern cities seen from a historical and humanitarian viewpoint; AMOS RAPOPORT, *House Form and Culture* (1969), an analysis of basic domestic forms in the light of cultural anthropology.

Techniques: LEWIS MUMFORD, *Technics and Civilization* (1934, reprinted 1962), a general view of the cultural role of technology; DEPARTMENT OF SCIENTIFIC AND INDUSTRIAL RESEARCH (London), *Principles of Modern Building*, 3rd ed., 2 vol. (1959–61), a study of building techniques and materials; CHARLES RAMSEY and HAROLD SLEEPER, *Architectural Graphic Standards*, 6th ed. (1970), the practicing designer's handbook of standards and equipment; MARIO SALVADORI and ROBERT HELLER, *Structure in Architecture* (1963), a clear, well-illustrated explanation of structural principles.

Expression and theory: (Sources): JOHN RUSKIN, *The Seven Lamps of Architecture* (1849, reprinted 1956), an aesthetic of architecture of the Romantic era allied to ethics; EUGENE VIOLLET-LE-DUC, *Entretiens sur l'architecture*, 2 pt. (1862–72; Eng. trans., *Discourses on Architecture*, 1875), a pre-modern architectural theory based on rational construction; VITRUVIUS, *Ten Books on Architecture* (Eng. trans. 1914), the only architectural treatise to survive from antiquity—exerted great influence on Renaissance and later design (*Modern*): CHRISTOPHER ALEXANDER, *Notes on the Synthesis of Form* (1964), design calculations for the cybernetic age; LE CORBUSIER, *Vers une architecture* (1925; Eng. trans., *Towards a New Architecture*, 1927); WALTER GROPIUS, *Scope of Total Architecture* (1955); and FRANK LLOYD WRIGHT, *Modern Architecture* (1931), the credos of the three most influential modern architects. (*Surveys*): PETER COLLINS, *Changing Ideals in Modern Architecture, 1750–1950* (1965), a survey of architectural principles; PAUL FRANKL, *The Gothic* (1960), a medieval style and its survivals through the centuries; GEOFFREY SCOTT, *The Architecture of Humanism*, 2nd ed. (1924, reprinted 1969), combines a critique of 19th-century theory with a psychologically based defense of Baroque design; JOHN SUMMERSON, *The Classical Language of Architecture* (1963), on the use of the classical repertory of motives through the ages; RUDOLF WITTKOWER, *Architectural Principles in the Age of Humanism*, 3rd ed. (1962), architectural thought in the Renaissance.

Ornament: A useful general survey of ornamental forms and designs is JOAN EVANS, *Pattern: A Study of Ornament in Western Europe from 1180 to 1900*, 2 vol. (1931). For an account of mimetic ornament and design, the works of E. BALDWIN SMITH are recommended: *Egyptian Architecture As Cultural Expression* (1938, reprinted 1968), *The Dome* (1950), and *Architectural Symbolism of Imperial Rome and the Middle Ages* (1956). For a theory of ornament as social function, identifying use and humanizing artifacts, see ALAN GOWANS, "The Unchanging Arts of Beautification: Commercial Design and Decoration," in *The Unchanging Arts* (1971).

Theory and criticism: LEONE BATTISTA ALBERTI, *De re aedificatoria* (1485), the first printed book on the theory of architecture—numerous English translations are available; REYNER BANHAM, *Theory and Design in the First Machine Age* (1960), an authoritative study of the theories of architecture developed in the second quarter of the present century; PETER COLLINS, *Architectural Judgement* (1971), a comparative study of decision making in architecture and law; CHRISTIAN NORBERG-SCHULZ, *Intentions in Architecture* (1963), an influential study of architectural theory based on linguistics; HOWARD ROBERTSON, *The Principles of Architectural Composition* (1924), a characteristic textbook of the first quarter of the present century; ROBERT VENTURI, *Complexity and Contradiction in Architecture* (1966), a proposal for a new theory of architecture based on recent art-historical interpretations of Mannerism and Baroque architecture; EDWARD DE ZURKO, *Origins of Functionalist Theory* (1957), a compendious historical analysis of the relationship between form and function, as conceived by philosophers and architectural theorists.

(J.S.Ac./Pe.C./Al.Go.)

Arctic Islands

The islands lying north of the Canadian mainland, forming one of the world's great archipelagos, extend about 1,500 miles east to west and 1,200 miles from the mainland to Cape Aldrich, the northernmost point of Ellesmere Island. Two large peninsulas, Boothia and Melville, extend northward from the mainland and will be discussed here as parts of the archipelago. The land area of this region exceeds 500,000 square miles (1,300,000 square kilometres), about one-seventh of the area of Canada.

Location and general characteristics

The Arctic Islands lie far north of the temperature line, or isotherm, along which 50° F (10° C) is the mean July temperature; this line coincides closely with the northern limit of trees on the mainland and commonly is used to define the southern limit of the world's Arctic region. Like much of the adjoining mainland, the islands are characterized by permanently frozen ground except for a thaw to the depth of about one foot during the summer.

Ice restricts sea access to most of the islands to the late summer and early winter. The few and widely scattered permanent settlements lie mainly on the southern islands. Although remote, the region has become accessible the year round through a network of airfields and air routes connecting centres of population to the south. For thousands of years the archipelago was peopled only by Eskimo, who evidently migrated to this land from Asia. Norsemen may have reached the easternmost islands in the 10th and 11th centuries, and from the 16th century numerous European expeditions carried out a prolonged search for a northwest passage through the islands to the Orient.

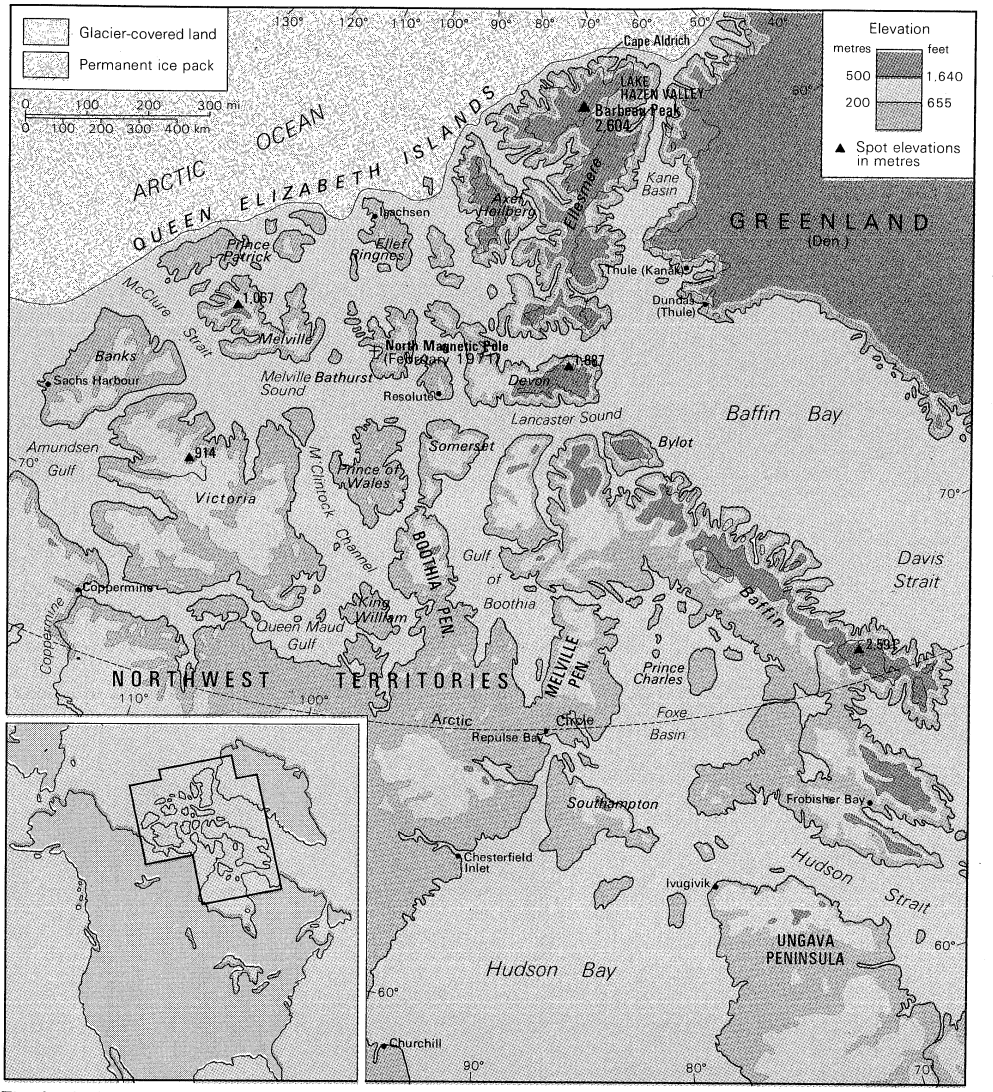
A few minerals and rocks have long been used by the aboriginal Eskimo—native copper and flint for knives and weapons and soapstone for lamps and cooking pots—but economically valuable minerals have played only a small part in the history of the Arctic Islands. Exploitation began with mining and shipping of "gold ore" (probably amphibolite with minor, bronze-coloured biotite) from Baffin Island in the 16th century by the English explorer Martin Frobisher. Graphite, mica, and coal have been mined on the same island. Recently, increased mineral exploration has resulted in promising discoveries of iron and lead–zinc. Most spectacular, however, was the modern oil and gas exploration program that followed discovery of a major oilfield in northern Alaska in 1968. An oil-producing zone has recently been found by drill, and three wells have been capped as gas wells.

The Canadian Arctic Islands and Boothia and Melville peninsulas form the District of Franklin, one of three administrative divisions of Canada's Northwest Territories. Government is by both elected and appointed councillors seated at the territorial capital, Yellowknife, on the mainland. (This article covers the physical environment and ecology of the region. For related information on its human geography, see NORTHWEST TERRITORIES; NORTHWEST PASSAGE; WESTERN ARCTIC CULTURES; CANADA; see also NORTH AMERICA.)

The physical environment. *General surface features.* The surface of the archipelago includes plains, lowlands, uplands, plateaus, highlands, and mountains. Most elevated surfaces are rolling and dissected by valleys and deep channels. Physiography generally expresses the type and structure of the bedrock: extremely resistant granitic and gneissic rocks form mountains, highlands, and uplands; folded sedimentary rocks form ridged mountains and uplands; slightly folded or flat-lying limestones and related rocks form extensive plateaus and lowlands; and weakly consolidated sandstones and shales underlie lowlands and plains.

Geology and geomorphology of the archipelago

Mountains and highlands lie along the eastern coasts of Baffin, Devon, and Ellesmere islands to form a high eastern edge of the archipelago. The surface slopes generally westward through plateaus and lowlands interrupted by upland and some mountain areas, then drops away to the low Arctic Coastal Plain that lies along the northwestern margin of the region. The eastern mountain belt, in which the highest land attains an elevation of about 7,000



The Arctic Archipelago of North America.

feet, is deeply dissected by fjords to form a region of high relief and rugged scenery. Another belt of rugged mountains forms Axel Heiberg Island and the northern edge of Ellesmere Island. The highest point in the Canadian Arctic lies in northern Ellesmere Island: Barbeau Peak, elevation 8,544 feet (2,604 metres). Plateaus in the central part of the archipelago reach average elevations of 800 to 2,000 feet, with hills and mountains of about 2,500 feet.

Structural regions. The Canadian Arctic comprises two geological provinces: the Central Stable region in the south and the Innuitian (from Eskimo *innuit*, "the people," the word for themselves) orogenic region (a once mobile belt) in the north. The Central Stable region includes the vast crystalline complex of the Canadian Shield, formed during the Precambrian (with radiogenic dates between 2,800,000,000 and 600,000,000 years ago), and large areas of overlying, little-deformed sedimentary rocks of Late Precambrian and Paleozoic (more than 225,000,000 years ago) eras. The Innuitian region is a sinuous belt of slightly to severely folded rocks ranging in age from the Late Precambrian Era to the Tertiary Period (*i.e.*, up to as little as 2,500,000 years ago).

The Innuitian region was the site of two major sedimentary basins: the Franklinian Geosyncline and the Sverdrup Basin. The Franklinian Geosyncline, the earlier of these tectonic depressions, or regional downwarps, received sediments nearly continuously from Late Precambrian to Late Devonian times (about 345,000,000 years ago). An episode of folding and faulting of the geosyncline in latest Devonian time is named the Ellesmerian orogeny, or mountain-building episode. Following this orogeny, much of the area was again depressed to form the Sverdrup Basin, which received sediments between Late Mississippian and Early Tertiary times. The rocks of the Sverdrup Basin were deformed during the Early Tertiary, some 65,000,000 years ago, by compressive forces of the Eurekan orogeny.

Ages and thicknesses of geological formations. The Canadian Shield of Precambrian metamorphic rocks forms a "basement" complex along the southern and eastern edges of the archipelago. Rocks of the Shield are mainly granite, gneiss, and schist. Late Precambrian to Tertiary sedimentary rocks overlie the shield to form northeast-trending belts to the north and northwest; these younger rocks include limestone, dolomite, shale, sandstone, gypsum, and volcanic rocks such as andesite and basalt. Bedded rocks of Late Precambrian Era attain thicknesses of 12,000 to 19,000 feet on Victoria and Baffin islands but are absent over large areas due to later erosion. More widely exposed on the southern islands are Early Paleozoic sedimentary formations with a maximum total thickness of 10,000 feet. The Early Paleozoic formations pass northward into a 40,000-foot-thick, mainly marine sedimentary sequence that occupies the Franklinian Geosyncline. The geosynclinal sequence passes from marine to nonmarine in the uppermost (Devonian) part, in which thin coal seams, as well as plant and fish remains, occur in sandstones. On northern Axel Heiberg and Ellesmere islands, volcanic rocks and related sediments occur along with a metamorphic complex that in part predates the geosyncline but may in part derive from

it. The Sverdrup Basin, overlying the folded and partially eroded geosyncline, contains some 50,000 feet of mainly marine sedimentary rocks. Nonmarine strata in the upper part of the sequence include coal and basaltic flows. Late Paleozoic anhydrite deep in the basin was forcibly intruded into overlying beds by orogenic pressures or by the weight of the superimposed rocks and now appears at the surface as conspicuous gypsum–anhydrite diapirs, or "domes."

The Arctic Coastal Plain is a narrow belt of very low relief underlain by Late Tertiary and Early Quaternary sand and gravel and is presumably continuous with the continental shelf, which extends about 100 miles into the Arctic Ocean.

Final shaping and drainage. The form and drainage pattern of the archipelago probably dates from Late Tertiary and Pleistocene times, or since about 25,000,000 years ago. The bedrock was deeply eroded after the Early Tertiary orogeny, and remnants of the erosion surface now stand at various elevations. Faults evidently border many islands, and many channels and major valleys probably are related to Tertiary faults. The subsequent drainage system was modified by Pleistocene glaciation and now is largely submerged in marine waters.

Glacial geology. Successive glacial advances and interglacial intervals of the Pleistocene Epoch followed worldwide climatic changes some 2,000,000 to 3,000,000 years ago. The Arctic Islands were largely ice covered during this epoch, as was most of the rest of Canada. The effects of the latest, or classical Wisconsin, glacial invasion (climaxing about 20,000 years ago) are most abundant and fresh, but earlier advances and interglacial events are recorded at many places. Some areas of the westernmost islands lack obvious evidence of glacial activity and may have escaped glaciation entirely.

Two main ice sheets occupied the islands: the Laurentide Ice Sheet flowing northward from the mainland to override the southern islands, and the Innuitian Ice Sheet, a coalescing series of glacier complexes that were especially well developed in the highlands and mountains of the eastern part of the archipelago. Markedly linear glacial features, such as drumlins (hills of glacier-deposited drift) and eskers (glacial stream deposits), are abundant on the lowlands of the southern islands, while glacial valleys, rounded rock surfaces, and other alpine glacial features are evident in the higher eastern islands.

Effects of the ice sheets

Layered sediments containing plant and woody material occur at several localities beneath glacial till, between till layers, and overlying older till beyond the recognized limit of the late Wisconsin Ice Sheet. Radiocarbon ages greater than 30,000 years have been obtained from the organic matter, which includes pollen of grasses and sedges, other herbaceous plants, and alder, willow, and spruce trees. Willow and spruce wood, some beaver gnawed, is found. Evidently the interglacial climate was somewhat more favourable than at present.

Marine shells yielding radiocarbon ages from 20,000 to more than 40,000 years occur in stratified deposits up to a few hundred feet above sea level and as scattered debris at higher elevations, as on upland surfaces. The dates are minimums, and the "old" shells indicate marine events prior to the last glaciation. Some of the older shells are broken and abraded and must have been transported and dispersed by glacial ice.

Deglaciation in the Arctic Islands was well advanced by about 9,000 years ago, as dated by shells from the highest beaches. Deglaciation of the islands, as elsewhere, appears to have been marked by fluctuations of the ice front. Scattered proglacial lakes formed as the ice abated, and the sea invaded the land, which was then still depressed because of the former ice load. Remarkable flights of raised beaches, now standing from a few feet to about 500 feet or more above sea level, attest to postglacial uplift following removal of the ice load. The greatest uplift occurs in a linear, nearly central zone trending northeasterly through the islands from Boothia Peninsula. The emerged marine features, however, present a complex pattern because of the rise in sea level during deglaciation, the variation in time of ice retreat—and hence of marine incursion—from area to area, and possible deformation or possible fault movement along some coasts.

Climate. The southern boundary of the Arctic climate region is usually placed at the tree line, which lies on the Canadian mainland. North of this distinct natural boundary, the growing season is too short and cold to allow tree growth.

The climate of the Arctic Islands is unique in many respects because of the high latitude, the barren nature of the land surface, and the maritime influence of the Arctic Ocean, Baffin Bay, and the interisland waterways. The winters are longer and the summers cooler than more southerly parts of Canada, but conditions in the Arctic are usually not so fierce or extreme as some popular accounts would suggest.

The Arctic region is snow and ice covered for more than half the year, subjected to a cold, dry climate. The daily average temperatures of the three coldest months range from −20° F (−29° C) in southern sections to −30° F (−34° C) in the north. The average temperatures generally remain below 0° F (−18° C) during six cold months and rise above the melting point only during June, July, and August. The record low temperatures of −55° to −60° F (−48° to −51° C) at Arctic stations are not as low as the −70° F (−57° C) and colder temperatures reported for parts of the mainland. The summer temperatures are cool, generally below 45° F (7° C), with occasional brief sunny intervals of 65° to 70° F (18° to 21° C). Winter sets in about the end of August in the highest latitude and in early or mid-September in the southern islands. The lowest temperatures occur between December and April, with January the most severe month in southern parts and February in the high latitudes. Thawing begins in late May to early June, and the snow has disappeared from much of the land by early July. The breaking up of sea ice, which attains thicknesses of about five to seven feet, occurs considerably later than the beginning of the thaw, but by late July it is well advanced in the southern and eastern periphery of the archipelago. Open water suited to ship navigation advances slowly into the archipelago during August and September, but the channels between the northwestern islands are ice choked the whole season (June to September).

Temperature ranges

The prolonged Arctic cold season is a result of the far northern latitude: incoming solar radiation is minimal during daylight because of the low angle of the sun's rays and is absent during the long Arctic night. In addition, although the amount of solar energy available during the periods of long days and continuous daylight is even greater than in southern latitudes, only a small percentage reaches the ground because of the high reflectivity of the snow and ice surfaces and cloud layers. The sustained cooling results in a continuous winter ice cover on the seas and channels, generally stable, cloud-free air, and consequent light snowfall. During the thaw season the open water and wet, thawing land result in high humidity and notorious, low-lying stratus clouds and coastal fogs.

Annual precipitation is low—about three to seven inches (75 to 175 millimetres)—throughout the archipelago except along the mountainous eastern coastal sections, where the height of the land and the proximity to the open waters of Baffin Bay, Davis Strait, and the North Atlantic result in anomalously high precipitation and persistent icecaps.

The worst climate in the Arctic Islands is that of the Hudson Strait region. Dominated by open water and frequent cyclonic activity, this region has the highest average temperature, the heaviest snowfall, the highest average wind speeds, and the greatest number of days of summer fog in the Canadian Arctic. (R.L.Ch.)

Vegetation and animal life. Plants and animals of the Canadian Arctic Islands are relatively few for so large an area but are of great interest through their adaptations to extremely severe conditions. Spitsbergen, lying far north of the Scandinavian Peninsula but bathed by the Gulf Stream, supports many temperate plants that cannot grow in more southerly Canadian islands, which are surrounded by cold water flowing from the Arctic Ocean.

The continuous ice and smooth wind-packed snow cause seeds and other reproductive structures to be blown great distances in winter gales. Many plants are circumpolar without local variation, indicating dispersal between Old and New World lands. There is evidence that some insects may be carried similarly.

Environmental hazards and adaptive mechanisms. Flowering plants must flower and fruit rapidly in a short, cold summer, freeze with impunity at any time, resume full growth as soon as they thaw, and withstand abrasion by gritty snow crystals in winter gales. All other Arctic plants—ferns, horsetails, mosses and liverworts, lichens, algae, and fungi—as well as insects—the primitive arthropods known as springtails, spiders, and mites—share the ability to freeze at any time and resume full metabolism as soon as the temperature permits it. This is the most important characteristic of all Arctic organisms except birds and mammals. A few crucifers, members of the mustard family, may mature their seed in a second year if snow cover is complete, while many fungi and insects mature gradually in two to several years, stopping growth at any stage and continuing the next spring.

Suspension and reactivation of metabolic functions

Although small animals may take shelter from gales, plants must either grow in sheltered sites or tolerate snow abrasion. A common protection is assumption of a compact cushion, in which the winter buds lie below the level of the old stems and leaf tips. The projecting parts cause eddies that slow down snow particles.

Mammals of or above the size of Arctic hares can carry fur heavy enough to give them excellent insulation, and they need little shelter. Such coats are unmanageable for such small species as ermine and lemmings, which must spend most of the winter below ground or below the snow. Snowfall is often very light, and large areas may be swept bare by wind. Soil temperature may then be so depressed that lemmings fail to reach breeding condition.

Native species. The most widespread mammals are Arctic fox, wolf, ermine, polar bear, Arctic hare, collared lemming, caribou, and musk-ox. The brown lemming (*Lemmus sibiricus*) occurs on the southern islands, while the red fox and wolverine occur on Baffin Island.

Most birds leave the islands before winter, but a few willow ptarmigan, snowy owls, and ravens (*Corvus corax*) remain. Common birds in the high-latitude Queen Elizabeth Islands include the red-throated loon, snow goose, old-squaw and king eider ducks, rock ptarmigan, a few sandpipers, long-billed jaeger, glaucous gull, Arctic tern, and a few buntings. About 35 other species, notably plovers and gulls, breed regularly in the southern islands.

The most important fish is the Arctic char, a staple food item of the Eskimo and increasingly popular for shipment south. No reptiles or amphibians reach the islands.

The insect life is of major consequence. A few moths and bumblebees reach northern Ellesmere Island, but the true flies, or Diptera, are by far the most abundant order. They include many midges, a few mosquitoes, and some muscoid flies. By their great numbers, they are important in pollination, and they are the main food for various birds. Beetles are few, except in the southern islands. Other orders are scarce or absent.

About 325 species of flowering plants grow in the region. Grasses, sedges, and rushes are common, but most other monocotyledons are absent. Dwarf birch and several willows reach the southern islands, but only one prostrate willow reaches northern Ellesmere. Other conspicuous species or groups include mountain sorrel, bistort, various Caryophyllaceae, buttercups, various Cruciferae, poppies, saxifrages, cinquefoils, mountain avens, crowberry, fireweed, louseworts, fleabanes, and dandelions.

Five ferns, three horsetails, and two clubmosses are recorded, all southern species that survive on sterile sites where competition is lacking. Mosses, liverworts, and some groups of algae are plentiful; but few species are endemic to the Arctic, and some are nearly cosmopolitan. Lichens are numerous and often conspicuous. They are important in delaying water loss from gravel, thus helping to establish mosses and flowering plants. Large lichens are major items in the diet of caribou.

Fungi are of great ecological importance, for they largely replace bacteria in breaking down plant remains. Various mushrooms and a few puffballs and large cup fungi grow in soil; and nearly 300 species of microfungi are parasitic or saprophytic on leaves and stems.

Abundance. Plant cover is governed by summer climate and by soil and rock type. The numbers of vascular plants correlate closely with the July mean temperature. At Hazen Camp (42.8° F; 6° C), 105 species occur and probably 115 in all Lake Hazen Valley; at Resolute (39.7° F; 4.3° C) 70 species occur; and at Isachsen (38.5° F; 3.6° C) there are 48 species. On the outer coast 40 miles northwest of Isachsen, where the July mean is 35.5° F (1.9° C), traces of only two species were found. Thus, the flora diminishes precipitously as the freezing point in summer is approached. In many of the western islands, low relief and rapidly weathering sandstone or limestone give extremely barren substrates with much less than 1 percent ground cover. Near Isachsen a pair of caribou with one fawn required 300 square miles for feeding, and, in some even more desolate areas, they probably could not exist.

Total productivity is not known for any site, but breeding-bird censuses give rough comparisons. Isachsen yielded 9.61 birds per square mile; Hazen Camp, 24.99; and grassy tundra in southwestern Baffin, 391 and 544. (In contrast, temperate forests and swamps range from 1,000 to 11,785.) Such figures emphasize the low productivity and precarious balance of Arctic ecosystems.

(D.B.O.S.)

BIBLIOGRAPHY

The physical environment: CANADIAN HYDROGRAPHIC SERVICE, *Pilot of Arctic Canada* (1959), a general account of geography, history of exploration, ice characteristics, and climate for navigators; ANDREW TAYLOR, *Geographical Discovery and Exploration in the Queen Elizabeth Islands* (1955), a résumé of geographical discovery and exploration in the northern islands to about 1940; R. THORSTEINSSON and E.T. TOZER, "Geology of the Arctic Archipelago," in the *Geological Survey of Canada, Geology and Economic Minerals of Canada* (1970), a summary account; H.S. BOSTOCK, "Physiographic Subdivisions of Canada," *ibid.;* H.A. THOMPSON, *Climate of the Canadian Arctic*, an augmented reprint from the *Canada Yearbook* (1967).

Vegetation and animal life: J.A. DOWNES, "Arctic Insects and Their Environment," *Can. Ent.*, 96:279–307 (1964), a review of recent studies; W.E. GODFREY, *The Birds of Canada* (1966); J.D. MCPHAIL and C.C. LINDSEY, *Freshwater Fishes of Northwestern Canada and Alaska* (1970); A.H. MACPHERSON, "The Origin of Diversity in Mammals of the Canadian Arctic Tundra," *Syst. Zool.*, 14:153–173 (1965), with range maps; A.E. PORSILD, *Illustrated Flora of the Canadian Arctic Archipelago*, 2nd ed. (1964), keys, descriptions, illustrations, and range maps; D.B.O. SAVILE, "The Botany of the Northwestern Queen Elizabeth Islands," *Can. J. Bot.*, 39:909–942 (1961), on glacial and postglacial plant geography; D.B.O. SAVILE, *Arctic Adaptations in Plants* (Canada Department of Agriculture, 1972).

(R.L.Ch./D.B.O.S.)

Arctic Ocean

The Arctic Ocean is by far the smallest of the Earth's five oceans, having only a little over one-sixth the area of the next largest, the Indian Ocean. Its area of 4,732,000 square miles (12,257,000 square kilometres), however, is five times larger than that of the largest sea, the Mediterranean. The deepest sounding obtained in Arctic waters is 18,050 feet, but the average depth is only 3,240 feet.

Distinguished by several unique features, including a cover of perennial ice and almost complete encirclement by the landmasses of North America, Eurasia, and Greenland, the north polar region has been a subject of speculation since the earliest concepts of a spherical earth. From astronomical observations the Greeks theorized that north of the Arctic Circle there must be a midnight sun at midsummer and continual darkness at midwinter. The enlightened view was that both the northern and southern polar regions were uninhabitable frozen wastes, whereas the more popular belief was that there was a halcyon land beyond the north wind where the sun always shone and people called Hyperboreans led a

peaceful life. Such speculations provided incentives for adventurous men to risk the hazards of severe climate and fear of the unknown to further geographic knowledge and national and personal prosperity. For related information, see BERING SEA AND STRAIT; ARCTIC ISLANDS; and articles on the individual seas of the region.

EXPLORATION

As 20 percent of their country lies north of the Arctic Circle it is not surprising that Russians were among the first and most active Arctic explorers. As early as the 17th and 18th centuries, privately financed trapping and trading expeditions made major discoveries in the far north. In 1773 an expedition was sent to the Arctic by Peter the Great, where it spent more than ten years searching for a northeastern sea route for mercantile traffic between Europe and Asia. Although the venture did not succeed in this mission, the explorers mapped the coastlines of northern Russia from the White Sea to the Chukchi Sea and discovered the Aleutian and Kuril islands and much of the northwestern shore of North America.

Expedition of Peter the Great

In contrast to this national effort of the Russians, smaller publicly and privately financed expeditions were sent to the Arctic from other countries. These efforts, however, were directed principally toward geographic exploration of coastal lands or searches for a lost expedition (*e.g.*, the Franklin Expedition), and no attempt was made to penetrate into the central reaches of the polar sea.

By the latter part of the 19th century most of the Arctic Ocean coastline had been fairly well charted, although the greater part of the ocean itself was still unexplored. The northward extent of Greenland and the island groups to its west was yet unknown, and opinions of geographers concerning the distribution of land and sea were divided. Many believed that Greenland extended across the North Pole to Wrangel Land (Wrangel Island), whereas others believed that the unknown central polar region consisted of numerous islands separated by shallow waters.

In 1879 United States Navy Comdr. G.W. De Long set out in the ill-fated ship "Jeannette" in search of possible new lands north of Siberia. His ship became entrapped in the ice at 71°35′ N, 175°06′ W, and drifted with the ice pack for nearly two years before finally being crushed and sunk at 77°15′ N, 154°59′ E. This drift demonstrated that Wrangel Land was not contiguous with Greenland.

Nansen's explorations

The discovery of flotsam from the crushed "Jeannette," driftwood from Siberia, and implements belonging to Alaskan Eskimos led the Norwegian explorer Fridtjof Nansen to conclude that if a ship of sufficient strength and proper design were allowed to be frozen in the ice near the vicinity in which the "Jeannette" had sunk, it would drift with the currents until released off the coast of Greenland. Such an operation would permit the exploration of a large part of the inaccessible central Arctic, and possibly even the North Pole could be reached. Despite the profound skepticism of most Arctic explorers, Nansen pursued his plan and in 1893–96, in the "Fram," successfully completed the drift from the New Siberian Islands to Spitsbergen.

In accord with popular opinion, Nansen expected to find only shallow water in the North Polar Basin. Consequently the "Fram" was equipped to make soundings to only 6,000 feet. Soon after the drift began, depths in excess of this were encountered. Unravelling one of the ship's steel-wire cables and joining the individual strands end-to-end permitted 11 successful soundings to be completed under most difficult conditions. These soundings gave depths ranging from 11,000 to 13,000 feet, showing that there was a deep basin under at least part of the North Polar Sea. These deep soundings mark the true discovery of the Arctic Ocean.

Following Nansen's deep soundings, additional information on the deep ocean basin was accumulated very slowly. During his dash to the North Pole in 1909, the American explorer Robert E. Peary completed several soundings at shallow and intermediate depths north of Greenland. His single attempt at a deep cast near the Pole failed when his line broke at 9,000 feet.

The interwar period saw several further explorations. Particularly notable was the effort made by the Soviet Union, which, after the 1917 Revolution, intensified its efforts with the initiation of an integrated, long-range scientific study. Lenin himself was instrumental in formulating this program, which began with the outfitting of ice-strengthened ships capable of penetrating deep into the ice pack, followed by the establishment of a network of scientific stations on islands in the Arctic Ocean and on the adjoining coast.

In 1937 the Soviet Union landed a ski-equipped airplane on the North Pole, establishing on the sea ice a drifting station, which it occupied continuously for nine months; in 1941 it established three airlifted, floating scientific stations on the ice pack over deep water in the vicinity of 83° N, 150° E.

Since 1955 the Soviet Union has continuously maintained one, and usually two, drifting scientific stations in the Arctic Basin, in addition to making numerous airlifted expeditions.

In contrast to the substantial Soviet effort, other nations were slow in showing much interest in further exploration, although during the early 1950s the United States made over 20 landings by air on the polar ice, establishing eight oceanographic stations and one scientific station (T–3) on an ice island. The latter was permanently occupied from April 1952 to May 1954.

Activity during the International Geophysical Year

Arctic Ocean research received new impetus during the International Geophysical Year (1957–58). In addition to aircraft activity, there have been, since 1957, three or more drifting stations in continuous operation, usually two Russian and one American. Another major advance occurred at this time: the advent of nuclear-powered submarines made it possible to explore beneath the ice cap, and in 1958 the USS "Nautilus" and "Skate" made successful polar cruises.

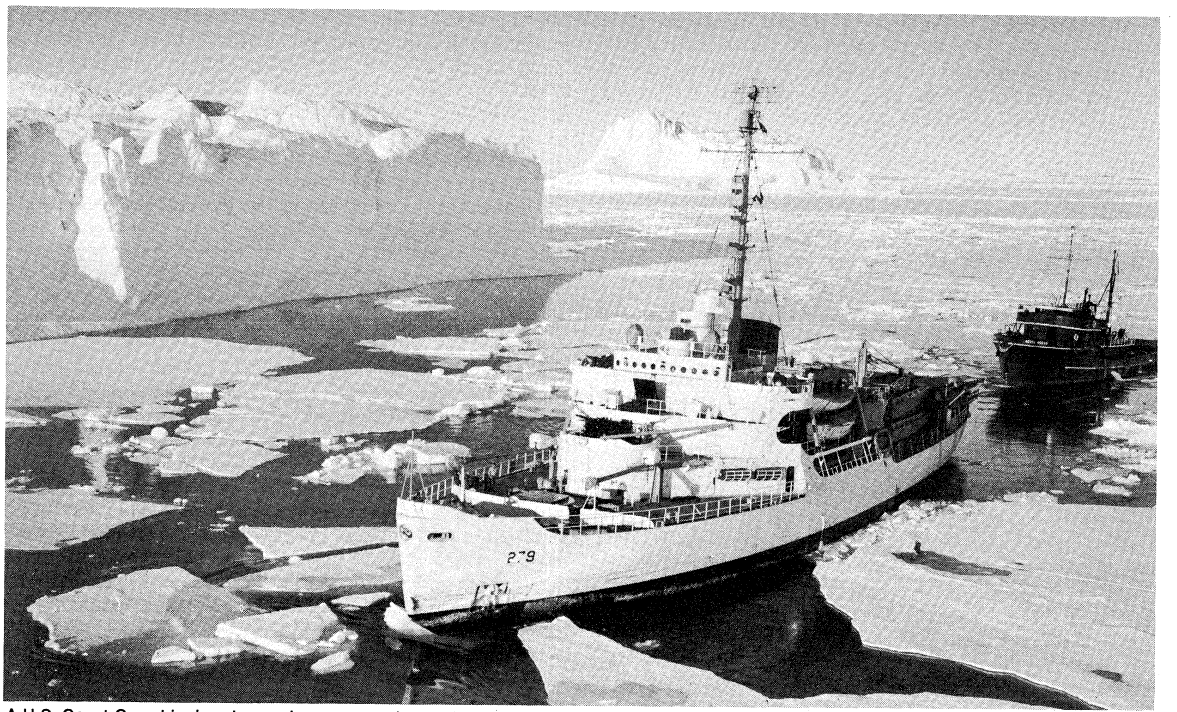
In 1958 the Canadian government organized a major and continuing program for investigation of the waters off its northern coast. The Polar Continental Shelf Project, as this effort is called, is providing a systematic and thorough investigation of the geology, geophysics, and oceanography of the region bounded by the extreme longitudes of Canada and extending to the geographic pole. Finally, the development of mineral resources in the far north is giving great impetus to intensified studies of ocean circulation and ice dynamics. Much of this effort is now being undertaken by industry, as witness the experimental cruises of the icebreaking oil tanker, "Manhattan," in 1969 and 1970.

PHYSICAL CHARACTERISTICS

The ocean floor. From the late 19th century, when Nansen first discovered an ocean in the central Arctic, until the middle of the 20th century, it was believed that the Arctic Ocean was a single large basin. Explorations since 1950 have revealed the true complex nature of the ocean floor. Rather than being a single basin, the Arctic Ocean consists of four deep basins formed by three transoceanic submarine ridges. The central of these ridges extends from the continental shelf off Ellesmere Island to the New Siberian Islands, a distance of 1,100 miles. This enormous submarine mountain range was discovered by Soviet scientists in 1948–1949 and reported in 1954. It is named the Lomonosov Ridge after the famous scientist, poet, and grammarian.

The Lomonosov Ridge has an average relief of about 10,000 feet and divides the Arctic Basin into two physiographically complex sub-basins. These are referred to as the Eurasia Basin on the European side of the ridge and the Amerasia Basin on the American side. The Lomonosov Ridge varies in width from 40 to 120 miles, and its crest ranges in depth below sea level between 3,100 and 5,400 feet.

The Eurasia Basin is divided into two smaller basins by a trans-Arctic Ocean extension of the Mid-Atlantic Ridge. This arctic segment of the global ridge system is called the Nansen Cordillera. Although it is a locus of active ocean floor spreading, its topography is less well developed than other oceanic ridges. The Fram Basin lies



A U.S. Coast Guard icebreaker and an oceangoing tugboat travelling through the Arctic Ocean on a mission during the brief Arctic summer.
By courtesy of the U.S. Coast Guard; photograph, Monkmeyer

between the Nansen Cordillera and the Lomonosov Ridge at a depth of 14,070 feet. The geographic north pole is located over the floor of the Fram Basin near its juncture with the Lomonosov Ridge.

The smallest of the Arctic Ocean sub-basins, the Nansen Basin, lies between the Nansen Cordillera and the Eurasian continental margin, and has a floor depth of 13,800 feet.

The Amerasia Basin is divided into two unequal basins by the Alpha Cordillera, a broad rugged submarine mountain chain that extends to within 4,600 feet of the ocean surface. The cordillera is believed to be a no longer tectonically active segment of the global rift system. The Makarov Basin lies between the Alpha Cordillera and the Lomonosov Ridge, and its floor is at a depth of 13,200 feet. The largest sub-basin of the Arctic Ocean is the Canada Basin which extends approximately 700 miles from the Beaufort Sea Shelf to the Alpha Cordillera. The smooth basin floor slopes gently from east to west where it is interrupted by regions of sea knolls. The average depth of the Canada Basin is 12,500 feet.

The Arctic Ocean is unique in that nearly one-third of its total area is underlain by continental shelf, which is asymmetrically distributed around its circumference. North of Alaska and Greenland the shelf is 60 to 120 miles wide, which is the normal width of continental shelves. In contrast, the Siberian and Chukchi shelves off Eurasia range from 300 to 1,100 miles in width. The edge of the continental margin is dissected by numerous submarine valleys. The largest of these, the St. Anna Trough, is 110 miles wide and 300 miles long.

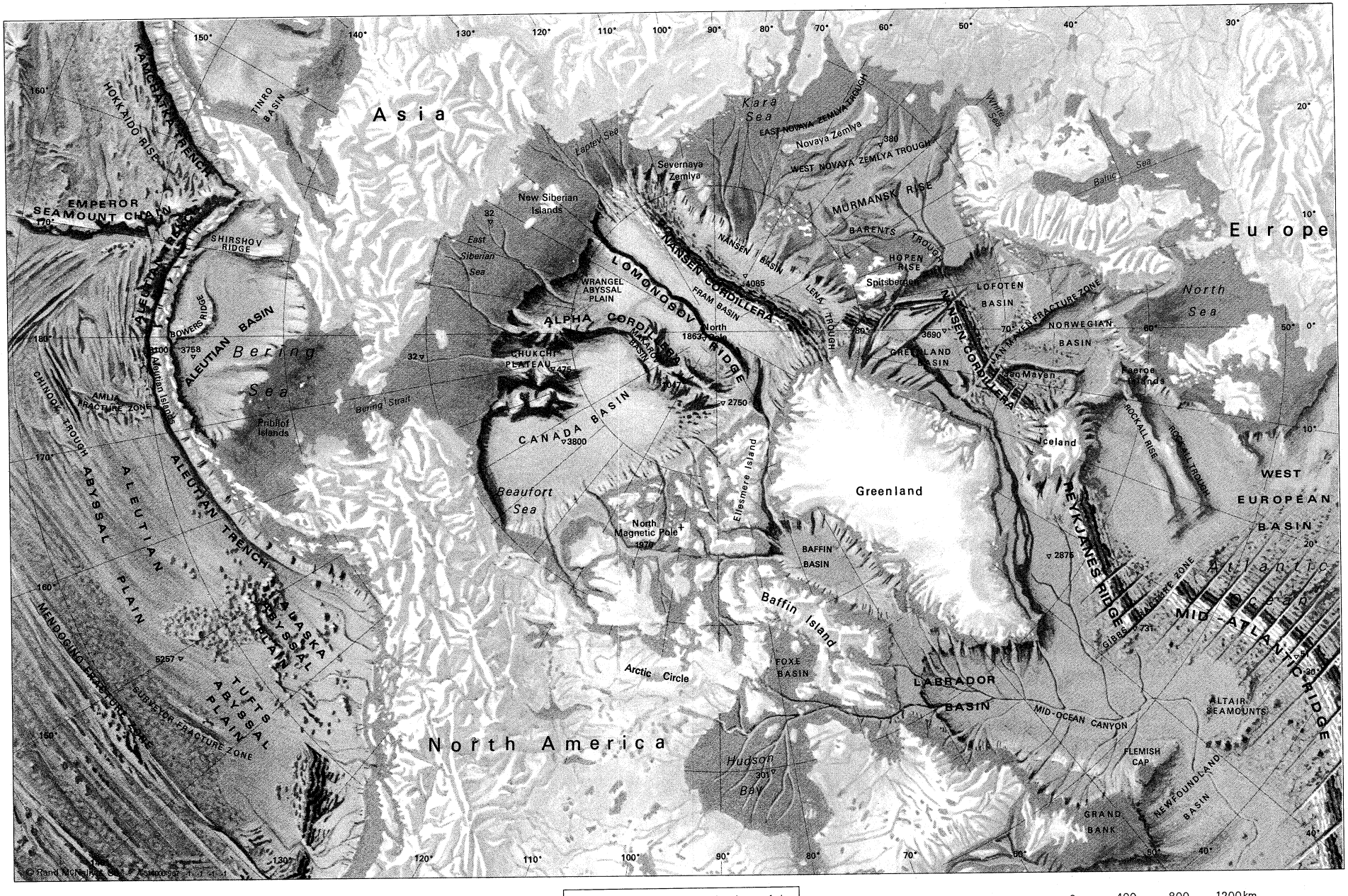
Oceanography. As an approximation, the Arctic Ocean may be regarded as an estuary of the Atlantic Ocean. The major circulation into and from the Arctic Basin is through its single, deep channel connection to the World Ocean: the passage that lies between the islands of Spitsbergen and Greenland. The connection of the Arctic Ocean with the Pacific is of minor importance. There are, however, factors in the Arctic that make its processes significantly different from those of the Atlantic. Most notable is the covering ice pack, which reduces the exchange of energy between ocean and atmosphere by about 100 times. In addition, sea ice greatly reduces the penetration of sunlight needed for the photosynthetic processes of marine life and impedes the mixing effect of winds. A further significant distinguishing feature is the high ratio of freely connected shallow seas to deep basins. Whereas the continental shelf on the North American side of the Arctic Ocean is of normal width (about 40 miles), the Eurasian sector is hundreds of kilometres broad, with peninsulas and islands dividing it into five main marginal seas: the Chukchi, East Siberian, Laptev, Kara, and Barents. These marginal seas occupy 36 percent of the area of the Arctic Ocean, yet contain but 2 percent of its water volume. With the exception of the Mackenzie River of Canada and the Colville River of Alaska, all major rivers discharge into these shallow seas. The combination of large marginal seas, with a high ratio of exposed surface to total volume, plus large summer inputs of freshwater, greatly influences surface-water conditions in the Arctic Ocean.

Main directions of flow

The balance of the Arctic Ocean's water budget (input and output) is still very poorly known. Although the largest volume of water transport is between Greenland and Spitsbergen, this mechanism is the least understood of all. The inflow of dense bottom water from the Atlantic occurs mainly on the eastern side of the passage and is estimated to be 78 percent of the total volume brought into the Arctic Ocean. This is balanced by the surface outflow of colder but less saline waters in the form of the East Greenland Current. Very few data are available on the transport volume of this current, but, from indirect evidence, it is estimated to account for 83 percent of the total discharge from the Arctic Basin. Of this figure, 2 percent is in the form of icebergs. Although the volume of water transported as ice is a small fraction of the total, it is of considerable importance to oceanic heat balance and poses a major problem to North Atlantic shipping.

The flow of water out of the Arctic Ocean through the Canadian Arctic Archipelago is about 17 percent of the total discharge, and, compared to loss through the Greenland Sea, the transport of ice is negligible. There is essentially no northward flow of water through the shallow channels separating the islands of the Archipelago.

The volume of water entering the Arctic Ocean through the Bering Strait is better known than that from the Atlantic and is approximately 20 percent of the total. The amount of freshwater entering the Arctic Ocean is about 2 percent of the total input. Precipitation is believed to be about ten times greater than loss by evaporation, although both figures can be only roughly estimated.



ARCTIC OCEAN Depths in metres

Colours used are thought to be those of the various rocks and sediments on the sea floors. Differences in relief are shown by relief shading.

Through all of these various routes and mechanisms, the exchange rate of the Arctic Ocean is estimated to be approximately 210,000,000 cubic feet per second.

All waters of the Arctic Ocean are cold. Variations in density are thus mainly determined by changes in salinity. On the basis of density, Arctic waters have a two-layer system: a thin and less dense surface layer is separated by a strong density gradient, referred to as a pycnocline, from the main body of water, which is of quite uniform density. This pycnocline restricts convective motion and the vertical transfer of heat and salt, and hence the surface layer acts as a cap over the larger masses of warmer water below.

Major water masses

Despite this overall similarity in gross oceanographic structure, the waters of the Arctic Ocean can be classified into three major masses and one lesser mass.

(1) The water extending from the surface to a depth of about 650 feet is the most variable and heterogeneous of all that in the Arctic. This is because of the latent heat of freezing and thawing; salt-brine addition from the process of ice freezing; freshwater addition by rivers, ice melting, and precipitation; and great variations in insolation (rate of delivery of solar energy) and energy flux as a result of sea ice cover. Water temperature may vary over a range of 7° F (4° C) and salinity from 28 to 34 grams of salt per kilogram of seawater (28 to 34 ‰).

(2) Warmer Atlantic water everywhere underlies Arctic surface water from a depth of about 650 to 3,000 feet. As it cools it becomes so dense that it slips below the surface layer on entering the Arctic Basin. The temperature of this water is about 34° to 37° F (1° to 3° C) as it enters the basin but is gradually cooled so that by the time it spreads to the Beaufort Sea, it has a maximum temperature of 32.9° to 33.1° F (0.5° to 0.6° C). The salinity of the Atlantic layer varies between 34.5 and 35 ‰.

(3) Bottom water extends beneath the Atlantic layer to the ocean floor. This is colder than the Atlantic water (below 32° F or 0° C) but has the same salinity.

(4) An inflow of Pacific water can be observed in the Amerasia Basin but not in the Eurasia Basin. This warmer and fresher water mixes with colder and more saline water in the Chukchi Sea, where its density enables it to flow as a wedge between the Arctic and Atlantic waters. The Pacific water, by the time it reaches the Canada Basin, has a temperature range of 31.1° to 30.8° F (−0.5° to −0.7° C) and salinities between 31.5 and 33 ‰.

Surface-water movement

Arctic waters are driven by the wind and by density differences. The net effect of tides is unknown but could have some modifying effect on gross circulation. The motion of surface waters is best known from observations of ice drift. The most striking feature of the surface circulation pattern is the large clockwise gyre (circular motion) that covers almost the entire Amerasia Basin. Fletcher's Ice Island (T–3) made two orbits in this gyre over a 20-year period, which is some indication of the current speed. The northern extremity of the gyre bifurcates and jets out of the Greenland–Spitsbergen passage as the East Greenland Current, attaining speeds of six to 16 inches per second. Circulation of the shallow Eurasian Shelf seas seems to be a complex series of anticlockwise gyres, complicated by islands and other topographic relief.

Circulation of the deeper Atlantic water is less well-known. On entering the Eurasia Basin, the plunging Greenland Sea water appears to flow eastward along the edge of the continental margin until it fans out and enters the Amerasia Basin along a broad front over the crest of the Lomonosov Ridge. There seems to be a general anticlockwise circulation in the Eurasia Basin and a smaller clockwise gyre in the Beaufort Sea. Speeds are slow—probably less than two inches per second.

The circulation of the bottom water is unknown but can be inferred to be similar to the Atlantic layer. Measured values of dissolved oxygen show that the bottom water is well ventilated, dissolved oxygen everywhere exceeding 70 percent of saturation.

Sea ice. The Arctic Ocean's cover of sea ice suppresses wind stress and wind mixing, reflects a large proportion of incoming solar radiation, imposes an upper limit on the surface temperature, and impedes evaporation. Wind and water stresses keep the ice pack in almost continuous motion, causing the formation of cracks (leads), open ponds (polynya), and pressure ridges. Along pressure ridges the pack ice may be locally stacked high and project downward into the ocean several tens of metres.

Besides its deterrence to the exchange of energy between the ocean and the atmosphere, the formation of sea ice generates vast quantities of cold waters that help drive the circulation of the world ocean system.

Sea ice rarely forms in the open ocean below 60° N but does occur in more southerly enclosed bays, rivers, and seas. Between about 60° N and 75° N the occurrence of sea ice is seasonal, and there is usually a period of the year when the water is ice-free. Above 75° N there is a more or less permanent ice cover. Even there, however, as much as 10 percent of the area is open water due to the continual opening of leads and polynyas.

In the process of freezing, the salt in seawater is expelled as brine. The degree to which this rejection takes place increases as the rate of freezing decreases. Typically, newly formed sea ice has a salinity of 4 to 6 ‰. Even after freezing the process of purification continues but at a much slower rate. By the time ice is one year old it is sufficiently salt-free to be melted for drinking. This year-old, or older, salt-free sea ice is referred to as polar pack. It can be distinguished by its smoother, rounded surface and pale-blue colour. Younger ice is more jagged and grayer in colour. Because the hardness and strength of ice increases as the salts are expelled, polar pack is a special threat to shipping.

Longevity of the ice pack

Important evidence concerning the long-term stability of the Arctic ice pack has been found by studying the sedimentary record of the ocean floor. This research has recently shown that there is strong evidence of an uninterrupted existence of sea ice on the Arctic Ocean during at least the past 3,000,000 years. Such evidence strongly refutes those theories of the cause of ice ages that postulate rapid variability of the Arctic ice cover.

PLANT AND ANIMAL LIFE

Plankton. The chain of marine food supply is founded on floating microscopic plants, the phytoplankton. The annual productivity of these plants in the Arctic Basin is less than 10 percent of that of other ocean areas. The factor limiting this primary production is the ice pack itself, which reduces the penetration of solar energy needed for photosynthesis. Recent studies have shown, however, that Pacific water entering the Amerasia Basin is rich in detritus and nutrients. This is presumably true of Atlantic water coming through the Greenland Sea.

The next link in the food chain are zooplankton, small animals that feed on phytoplankton. Available information on the zooplankton under the permanent ice pack is even more fragmentary than for phytoplankton. But many of the Arctic zooplankton (and fishes) grow larger than their near relatives in warmer seas. This is believed to be due to their delayed sexual maturity in cold water rather than to their speed of growth.

Fish. Free swimmers, or nekton, depend upon zooplankton as their food base. In truly Arctic waters fish play a minor role in human economy compared to marine mammals. The reverse is true in the bordering seas, in which there is an admixture of warmer Atlantic or Pacific waters and in which fisheries abound. The most important fish is the Arctic char (*Salvelinus alpinus*), a member of the salmon family. It spends most of its life in freshwater and comes into the sea for a few weeks in the summer. Only the polar cod (*Boreogadus*) has been taken from the central Arctic Basin, although bottom photographs have shown at least one other species of fish and shallow sonic scattering, apparently from fish schools, has been noted during summer months.

Mammals. The large sea mammals are the most conspicuous polar wild life. All are important to the hunting natives, and some of them are of commercial interest. These include the suborder Pinnipedia, including the eared or fur seals, the hair seals, and the walrus. Fur seals do not occur as close to the pole as the other groups, reaching only the southern edge of the Arctic. The Pribi-

lof Islands in the Bering Sea are one of the main breeding areas. Hair seals include the harbour seal, the jar or ring seal, the bearded seal, the harp seal, and the hooded seal. The first three of these are widely distributed in the Arctic throughout the year and are of importance to the Eskimo. The latter two are migratory, and have been exploited commercially. The walrus is divided into two species, Atlantic and Pacific, and inhabits shallow waters, feeding on clams and other bottom fauna, which it can root up with its ivory tusks.

A wide variety of whales frequents Arctic waters, either for the entire year or during the summer months. They include sperm, killer, white, bottlenose, finback, humpback, blue, and narwhales.

Polar bears range throughout the Arctic, including the central ocean basin, throughout the year. They are frequently followed by the scavenging Arctic fox.

BIBLIOGRAPHY

Exploration: T.E. ARMSTRONG, *The Northern Sea Route: Soviet Exploitation of the North East Passage* (1952); L.H. NEATBY, *In Quest of the North West Passage* (1958); N.A. OSTENSO, *Geophysical Investigations of the Arctic Ocean Basin* (1962).

Geology and geophysics: AMERICAN ASSOCIATION OF PETROLEUM GEOLOGISTS, *Proceedings of the Second International Symposium on Arctic Geology* (1971); N.A. OSTENSO and R.J. WOLD, "Aeromagnetic Survey of the Arctic Ocean: Techniques and Interpretation," *Mar. Geophys. Res.*, 1:178–219 (1971). *Oceanography:* NATIONAL ACADEMY OF SCIENCE, *Polar Research: A Survey* (1970); ARCTIC INSTITUTE OF NORTH AMERICA, *Arctic Basin Symposium, 1962* (1963); *The Arctic Basin*, rev. ed. (1969); J.E. SATER, *Arctic Drifting Stations* (1968). Two significant recent references are Y.Y. GAKKEL and V.D. DIBNAR, "Bottom of the Arctic Ocean," in S.K. RUNCORN *et al.* (eds.), *International Dictionary of Geophysics*, vol. 1, pp. 152–165 (1967); and R.M. DEMENITSKAYA and K.L. HUNKINS, "Shape and Structure of the Arctic Ocean," in A.E. MAXWELL (ed.), *The Sea*, vol. 4, pp. 223–249 (1971).

Sea ice: ARCTIC SEA ICE CONFERENCE, *Arctic Sea Ice* (1958); E.R. POUNDER, *The Physics of Ice* (1965); ARCTIC INSTITUTE OF NORTH AMERICA, *Naval Arctic Manual ATP-17 (A), Part I* (1967).

Life: P.D. BAIRD, *The Polar World* (1964); *Arctic*, vol. 22, no. 3 (1969).

(N.A.O.)

Arctic Peoples and Cultures

Two major manners of earning a living are found among the cultures of the Arctic—that of the hunter and that of the reindeer herder. In the Western, or New World, Arctic only hunters are found and all are Eskimos (including the closely related Aleuts of the Aleutian Islands). Hunting cultures also occur in the Eastern, or Eurasian, Arctic; but more typical are reindeer herders distributed in a great arc across northern Eurasia, from the Chukchi of northeastern Siberia to the Lapps of northern Europe.

When defined as a culture area, the Arctic includes some regions that are sub-Arctic in climate. Physical geographers often place the southern boundary of the Arctic either at the Arctic Circle or at a line marking the regions where the warmest month has a mean temperature not above 50° F (18° C); in Eurasia this coincides closely with the northern limit of trees and represents a real division in vegetation, climate, and habitat. But this geographical boundary is not a cultural boundary, since most of the Eastern Arctic cultures are found in both the treeless desert-like tundra of the Arctic and the northern reaches of the swampy, coniferous forests, or taiga, of the sub-Arctic. Some groups live in the taiga and some on the tundra, and others are seasonal nomads (transhumants), wintering their reindeer in the taiga and moving them in spring and summer to graze on the tundra along the Arctic Ocean. Considered as a culture area, the Arctic is determined by relative uniformity of practice and custom, not solely by geography.

In many ways the Eskimos are the people most adapted to the Arctic, the people most specialized for Arctic living. Most Eskimos are coastal dwellers, specialized in hunting sea mammals, particularly the several species of seals, and in some places walrus and whales. Fish and birds also are important. Although there are some inland Eskimos in northern Alaska and Canada who depend especially on hunting the caribou, a wild relative of the domesticated reindeer of Eurasia, an orientation toward the sea characterizes the patterning of most Eskimo cultures. And although some Eskimos live in the forested areas of the sub-Arctic (in Labrador and southern Alaska), the characteristic Eskimo settlement is along a coast of the Arctic Ocean or along some northern strait with tundra at the back of the village. Most Eskimo villages are between latitude 60° and 72° N. The most northerly of the Eskimos—and most northerly of all the world's peoples—are the so-called Polar Eskimos of northwestern Greenland, who are found as far north as 79°.

ECOLOGY AND CULTURAL ADAPTATIONS

Character of the Arctic environment

The most demanding environmental conditions for Arctic peoples are the cold and high winds, the seasonal variations in the hours of daylight, the limited numbers of natural species, both flora and fauna, and the restricted range of raw materials for making houses and tools. Contrary to widely held impressions, the tundra dwellers are not buried in snow. Precipitation is very light in much of the Arctic (often less than four inches of water a year). The Arctic would be recognized as a desert except that the long, cold winters and short summers reduce evaporation and prevent runoff of groundwater or ice; much of the land is permanently frozen soil, or permafrost. Expressed in temperatures, the cold is more intense in the boreal forests of the interior taiga; but on the tundra high winds persist throughout the year, and in winter the resulting windchill makes exposure critical and the needs for protective housing, clothing, and equipment more demanding. Winter along the high Arctic coast brings some days with temperatures reaching −50° F (−46° C), some days considerably warmer, and a mean winter temperature of about −30° F (−34° C). In the boreal forests of both the Eastern and Western Hemispheres, temperatures of −70° F (−57° C) are not uncommon.

Arctic habitations, clothing, and equipment. For those peoples who still follow their traditional way of life, survival in such conditions requires highly functional housing and clothing. The provisions to meet these requirements are particularly revealing. The construction materials have to be those readily available in the immediate vicinity. Some Eskimos use snow blocks to build the domed house widely identified as an igloo; but most Eskimos (before the introduction of prefabricated structures in the Arctic) lived in earth- or sod-walled structures built on a frame of driftwood, sometimes with large whale bones as structural members. Many of the design features to offset cold and wind are common to both the snowhouse and the semi-subterranean sod house: both include an entrance shelter or storm shed, frequently long and narrow, to block cold winds, and a raised bench ("sleeping bench") to keep the occupants up out of the colder levels of the house. The sleeping bench is the centre for many activities other than sleeping during the long winter nights, activities such as eating, making tools, and mending equipment. In summer, the Eskimos move out onto the tundra, living in skin tents.

Types of housing

For the nomadic or semi-nomadic reindeer herder of the Eurasian Arctic, the winter habitation is more apt to be a tent, in many cases a double-skin tent (subterranean houses are possible only among some taiga pastoralists who can winter their herds in a sheltered area and keep them in one place for several weeks or more). The Eurasian tent, which is used both summer and winter, is conical (somewhat resembling the tepee of the North American Plains Indians) and is covered with reindeer skins on the tundra and with bark in the sub-Arctic forests. The winter tent is sometimes banked with sod and often, if not always, with snow. Some are fitted with raised benches and with storm sheds. The main tent surface may be doubled with insulating materials between the layers; in other cases, as among the Chukchi and Koryaks, there may be separate boxlike sleeping shelters within the main canopy, each for several occupants. To-

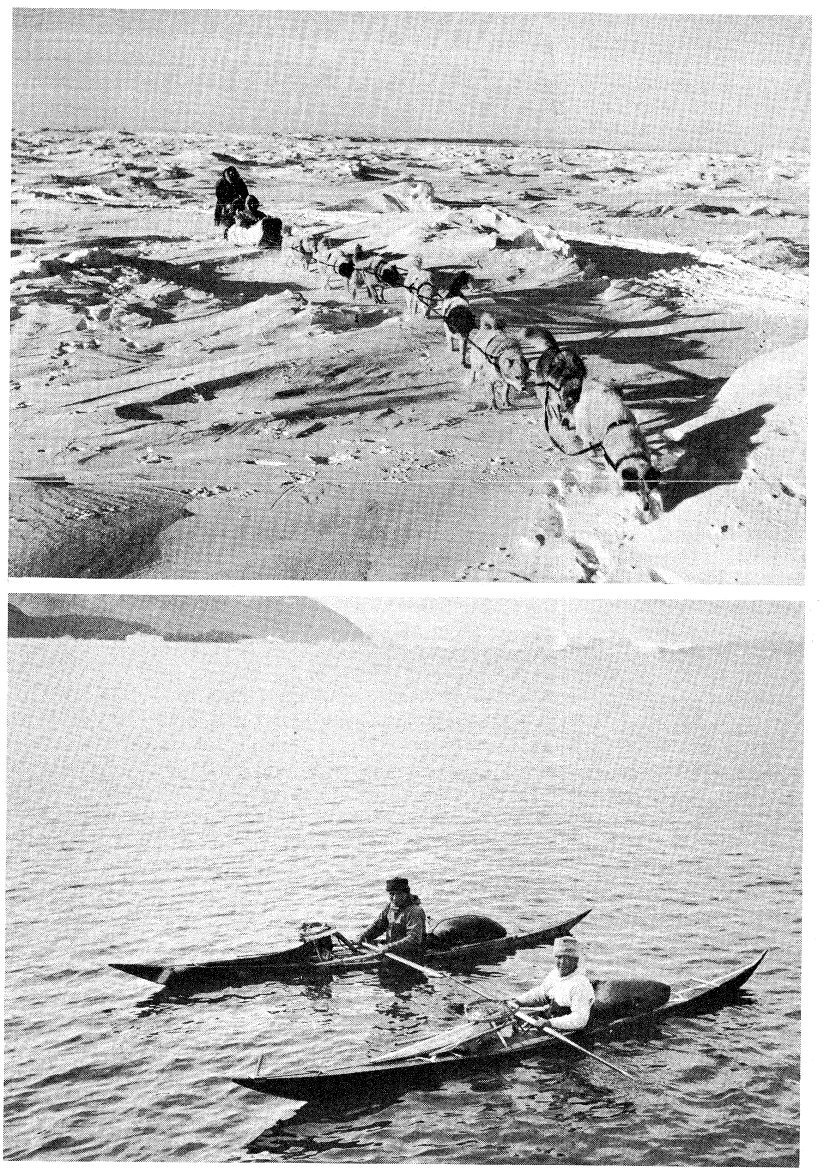
ward the western end of the Eurasian area, approaching northern Europe, the winter house is more akin to European prototypes, such as the Scandinavian log house, usually with a flat roof that is earth-covered; some Lapp herders of Scandinavia, however, spend winters in conical tents banked with sod and snow. The Arctic Eurasians who do not depend on reindeer but are sea-mammal hunters and fisher folk, and thus largely sedentary, live in semi-subterranean winter houses.

Arctic clothing

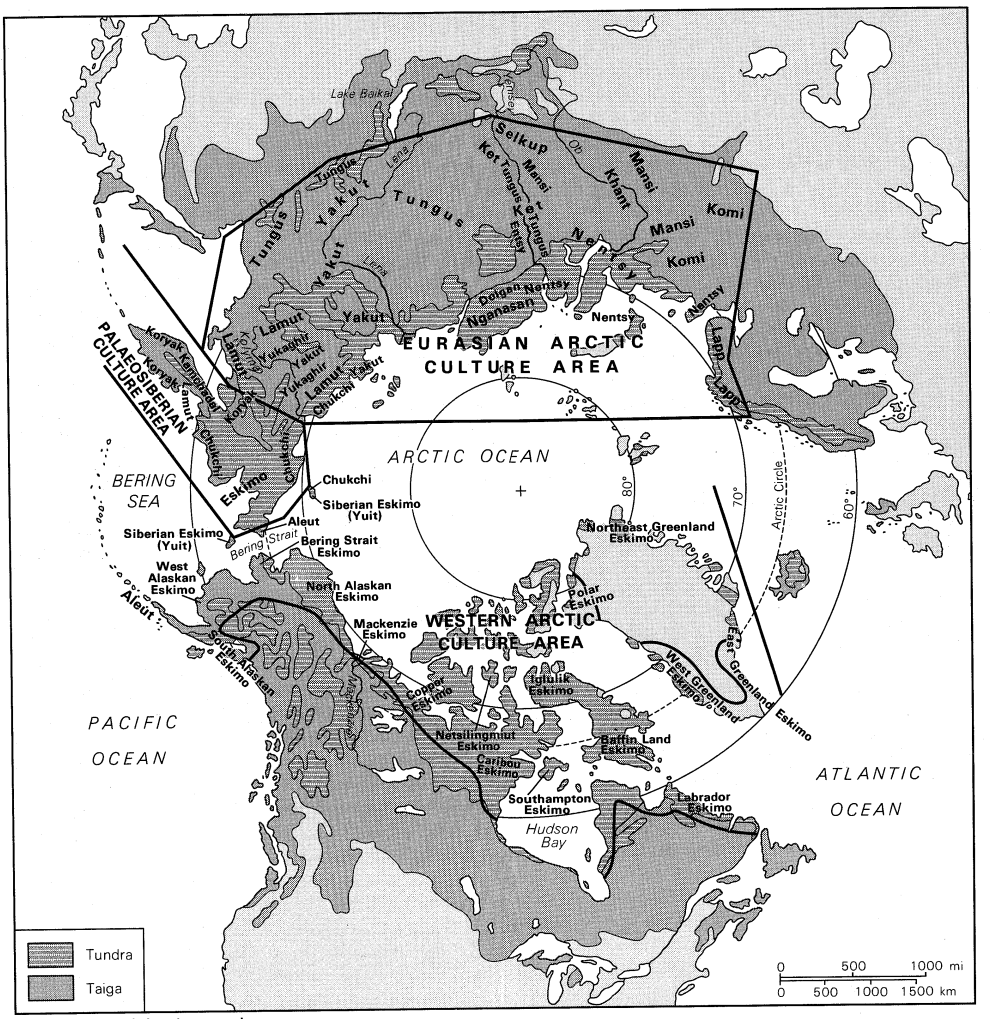
Clothing is no less critical (and perhaps more critical) than housing for survival in the Arctic. Eskimo clothing, similar in general to clothing in the Eurasian Arctic cultures as well, is so suited to the environmental conditions that until recently none made elsewhere could equal it, and explorers and others from the outside world adopted it after a short period in an Arctic winter. The general aspects of Eskimo clothing are well-known—garments fashioned of the skins of fur-bearing animals with emphasis on a pullover jacket with an attached hood (parka) and bootlike footgear with watertight seams. The parka and trousers (worn by both sexes) are double-skin garments—two layers of skins placed back to back with the fur surface of the outer layer on the outside and the fur or hair of the inner layer inside, against the wearer's skin. Caribou skins with the hairs at just the right stage of development are selected. Not only does the hair of the inner layer provide warmth, but air trapped between the short, fine hairs of the inner caribou skins provides additional insulation. Other functional aspects of design provide for ventilation and perspiration control. Even at temperatures well below zero, a man exerting considerable effort, as in running after a dog sledge, often perspires. If he is wearing conventional undergarments, these inner garments become damp with perspiration; and when the activity rate is decreased and the body cools, the perspiration-dampened clothing freezes. In the fully developed Eskimo garments, drawstrings are provided at the end of each sleeve and at the face opening. A belt closes the lower edge of the parka. If the wearer perspires, he loosens these bindings, permitting the outside air to percolate through the hairs of the inner fur layer. The cold, dry, ambient air, warmed by passage over the man's body, becomes extremely dry and quickly absorbs the moisture. With the perspiration stopped, the wearer pulls the bindings tight again to prevent overcooling. Other items of Arctic clothing include variant footgear for differing weather, outer rain garments to keep the exterior fur layer dry, and mittens, each item with design features providing similar functional fit and solving particular environmental problems.

Technology and equipment

It has become customary to state that Arctic dwellers were able to occupy their difficult environment only through application of much ingenuity. Other references credit the Eskimos, for instance, with the invention of the toggle-headed harpoon (one whose spurred head detaches from the shaft after piercing a struggling animal); the semi-subterranean house; the composite bow (one of wood, bone, and sinew glued and lashed together); the



Modes of Eskimo transportation. (Top) Copper Eskimos drive a dog sledge near Cambridge Bay, Victoria Island, Canadian Arctic Archipelago, Northwest Territories. (Bottom) West Greenland Eskimos in kayaks on Baffin Bay, off Greenland.
(Top) National Film Board of Canada, bottom) Rutherford Platt



Distribution of Arctic peoples.

throwing stick, or atlatl (a rodlike device for hurling darts or spears); and other items of material culture. In terms of skill in manufacturing and eagerness to adopt new items, Arctic peoples were indeed ingenious; but this does not mean that they invented all of the items that they used. With the exception of the domed snowhouse, a trait exclusive to the Eskimos, even the most typical items in the material culture of Arctic peoples were widely shared with other peoples in more southerly areas.

A list of material items used in both the Eurasian and Western Arctic would include the semi-subterranean house with entranceway and sleeping bench; tailored clothing, including the parka; the toggling head and other harpoon parts; blunt arrowheads for birds and small game; the atlatl; barbed darts for taking birds; the semicircular slate knife (ulu); snow goggles with slits as eye openings; sledge forms; dog teams for traction; skin boats both large and small (kayak); the bow drill; the composite bow; ice picks for harpoon shafts; the tambourine-type drum used in healing rites; and so on. The semi-subterranean house is shared with American Indians south of the Eskimo area and distributed across the Eurasian Arctic. The hooded parka was used as far west as northern Europe and can be identified in art objects far back in time (as in the Siberian Neolithic Period). The toggling harpoon head was used in hunting sea mammals by some prehistoric and later cultures of northern Europe.

Arctic livelihoods. Although the inroads of modern civilization have considerably altered the modes of subsistence in certain areas (see below *Modern populations in the Arctic*), a substantial number of Arctic peoples still cling to their traditional culture patterns and obtain their livelihood, as previously noted, primarily by hunting and herding. The Eskimos have traditionally organized their lives for hunting the mammals of the sea (four or five species of seals and, in certain regions, beluga, walrus, and whales), the mammals of the land (principally caribou but also polar bears, hares, and marmots), and the fish of seas and streams. Thus, in winter, a dozen or more families congregate in a small community from which the men go forth each morning to harpoon the seals that expose their heads in lanes of open water or rise to breathe at the small holes that they keep open for themselves in the sea ice. During the long days of spring these communities break up. Some families pursue the seals over the open water in kayaks; others hunt whales; and still others wander off to fish through the ice of the still-frozen lakes or to trap the fish of streams and rivers. Later, in summer, most families hunt the caribou when the fur of that animal is in its prime; and some carry on this hunt into the late autumn when the herds are migrating to new southerly pastures. Then, with the approach of winter, the families draw together again, bringing the yearly cycle to a close. This cycle varies somewhat, however, from region to region.

The seasonal nature of Arctic life

Compared to the Western Arctic, the patterns of ecological adaptation in the Eurasian Arctic have been more varied, influenced by non-Arctic cultures of greater variety and higher development, by differences in natural environment (particularly differences in ocean conditions), and by the presence of domestic reindeer. Although there are some sedentary hunters and fishers to be found, the reindeer in fact has set the ecological tone for the Eurasian Arctic. Many of the hunters use deer for traction, riding, and packing; and for many groups (such as the Reindeer Chukchi, Reindeer Koryaks, many Nentsy, Entsy, and Lapps) travelling with the herds is a way of life. One of life's realities for a culture depending on reindeer is that the herd moves many miles each day through much of the year. The Reindeer Chuk-

chi and Koryaks move camp 20 or more miles each day as the herd forages for food or, in summer, flees from insect pests (particularly mosquitoes and biting flies). Except in a few cases in which a small number of deer are kept in pens, mobility is a prime condition for all Eurasian reindeer keepers. As an alternative for moving the entire social group with the herd, special herdsmen may be assigned to look after the animals, a solution most prevalent toward the western end of Eurasia; but even among the Lapps and the herders among the Komi, entire groups move with the reindeer during the seasonal movements from wintering areas to summer pastures. There are, of course, great variations in patterns of reindeer keeping and some variations in types of animals. Chukchi and Koryak herds sometimes number 10,000 or 12,000 reindeer, whereas Tungus or Yakut herds are not apt to number more than 20 or 25. The Chukchi reindeer are too small to be ridden; and the half-wild animals are herded by men on sleds drawn by reindeer; Tungus deer are larger and can be used as riding and pack animals. Most Tungus acquire their food largely from hunting wild reindeer and other game, whereas Lapps depend on their herded reindeer for milk, meat, and skins.

ARCTIC ETHNIC GROUPINGS

Social organization among Eskimos

Western Arctic. There are no tribes of Eskimos in the sense of social groupings with internal political organization or even in the sense of populations conscious of their own tribal identities as opposed to other tribal groups. The Eskimos are conscious of being Eskimos (*iniut*, or "real men"), in contrast to non-Eskimos, and particularly in contrast to American Indians, who are their nearest neighbours; but this is a general identification and involves no elements of structure or organization. The only Eskimo social units with validity from the Eskimo point of view are the immediate family, the household (often made up of two families), and the local residence group, which is identified by a designation for the location, or place, with the suffix "miut," meaning the people of that place. Thus, the Birnirkmiut are those people who dwell at Birnirk (a location in the bend of a lagoon). The designators for larger groupings that have become standard in Eskimo studies are conventions introduced by field workers and scholars (see the accompanying map for the location of these conventionally named groups). Some names, such as Iglulik, Netsilingmiut, and Caribou Eskimos, refer to a particular subsistence characteristic (*iglu*, or "igloo"; *net*, or "seal"; caribou). Others have territorial referents, such as the North Alaskan Eskimo, Bering Strait Eskimo, and South Alaskan Eskimo. Whether in the first or the second category these conventionalized groupings have become established because they reflect various ecological adaptations within the general Eskimo way of life. The major groupings are as follows:

Major Eskimo groupings

Siberian Eskimo (Yuit), including St. Lawrence Island Eskimo, hunters of walrus, whales, and seals, using the whaleboat (umiak) for whaling and voyages.

Aleut, fishers and hunters of seals, sea otters (for furs), and other sea mammals, using kayaks with two cockpits.

South Alaskan Eskimo, fishers (chiefly of salmon) and coastal dwellers in a forested area, using kayaks.

West Alaskan Eskimo (Nunivak Island and adjacent coast, Norton Sound), hunters of seals, walrus, and seabirds and fishers of salmon and other fish.

Bering Strait Eskimo (including Diomede Island and Cape Prince of Wales), hunters of walrus, whales, and seals, using whaleboats (umiak).

North Alaskan Eskimo (Point Hope, Point Barrow, and the North Alaskan Interior Eskimo), primarily whalers but also hunters of walrus, seals, waterfowl, and caribou.

Mackenzie Eskimo, hunters of whales and seals and fishers of salmon.

Copper Eskimo (including Victoria Island) and Netsilingmiut Eskimo (including Boothia Peninsula), hunters of seals (in winter, at breathing holes, and, in summer, from kayaks in open water) and of caribou.

Iglulik Eskimo, hunters of seals, walrus, and caribou.

Baffin Land Eskimo, hunters of seals and sea fowl.

Caribou Eskimo (Barren Grounds, Chesterfield Inlet, Back River), primarily hunters of caribou, but also of musk ox and, rarely, sea mammals, using no boats.

Southhampton Island Eskimo (extinct since 1904), hunters of caribou and seals, using no boats.

Labrador Eskimo, hunters of seals (from the ice edge and in open water from the kayak) and of caribou, and dwellers in or near sub-Arctic forests.

West Greenland Eskimo, hunters of seals, using kayaks.

East Greenland Eskimo (Angmagsalingmuit), hunters of seals, using kayaks.

Polar Eskimo (northwest Greenland), hunters of seals, walrus, and seabirds, using no boats.

There is general agreement among specialists in studies of the indigenous peoples of North America that all of the Eskimo area, including the Aleutians, makes up one culture area in contrast to the various American Indian culture areas south of the Arctic. There also is considerable agreement among specialists that the Eskimo, or Western Arctic, culture area is divided into two major subcategories: a central and eastern subculture area, which extends across the northern Canadian archipelago and includes Greenland and the coast of Labrador, and a western subculture area, which coincides closely with the Alaskan regions, including the Aleuts and also the Siberian Eskimo (Yuit) on the northeast tip of Siberia. Various traits have different forms in the two subregions. Dog teams, for instance, are hitched differently. In the western subregion the dogs are attached in tandem along a central trace (a mode of hitching also used in the Eurasian Arctic). In the central and eastern subregion, dogs in a team are attached to separate traces of varying lengths and run in a fan formation. Kayaks, harpoon heads, house design, and other things show similar contrasts. There are a series of traits, not geared to mere survival but to more sophisticated living, that are found in the western area but not in the central and eastern region. Among these are labrets (ornaments worn in the perforated lip), dance masks, coiled basketry, pottery, grave monuments, mourning feasts, property distributions, war parties, and kin groups larger than the family. Collectively, these make the western cultures more elaborate or complex than those of the central and eastern region. One interpretation is that the western elaboration represents traits inherited from ancestral Asian cultures and that many such traits were sloughed off as the Eskimos migrated eastward and had to cope with a harsher environment that left them little opportunity for partly superfluous traits and practices.

Major Eskimo subculture areas

Some studies further divide the central and eastern areas into separate categories. The people in the central area are largely nomads who dwell in snowhouses; the people in the east, particularly in Greenland, live in semi-subterranean houses in fixed villages of considerable size. It is the nomad of the central groups (especially Iglulik, Netsilingmiut, Copper, and Caribou Eskimo) who provides the popular stereotype of Eskimo life; but actually most Eskimos, east and west, are not nomads living in snow winter houses but instead are sedentary village dwellers living in subterranean sod-walled houses. Village living is made possible for the Eskimos by their efficient transportation—dog teams in winter and skin boats (kayak and umiak) in summer. Only hunters with good transportation can range far enough to gain sufficient food for sedentary village living.

Eastern, or Eurasian, Arctic. Although many hunters and fishermen are found in Arctic Eurasia, it is the keeping of domestic reindeer that defines the culture area and gives it unity. The herds and correlated traits, in harmony with the required mobility, set the Eurasian Arctic apart from the Western Arctic, despite many common circumpolar traits. The reindeer complex also differentiates Arctic from non-Arctic cultures in Asia; but some boundaries are blurred, particularly in southern Siberia, by transitions between the herding of reindeer and the herding of other domestic animals, such as cattle, or by transitions to agriculture within single ethnic groups.

The peoples of the Eurasian Arctic more genuinely divide into different ethnic groups (in contrast to the Eski-

mo groupings); each group has its own language (though many are linguistically related), and each considers itself as having its own name and tradition. The major Eurasian Arctic groups number 19 and go by various names (the alternative names appear below in parentheses):

Major Eurasian Arctic ethnic groups

Lapp (Saamian, Lopari), reindeer herders and fishers, speaking a West Finnic language (of the Finno-Ugric family, a subdivision of Ural-Altaic).

Komi (Zyryan), various groups of reindeer herders, cattle herders, hunters, and agriculturalists, speaking a Permian language (of the Finno-Ugric family).

Mansi (Vogul), reindeer herders, hunters, and fishers, speaking an Ugrian language (of the Finno-Ugric family).

Khant (Ostyak), reindeer herders, hunters, and fishers, speaking an Ugrian language (of the Finno-Ugric family).

Selkup (Ostyak-Samoyed), reindeer herders, hunters, and fishers, speaking Selkup (of the Samoyed family, a subdivision of Ural-Altaic).

Entsy (Yenisey-Samoyed), reindeer herders, hunters, and fishers, speaking Entsy (of the Samoyed family).

Nganasan (Tavgi Samoyed), reindeer herders, hunters, and fishers, speaking Nganasan (of the Samoyed family).

Nentsy (Samoyed, Yurak), reindeer herders, hunters, and fishers, speaking Nentsy (of the Samoyed family).

Yakut (northern Yakut: Sakha, Jeko), chiefly cattle and horse herders but also some reindeer herders, hunters, and agriculturalists, all speaking a Turkic language (of the Altaic family, a subdivision of Ural-Altaic).

Dolgan, reindeer herders, hunters, and fishers, comprising ethnically a Tungus group but now speaking the language of the Yakut.

Lamut (Even, Oven), reindeer herders, hunters, and fishers, speaking a Manchu-Tungus language (a subdivision of Ural-Altaic).

Tungus (Evenk), reindeer herders, hunters, and fishers, speaking a Manchu-Tungus language (a subdivision of Ural-Altaic).

Yukaghir (Odul), a Paleosiberian group, formerly primitive hunters but now largely reindeer herders, speaking Yukaghir (an unaffiliated language).

Koryak (Nymylan), a Paleosiberian group of reindeer herders and hunters and of sea hunters and fishers, speaking Koryak (of the Chukchi-Koryak-Kamchadal family).

Chukchi (Chukchee, Luorawetlan), a Paleosiberian group of reindeer herders and hunters and of sea hunters and fishers, speaking Chukchi (of the Chukchi-Koryak-Kamchadal family).

Kamchadal (Itelmen), a Paleosiberian group of sea hunters and fishers, speaking Kamchadal (of the Chukchi-Koryak-Kamchadal family).

Eskimo (Yuit), Asiatic Eskimo sea hunters and fishers, speaking Eskimoan.

Aleut (Unangan), Asiatic Aleut sea hunters and fishers, speaking Aleut.

Ket (Yeniseian, Yenisey-Ostyak), western Paleosiberian hunters, speaking Ket (an unaffiliated language).

At the present time there is no consensus on how to divide Arctic Eurasia into subculture areas comparable to those customarily assigned to the Western Arctic. The coastal regions of northern Scandinavia (particularly northern Norway) might be considered separate from the remainder of the Arctic culture area; even though the coasts reach far north of the Arctic Circle, they are warmed by the North Atlantic Drift (Gulf Stream) and inhabited by villages of sedentary fishermen who live more like other Scandinavians than like any of the other peoples of the Eurasian Arctic. This frequently is not indicated in culture area charts because of mapping problems in separating such a narrow region from the interior area of Lapp herders.

Paleosiberian cultures

More frequently, the northeastern extremes of Siberia are designated as a separate culture area, labelled Paleosiberian. The prefix "paleo" implies that the peoples are remnants of earlier, more extensive populations who have been pushed into northeastern Siberia by incoming peoples of later times (Neosiberians). Paleosiberian, when used as a culture area, primarily includes the Chukchi, Koryaks, and Kamchadals but, when used as a linguistic category (lumping together those Siberian languages that do not have close relationships with other known languages), also includes the Yukaghirs and the Kets to the west. The Kamchadals are atypical of Arctic peoples, being largely sedentary fishermen in a maritime, temperate environment. They are included because they live interspersed with Koryak settlements. (A general problem for culture area definitions in Arctic Eurasia is that groups are discontinuous and interspersed—typical results of physical mobility associated with pastoralism.)

The distinction between the Paleosiberian area and the remainder of the Eurasian Arctic area is buttressed by differences in methods of caring for the reindeer, differences in uses for deer, and differences in the types of reindeer kept. Chukchi and Koryak deer are small, half-wild, and kept in large herds by herders who use them to drive sledges. The deer are neither ridden, nor used as pack animals, nor (customarily) milked. In the greater Eurasian Arctic culture area there are variations in many of these factors, but in general the deer are tamer (more domesticated) and are herded by men riding reindeer and aided by dogs. Another cultural factor setting the Paleosiberian area apart is the presence of maritime varieties of the ethnic groups (Maritime Chukchi, Maritime Koryaks), who are sea-mammal hunters living in a manner similar to the Bering Strait and Siberian Eskimos.

Finally, another boundary for subregions, as suggested in some published sources, may be located in the vicinity of the Ural Mountains, with the Komi and Lapp herders of Europe, to the west, separated and differentiated from the reindeer herders of Asia, to the east. Considered together, the Lapps and Komi, who speak Finno-Ugric languages, contrast with and are culturally distinct from the Khants, the Mansi, and the Samoyedic speakers.

PHYSICAL TYPES AMONG ARCTIC PEOPLES

Two of the major varieties of mankind are represented among the Arctic peoples—Mongoloid (Asian) and Caucasoid (European). The idea that mankind varies in continuous distributions of physical characteristics, which should be plotted in terms of frequency occurrences (clines) of these traits rather than broken into discrete categories (races and subraces), is demonstrated in a general way among the Arctic populations. There is more or less a cline from Mongoloid to Caucasoid extending from Arctic America and eastern Siberia across northern Asia and into northern Europe, with the more generally Mongoloid nearer to the American and Siberian end and the more generally Caucasoid nearer to the European end. Comparative studies of the ethnic populations reveal, however, that the distribution of traits is not strictly continuous; some groups not fitting the pattern intrude here and there, and the cline extremes are not truly at the geographical ends of the distribution.

Biology of major groups. In regard to physical type, the Eskimos are much more closely related to northern Asians than they are to American Indians, as shown by blood types as well as by morphological features (principally facial features and body proportions). Sources variously refer to an "Eskimo" subrace of Mongoloids or to an "Arctic" subrace. When "Arctic Mongoloid" is used as a category, the Chukchi and some other groups of northeastern Siberia are included; however, field workers report that the Chukchi and their neighbours, the Siberian Eskimos, differ enough so that most individuals can be identified without reference to language or other cultural traits; and early sources reported that skulls unearthed from ancient tundra burials showed considerable differences, with Eskimo crania being higher and Eskimo facial breadths greater than those for the Chukchi.

The Tungus and Yakuts

West of the Arctic Mongoloids, the Tungus and Yakuts show some of the highest frequencies of extreme Mongoloid traits found anywhere. They are among the most Mongoloid of all peoples, with low-bridged noses and forward-projecting cheekbones (malars) that are covered with pads of fatty tissue, resulting in extreme facial flattening. Eye forms show a high frequency of epicanthic (or "Mongoloid") eye folds. Hair is straight, black, and large in cross section; it is abundant and long on the head but sparse in beards and body hair. Skin colour is yellow

or yellow brown, and eye colour is consistently brown. In stature the Tungus and Yakuts are short, with the shortness accentuated in lower limbs. Legs below the knee (tibia and fibula) are short relative to the legs above the knee (femur); and elbow to hand lengths (radius and ulna) are short relative to upper arms (humerus). Resulting total limb lengths are short relative to trunk lengths. The type is labelled "classic" Mongoloid or the "extreme" Mongoloid. Broad heads, relative to head lengths (brachycephalic), are also stated to be characteristic of extreme Mongoloids in most discussions of racial types; but this ratio depends on a combination of factors, and changes through time occur within ethnic populations. Many Tungus are longheaded (dolichocephalic), and in recent times there has been a trend in all northeastern Asian Mongoloids toward an intermediate head form (mesocephalic).

The Tungus and Yakuts have been frequently cited as extreme Mongoloids who are placed geographically between the Caucasoids of the Urals and the Caucasoid-like Ainu of the northern Japanese islands; moreover, the Tungus and Yakuts are equally intrusive between the less Mongoloid Arctic peoples who are their immediate neighbours, particularly the Yukaghirs and the Kets. The physical characteristics of these various groups and their distributions are consistent with the hypothesis that the extreme Mongoloids derived from a people who originally had more generalized Mongoloid characteristics (and were more like Caucasoids) and who, isolated in Siberia during a period of extreme cold in the Ice Age, developed the distinctive facial and body characteristics of the extreme Mongoloid as adaptations to the glacial environment. Prior to the development of the extreme Mongoloid, some people of the generalized Mongoloid type had migrated across the Bering Strait area to become ancestors of the American Indians, while other generalized Mongoloids remained in scattered populations elsewhere in Asia. After the ice receded, the physical characteristics of the extreme Mongoloids, though adaptive to very cold conditions, were no handicap in warmer environments and were continued and spread by migrations; but the centre of highest frequency for extreme Mongoloid traits remained near the point of origin. Seen relative to this hypothesis, the Tungus and Yakuts are representative of the extreme Mongoloid type, and the Yukaghirs and Kets are near the generalized Mongoloid in physical type. In appearance the Kets are suggestive of some American Indians. Old photographs of the Yukaghirs show Caucasoid-like facial features, but this is not inconsistent with the idea of a generalized Mongoloid type.

West of the Yenisey River, the Khants and Mansi are basically Caucasoids (as are the Komi and Lapps farther to the west), although the frequency of blood type B is somewhat higher (a Mongoloid trait), and individuals with epicanthic eye folds are more common than in European populations. The Khants, Mansi, and Komi are usually classified as belonging to a "Ural" subrace of Caucasoids. (Soviet sources refer to a "Europoid" type.) The Ural type is characterized by straight, brown or light-brown hair, fine in cross section; light skin colour; mostly light or mixed eyes; moderately thick beards; comparatively short and moderately wide faces; concave or upturned noses; thin lips; medium ratio of head breadth to head length (mesocephaly); and short stature (with a mean of about five feet, three inches [160 centimetres] for adult males).

Compared to their neighbours, the Khants and Mansi, the Samoyedic speakers—Nentsy, Nganasans, Entsy, and Selkups—have a higher frequency of Mongoloid features, including particularly the epicanthic eye fold, and they are often identified as Mongoloid outliers in Europe and western Siberia. Compared with the extreme Mongoloids, however, they are clearly closer to Caucasoids in many features. They lack pronounced malars and have higher nasal profiles, and their facial proportions are more comparable to Europeans than to Asian types. Clearly, they are intermediate in the continuum found among Arctic peoples.

The Lapps of the Scandinavian countries and northwestern Russia are now recognized as Caucasoids. Formerly, they were not always so classified, probably mainly due to the marked contrast with their Indo-European-speaking Scandinavian neighbours, who show a high proportion of tall Nordic types. Whereas the Scandinavians, particularly the people of Sweden, are the tallest of Caucasoids and among the tallest groups of mankind, the Lapps are among the shortest, with a mean stature of only five feet, three inches (160 centimetres) for males and four feet, ten inches (147 centimetres) for adult females. Some sources formerly classified the Lapps as Mongoloids, but they lack most sorting criteria for Mongoloids, having only a moderate incidence of B-type blood (an Asian trait), shovel-shaped incisors (also Asian), and other dental characteristics to suggest this affiliation. Some authorities have even assigned the Lapps to their own individual major racial category, thinking of them as a remnant of some archaic Caucasoid-like people. In most traits, however, they are close to the Ural type, and systematic assessment places them among the Caucasoids. Their Mongoloid traits are terminal effects of the Arctic Mongoloid-Caucasoid cline.

Caucasoid features of Lapps

In most physical characteristics, the Eskimos are close to the extreme Mongoloid type. They also share the general characteristics of short stature, thick torso, and small extremities found among all the Arctic peoples. Uniquely Eskimoid features include very broad faces, thin noses, and high cranial vaults with sloping sides ("gabled" parietals). Eastern Eskimos and the crania from old Eskimo archaeological sites are longheaded (dolichocephalic), and for this reason many summaries characterize the Eskimos as the longheaded subtype among Mongoloids. Not all Eskimos, however, are dolichocephalic. Group means vary from dolichocephaly through mesocephaly to the lower range of brachycephaly. Anthropometric studies also report variations in stature from group to group (from about five feet, two inches, to five feet, six inches, for adult males).

General adaptive features. Considered as a whole, the Arctic peoples share certain physical features regardless of the predominance of Mongoloid or Caucasoid traits. The most evident of their shared traits is their small size —or, more correctly, short stature, since such people as the Chukchi are described as being massive in build, though only two or three inches above five feet in mean stature for adult males. Other shared characteristics include large girths relative to body lengths, short limbs relative to torso lengths, short lower legs and lower arms relative to upper limb segments, and small hands and feet. All of these characteristics are in harmony with general biological principles regarding adaptation to cold environments. Experimental studies, particularly since World War II, have identified at least two specific cold-adaptive body mechanisms (largely vascular) that have become hereditary among the Arctic populations. One of these results in increased blood flow in exposed body parts, such as the hands, and heightened basal metabolism to keep blood temperatures within tolerable limits (found among Eskimos and other Mongoloids in the Arctic). The second form of cold-adaptation involves shunting of blood from one artery to a paired artery and warming of venous blood leaving the body extremities before it returns to the heart. Thus, the extremities can be reduced in temperature without affecting more critical body parts such as the brain. (This second cold-adaptive characteristic has been found among the Lapps.) Even before these cold-adaptive features were identified by experimental physiology, they were acknowledged by ethnographers in statements asserting that the Chukchi men could work for considerable periods with bare hands when temperatures were −30° F (−35° C), or that the Chukchi would strip to the waist and complain of the heat if the temperature rose much above 40° F (4° C).

Cold-adaptive body mechanisms

ORIGINS AND HISTORY OF ARCTIC PEOPLES

Eurasian Arctic prehistory. The polar regions were not populated during much of the Pleistocene Epoch (from 2,000,000 to 12,000 years ago), when most of man's prehistory was unfolding. Man began as a tropical

Lateness of Arctic settlement

animal, and the Arctic was the last area (except for the Antarctic, only now being explored) to become a habitation area. Areas just south of the Arctic, which, in Pleistocene times, had climates similar to the Arctic, do show significant early developments. At archaeological sites on the Upper Yenisey River and at Lake Baikal and at other locations across the steppes into European Russia, for instance, there have been found the world's oldest known houses—semi-subterranean dwellings with central fire pits, probably ancestral to the earth dwellings of the Eskimos and other Arctic peoples. (For such a functional trait, the possibility of independent invention in different times and places must be considered.) Associated with these finds were flint points and scrapers, fine blade tools, and bone, antler, and ivory tools.

The Lake Baikal region also supplies evidence for later developments (this time supplemented by some discoveries within the Arctic itself). Around Lake Baikal are to be found a number of stages of development. The earliest of these stages, beginning between 5000 and 4000 BC, is marked by a hunting culture using the bow and arrow. In the second stage, after 4000 BC, pottery and ground-stone tools are present, and in the third, beginning in the 3rd millennium BC, reinforced bows (wooden with bone plates), finely worked implements of bone and ground stone, and effigies of animals (especially fish) are suggestive of artifacts found in later periods all across the Arctic. These materials and additions from later periods in Siberia are reflected in old cultures around the Bering Strait and in the American Arctic.

The Arctic sites contain no evidences of cultivated plants and no evidence of domestic animals other than the dog, so the use of "Neolithic," or New Stone Age, to describe the sites is based on the presence of other Neolithic traits such as ground stone implements and pottery. Even reindeer seem to have been hunted in the wild only, until later times when they were substitued for other domestic animals by groups moving northward, who had kept horses and cattle in warmer regions.

Prehistorical migrations in the Eurasian Arctic

The origin and movements of the Eurasian Arctic peoples are not clear and are subject to scholarly conjecture and controversy. In general, however, rather recent movements from south to north are probable. The Chukchi, Koryaks, and Kamchadals are believed to have entered northeastern Siberia about 2,000 years ago, finding the ancestors of the Eskimos already living along the coasts. Because the Yakuts and Tungus were still expanding into the northlands after the Russians arrived, it was formerly held that they had migrated there only recently; but the archaeological evidence renders this questionable. Some sources refer to movements of the Yakuts from south to north and of the Tungus from north to south; however, the most tenable hypothesis at present suggests that both the Yakuts and the Tungus developed from groups associated with the Baikal region (or from groups further to the south and east), with the Yakuts expanding north and west over a long period, arriving in the middle Lena by the 17th century and assimilating earlier Arctic residents along the way, while the Tungus expanded both north and south.

Among those assimilated and replaced were many Yukaghirs, who are thought by Soviet prehistorians to have been in the area from about 4000 to 3000 BC and formerly to have extended westward to the Yenisey. The Kets also are held by some scholars to be remnants of a cluster of groups who were early Arctic dwellers, although some Ket traits suggest recent south-to-north movements along the Yenisey. For the Samoyedic speakers (Nentsy, Nganasans, Entsy, and Selkups), different categories of data provide conflicting conclusions. Linguistic distributions and other language studies suggest that they once lived west of the Urals in the Pechora Basin (now occupied by the Komi and Russians). Their physical type, however, suggests Asian origins. A tenable hypothesis held by some scholars is that a Caucasoid people speaking a Samoyedic language migrated eastward through the Urals about the time of Christ and intermixed with Mongoloids who had moved westward into northwestern Siberia. Beginning about AD 1000, the Khants and Mansi, under pressure from the Komi, followed the same migration route, completing their movement across the Urals about 1400. The Komi, themselves under pressure from the Slavs, had moved northward and arrived in their present area about 800 years ago.

More study by linguists, physical anthropologists, ethnologists, and archaeologists has been devoted to the origins of the Lapps than to any other Arctic people (with the possible exception of the Eskimos)—but with no certain results. Some scholars conclude that the Lapps originated in central Europe, where they were adapted to an Alpine habitat. Others consider them to be an ancient Arctic people who adapted to polar conditions during the Pleistocene. So far, archaeological studies of their shifting camps have shed little light on the question. The most developed hypothesis maintains that the modern Lapps are descendants of a people of Caucasoid physical type who were Stone Age hunters living south and east of their present area and outside the present Arctic but in an area of similar environmental conditions. When the Pleistocene ice retreated, they moved north and west and were the earliest Finnic-speaking people in northern Europe. They adopted reindeer herding according to patterns of animal domestication used for cattle by Europeans further south, augmented with ideas of reindeer herding diffused across the Urals from Asia.

Western Arctic prehistory. Two general problems are dominant in archaeological studies in the American Arctic. These are (1) the origin of the Eskimo, and (2) the migrations of the ancestral American Indians as they crossed the Bering Strait region from Asia and moved through Arctic America to spread out in other parts of the continent.

Origins of the Eskimo

Two major theories have dominated discussions of Eskimo origins: (1) that the Eskimos evolved from a northern group of American Indians who began to emphasize their winter living patterns and became specialized to tundra dwelling and sea-mammal hunting, or (2) that the Eskimos adopted their way of life in Asia, later migrating to fill a northern void in the New World at a time much later than when most of the ancestors of the American Indians left Asia. In the latter hypothesis, the Asian Eskimos (Yuit) are seen as a remnant along the migration route. In the former hypothesis, the Siberian Eskimos are regarded as a late backwash in counterdirection to the major migrations. Archaeological discoveries, almost from the beginning, have seemed to support the Asian origin, for the earliest sites clearly in the Eskimo tradition have been found around the Bering Strait (Okvik-Old Bering Sea) and in Siberia (Uellen).

There is little evidence of man in the American Arctic in the period between 25,000 and 10,000 years ago, and there are only a few scattered artifacts for the period between about 8000 and 5000 BC, even though the ancestors of the American Indians passed through the Alaskan area in the earliest years of these periods. For the period from 5000 to 3000 BC, however, in the Seward Peninsula and in the Brooks Range of Alaska, there are indications of a land-hunting and, to some degree, sea-mammal-hunting group that may have been related to some of the forest cultures to the south. Then, between 3000 and 2000 BC, the so-called Arctic Small Tool Tradition developed in northwestern Alaska. It was based on the hunting of caribou and other tundra animals, along with some hunting of sea mammals. This culture included elements from the northern movement of the American Plains hunters to northwestern Canada but was primarily derived from northeastern Siberia. It gradually spread eastward in the Canadian tundra to the northwest and northeast side of Hudson Bay, into extreme northeastern Canada, and to western and northern Greenland. This eastern spread was accomplished by 1000 BC.

Arctic Small Tool Tradition

The Arctic Small Tool Tradition is considered by most prehistorians to be too distinctive to be called Eskimo. The oldest known Eskimo culture is that in the so-called Aleutian and Southern Tradition, dating from as early as 2000 BC (though the earliest archaeological discovery, stone implements found at Anangula Island in the Aleutians, dates back about 8,000 years). Prehistoric south

Alaskan and Aleutian culture was basically Eskimo in character but specifically different in many respects from that of western and northern Alaska. It was somewhat simpler technologically; its most important achievements were the production of things made of stone (lamps, chipped knives, and points), bone (harpoon heads, fishhooks, and wedges for splitting driftwood), and ivory (lip ornaments, figurines, and needles). Although the Aleutian and south Alaskan region had the greatest population density of any part of the Eskimo territory, its culture was remarkably stable, in contrast to the Bering Strait region, where culture changes were frequent and pronounced.

In western and northern Alaska is to be found what is probably the next most ancient Eskimo cultural sequence —that of the Old Whaling, Choris, Norton, and other cultures, each known from only one or a few sites north of Kotzebue Sound near the Bering Strait, the earliest dating from about 1800 BC. Included in this sequence is the Ipiutak culture, which, discovered in 1939, is a major enigma in Eskimo prehistory. Ipiutak, known from the type site at Point Hope and a few other western locations, contains many traits of the Eskimo tradition, including toggle-headed harpoons and other items for sea hunting (although harpoon heads are outnumbered by arrowheads for hunting caribou); but the culture lacks many typical Eskimo items and stands by itself in the large size of the village and the regular arrangement of houses along streets. The Ipiutak site contains more than 600 shallow semi-subterranean dwellings with central fireplaces. The slight accumulations of debris argue for a short period of occupation (estimated to be about 100 years). Dating also is somewhat uncertain. Some radiocarbon dates for Ipiutak are as late as AD 1000, whereas some artifacts suggest overlap with the earlier phases of the Northern Maritime Tradition. Most summaries conclude that Ipiutak dates from about AD 300 to 400 and drew upon both the Northern Maritime Tradition (for seal-hunting techniques) and the northeastern Asian cultures (for certain implements and ceremonies).

Dorset culture

In the eastern parts of the American Arctic—including the Hudson Bay region, Labrador, Newfoundland, and Greenland—there is another very early culture, known as Dorset, dating from about 1000 BC and lasting to about AD 1000 or 1200 (when Eskimos in the Northern Maritime Tradition overran the area). The relationship between Dorset and other Eskimo traditions is somewhat enigmatic. Dorset sites contain semi-subterranean houses and some Eskimo traits such as the toggle-headed harpoons, lamps, and ivory carvings; but many of the artifacts were somewhat different in form, smaller in size, and more delicate; moreover, these sites lack dogsleds, bone arrowheads, lip ornaments, and other typical Eskimo traits. Many prehistorians think that the Dorset culture developed from the Small Tool Tradition and marks the migration of people who began moving from the western end of the American Arctic about 3000 BC, bringing with them implements in the Arctic Small Tool Tradition, and who by 1000 BC had reached the eastern extent of the Eskimo area and developed a largely independent culture. Some materials in later Dorset sites, however, resemble similar items in the Northern Maritime Tradition. Most summaries of Eskimo prehistory conclude that the Dorset Eskimos were overrun by the eastward migration of the Thule Eskimos, who were carriers of the Northern Maritime Tradition. Some scholars believe that the Eskimos of Southampton Island, extinct in 1904, and somewhat different from neighbouring Eskimos, represent a late survival of the Dorset peoples.

Northern Maritime Tradition

The best known early Eskimo cultural sequence is that of the Northern Maritime Tradition, which began in the area from the Bering Sea to Point Barrow about 500 BC. In this tradition, the most typical traits of Eskimo culture, including the semi-subterranean house and the toggle-headed harpoon (with all other parts of the seal-hunting complex), are present from the start. Although some less typical traits were added as the tradition developed and the emphasis in some areas changed (as from chipped-stone to ground-stone implements), the phases in the Northern Maritime Tradition are mainly demarcated only by stylistic changes, as in the design of harpoon parts, arrowheads, needle cases, and art objects.

In art the general trend is from more elaborate to less elaborate. Tools and other implements also become plainer and more utilitarian in later periods, although more efficient. The Birnirk and Thule cultures are the most widely distributed examples of the Northern Maritime Tradition, with Birnirk sites reported from the mouth of the Kolyma River in Siberia to the Mackenzie Delta in Canada, and Thule even more widely distributed, reaching from the Bering Strait to Greenland. The spread of Thule, after it developed out of Birnirk some time before AD 1000, has been traced eastward from Alaska, arriving in Greenland about 1200, where it was influenced by the medieval Norse settlements in the subsequent century. Later there was a resurge of Thule back toward the west, reaching all the way to Bering Strait and accounting for the relative uniformity in language and other aspects of Eskimo culture in early contact times.

At many locations throughout the area of the Northern Maritime Tradition, archaeological materials have been studied that trace the development of Thule into historically identified groups of Eskimos. Although there are many specific problems yet to be resolved, in general the Northern Maritime Tradition has the feeling of history.

Modern populations in the Arctic. Although many nations participated in the exploration of the Arctic, only a few have contributed settlers, and all are Indo-European speakers. Russians predominate in Eurasia (and numerically throughout the Arctic), although they were preceded in Arctic Europe by Scandinavians. Currently, Norway, Sweden, and Finland all have settlements north of the Arctic Circle. In Arctic Alaska and Canada, English-speaking peoples are found in small numbers; and in Greenland, Danish settlers have intermixed with the Eskimos, though Eskimo speech has been retained.

Russian settlement of Siberia

Russians began crossing the Urals in the 16th century and continued in flood tide after the abolition of serfdom in the Russian Empire in 1861. Settlement of the northlands was further accelerated by the building of the Trans-Siberian Railroad in the late 19th and early 20th centuries and was continued as state policy by the Soviet government after the Revolution. In the 1970s there are more than 5,500,000 people in the parts of the Soviet Union north of latitude 60° N, and there are five cities of more than 100,000 people. A condition of note for the modern populations of the Arctic Soviet Union is that 75 percent live in urban concentrations, whereas only 50 percent of the people of the Soviet Union as a whole are urban. Concentrated living in the north shows that the modern populations are mainly involved in extractive industries, such as mining, petroleum, and lumbering (in the sub-Arctic). The Russian settlers, who outnumber the indigenous peoples by 3 to 1, are the inhabitants of the Arctic cities; but some of the native peoples are also becoming urban dwellers. The Komi were reported to be 59 percent urban in 1970. About 70 percent of the people in the Soviet northlands are in Arctic Europe.

American Arctic settlements

In the American Arctic there are no comparable populations. Canada has 1,500,000 people in the Arctic and sub-Arctic, almost all of them in the sub-Arctic; and only one Canadian town north of 60° has a population approaching 5,000 (Whitehorse). In Alaska the population of 302,000 (1970) is mostly concentrated in the sub-Arctic, although of three cities over 15,000, one (Fairbanks) is close to the Arctic Circle. (In this comparison, however, it should also be noted that the "Soviet North" includes much of the sub-Arctic, and most of the Russian settlers live outside the Arctic proper.) A process of population concentration has, in fact, occurred in the American Arctic, even though it has not resulted in cities. In Alaska and Canada, the Eskimos of the north have abandoned many small settlements and concentrated in fewer towns—a response to opportunities for employment and other appeals of living in larger communities. Only in Greenland, where a population increase of 2 percent a year has resulted in population pressures, has there been an increase in the number of communities in recent years.

BIBLIOGRAPHY. H.B. COLLINS, *Arctic Area* (1954), contains a general archaeological summary and an ethnographic sketch of the American Arctic. For general introductions to the entire Arctic area and to the culture area orientation, see M.G. LEVIN and L.P. POTAPOV (eds.), *The Peoples of Siberia* (1964; orig. pub. in Russian, 1956); M.A. CZAPLICKA, *Aboriginal Siberia* (1914); WALDEMAR JOCHELSON, *Peoples of Asiatic Russia* (1928); WALDEMAR BOGORAS, "Elements of the Culture of the Circumpolar Zone," *A. Rep. Smithson. Instn. 1930*, pp. 465–482 (1931), and "New Data on Types and Distribution of Reindeer Breeding in Northern Eurasia," in *Proc. 23rd Int. Congr. Americanists*, pp. 403–410 (1930); A.L. KROEBER, *Cultural and Natural Areas of Native North America* (1939); and VILHJALMUR STEFANSSON, *The Northward Course of Empire* (1922). RONALD ST.J. MACDONALD (ed.), *The Arctic Frontier* (1966), contains a recent, authoritative symposium. CARLETON S. COON, *The Origin of Races* (1962) and with EDWARD E. HUNT, *The Living Races of Man* (1965); and E.A. HOOTON, *Up From the Ape*, rev. ed. (1946), are good for information on physical types in the Arctic in context of all the world's peoples. R.F. SPENCER *et al.*, *The Native Americans: Prehistory and Ethnology of the North American Indians* (1965), includes a summary of Eskimo archaeology on pp. 88–91 and a summary of Eskimo ethnography on pp. 120–154. WALDEMAR BOGORAS, *The Chukchee*, 3 pt. (1904–09, reprinted 1966) and *The Eskimo of Siberia* (1913), are among the many monumental, basic ethnographic monographs on Arctic peoples. ELMAN R. SERVICE, "The Copper Eskimo" and "The Reindeer Tungus of Siberia," in his *Profiles in Ethnology*, rev. ed. (1971), contains some very readable secondary ethnographies of individual Arctic cultures. For archaeological summaries, see JESSE D. JENNINGS, *Prehistory of North America*, pp. 287–320 (1968); and GORDON R. WILLEY, *An Introduction to American Archaeology*, vol. 1, pp. 69–72 and 410–453 (1966). See H.B. COLLINS, *Archaeology of St. Lawrence Island, Alaska* (1937); HELGE LARSEN and FROELICH RAINEY, *Ipiutak and the Arctic Whale Hunting Culture* (1948); and S.I. RUDENKO, *The Ancient Culture of the Bering Sea and the Eskimo Problem* (1961; orig. pub. in Russian, 1947), for some of the basic archaeological monographs. For aspects of the contemporary Arctic, see TERENCE ARMSTRONG, "The Administration of Northern Peoples: The USSR," and MARGARET LANTIS, "The Administration of Northern Peoples: Canada and Alaska," in RONALD ST.J. MACDONALD, *The Arctic Frontier*, pp. 57–119 (1966); TERENCE ARMSTRONG, *Russian Settlement in the North* (1965); and DIAMOND JENNESS, *Eskimo Administration*, pt. 1–3, *Tech. Pap. Arct. Inst. N. Am.* (1962–65).

(W.K.C.)

Arecales

The palm order, Arecales (or Principes), consists of a single family, the Arecaceae (or Palmae), and is generally considered to be second only to the grasses in economic importance. Palms are distinctive in their habit and geography, being characteristically woody in a class of plants (Monocotyledoneae) that is otherwise mostly herbaceous, and being chiefly restricted to the tropics or subtropics, where they are a prominent part of the vegetation. There is a uniformity of aspect to palm leaves that characterizes members of the order despite great differences in the size of various species. Stems may vary from the height and diameter of a pencil to a height of 190 feet (58 metres) or a diameter of three feet. Leaves of a few inches long to the largest in the plant kingdom, more than 30 feet (9 metres) long are found in this order, and seeds from the size of buckshot to the largest known, the 20-pound (9 kilograms) double coconut or coco-de-mer (*Lodoicea maldivica*), are found.

GENERAL FEATURES

Palm distribution

The great centres of palm distribution are in America and in Asia from India to Japan and south to Australia and the islands of the Pacific and Indian oceans, with Africa and Madagascar as a third but much less important palm region. These centres are bounded roughly by latitudes 44° north and 44° south, though the greatest abundance lies primarily between latitudes 30° north and 30° south. These distributions correspond more or less with zones having mean annual temperatures of 60° and 70° F and, with rare exceptions, to areas having 20 or more inches of rainfall per year.

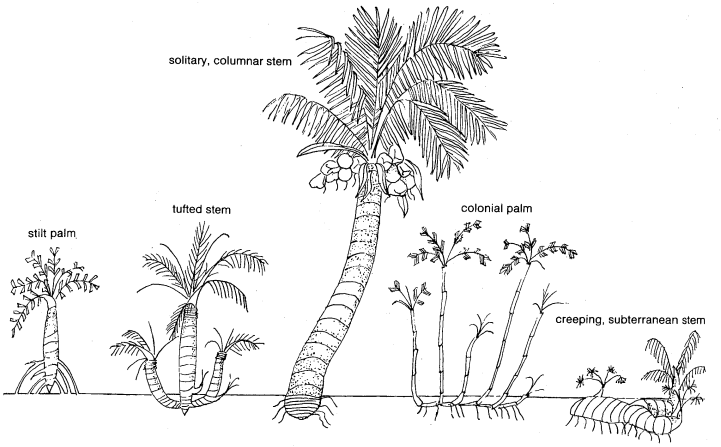
The northernmost palm is *Chamaerops humilis*, which grows about the Mediterranean in Europe and North Africa; the southernmost is *Rhopalostylis sapida*, of New Zealand and the Chatham Islands. Although there are species with extensive ranges, especially in America, most are restricted in range, and those of islands, in particular, are frequently found nowhere else. At least one species, *Liberbaileya gracilis*, is limited to a few limestone outcropings in the Langkawi Island off the Malay Peninsula. No species, except *C. humilis* and the pantropical cultivated coconut, occurs on more than one continent; the genera transcending continental bounds are *Chamaerops* in Europe and Africa, *Elaeis* and *Raphia* in Africa and America, and *Borassus*, *Calamus*, *Hyphaene*, and *Phoenix* in Africa and Asia. Numbers of individuals of a species may be few or many; *Copernicia alba* occurs in solid stands hundreds of square miles in extent in the northeastern section of the Paraguayan Chaco and adjacent Bolivia and Brazil, the largest stands in this region alone containing possibly 500,000,000 plants.

Economic uses

The palms most important in world commerce are the coconut (*Cocos nucifera*) and the African oil palm (*Elaeis guineensis*); both are prime sources of vegetable oil and fat. Probably no plant is as versatile as the coconut. The husk of the fruit is the source of coir, used for ropes and mats; the hard inner fruit layer (endocarp) is used as fuel and to make charcoal, cups, bottles, and trinkets; the coconut milk (liquid endosperm) is used as a beverage and in cooking; the coconut flesh (solid endosperm) is used as raw food or dried to form copra, a source of oil and oil cake, the last used as cattle feed. The sap obtained from tapping the inflorescence, or flower stalk, is drunk unfermented or fermented (toddy) and is a source of sugar, alcohol, and vinegar. Trunks are used in construction and furniture making, and leaves are used in a variety of ways in domestic economies. The African oil palm is important chiefly for the palm oil obtained from the fruit coat and for kernel oil from the seed.

Other palms are used extensively, especially in the Old World. Sugar and alcohol are obtained by tapping inflorescences of the sugar palm (*Arenga pinnata*), the palmyra palm (*Borassus flabellifer*), the wild date (*Phoenix sylvestris*), the fishtail palm (*Caryota urens*), the nipa palm (*Nypa fruticans*), and the buri and talipot palms (*Corypha elata*, and *C. umbraculifera*). Wine is made from species of *Raphia* in Africa and from *Acrocomia* and *Jubaea* in America. *Metroxylon sagu* and, to a lesser extent, *Arenga pinnata* and *Corypha elata* are sources of starch obtained from the pith. The fruit of the date palm (*Phoenix dactylifera*) is a staple in parts of Asia Minor, Arabian Peninsula, and North Africa and is also an article of commerce, while that of the peach palm (*Bactris gasipaes*) is eaten in America. Fruits are also widely used fresh, as sources of oil, and in the preparation of refreshing drinks and ice cream. Throughout much of the Asiatic tropics and even in parts of East Africa, the seed of the betel palm (*Areca catechu*) is used, with lime and the leaf of the betel pepper (*Piper betle*), as a chewing substance. Trunks and leaves serve in local construction, in the making of weapons, and as sources of wax (*Ceroxylon*, *Copernicia*). Leaves of *Corypha* are made into umbrellas and books; others provide material for rain capes, baskets, raffia (*Raphia farinifera*), hats, hammocks, and the fibre known as piassaba.

Minor uses are so diverse as to defy complete listing, but as examples some of the uses of the sugar palm, *Arenga pinnata* may be cited. Leaf stalk bases are used for cordage and thread; leaflets for wrapping material, brooms, basketry, fishing implements, musical instruments, handles, and holders. Leaf midribs and heavy fibers are used for styli, pens, darts for blowguns, and the principal leaf axes are used for lathing, and walking sticks. Fibres of roots are used for weaving, twine, whips, and matting; the hard fibres of the trunk are used for furniture, utensils, swagger sticks, axe handles, and canoes. Split trunks are used for drain pipes, troughs, gutters, conduits, and drums. The feltlike fibres of the leaf sheath provide tinder, caulking, and pillow fillers, and the outer coating of the fruit is used as hog feed and as a source of poison. Seeds are used as touchstones and toys, and other parts are used in local medicine and magic.



Growth habits of representative palms.
Drawing by R. Findahl based on (centre) G.H.M. Lawrence, *Taxonomy of Vascular Plants*, copyright 1951 by The Macmillan Company, reprinted with permission of The Macmillan Company; (others) P.B. Tomlinson, *Principes vol. 5* (1961)

Apart from commercial or local uses, palms are extensively planted for ornament in warm regions or indoors when a tropical effect is desired. Ornamental palms for out-of-doors use number several hundred species.

NATURAL HISTORY

Life cycle. The life cycle of a palm is like that of most flowering plants except that the early vegetative phase is often prolonged because the palm stem generally attains its maximum girth below ground before it begins to grow upward. This is because palms, unlike woody broad-leaved plants, do not have the means for growing a thicker trunk; moreover, they have only a single growing point. There is usually an accompanying change from the small and sometimes strap-shaped leaf of the seedling to the leaf of the adult. Some palms grow rapidly, others grow very slowly. Only by knowing how many leaves a palm produces in a year and by counting leaf scars can an approximation of age be arrived at.

Determination of age in palms

Reproduction may begin very soon in small palms—within two years of germination—or after several years in larger palms. In some it may be suppressed for many years as in the talipot palm (*Corypha umbraculifera*) or the fishtail palm (*Caryota urens*), which flowers once, then dies. In some palms, flowering and fruiting may be continuous or nearly so, but in many there appears to be a periodicity perhaps related to the onset of rains or to light intensity or both. The African oil palm apparently responds cyclically to rain, light, and nutritional levels.

It has long been assumed that palms are wind-pollinated, and doubtless many of those that occur in open places or that overtop the forest canopy are so pollinated. Syrphid flies, however, are known to pollinate *Asterogyne martiana*, curculionid beetles to pollinate some species of *Bactris*, and trigonid bees have been seen to visit flowers of both sexes in *Archontophoenix* cultivated in Panama. Obvious adaptations to insect pollination may be observed in other palms, despite lack of documentation. It is noteworthy that among palms, the male flowers usually open and shed pollen before the female flowers become receptive, though *Nypa*, *Bactris*, and probably others are exceptions. Thus outcrossing is promoted, though self-pollination in palms having flowers of both sexes on one plant is possible when several flower clusters have opened on the same plant.

Upon fertilization, usually only one ovule matures into a seed, but in a few genera two or three or up to 10 may form seeds. Palm fruits for the most part are fleshy, black or brightly coloured, and appear to be adapted for dispersal by birds and mammals. Bats, for example, are said to be fond of the fruit of the talipot palm, and birds that are insect eaters in the north may become fruit eaters in the tropics.

Ecology. Palms have adapted to many habitats within their geographic limits. To the dweller in temperate regions, they tend to be associated with jungle or with tropical shores, where the coconut is often a prominent feature of the landscape. Most palms are inhabitants of tropical forests. There they occur mostly on well-drained acid soils in regions of high though often seasonal rainfall. They range from the lowlands to mountain and cloud forests up to altitudes of 6,000 feet. Rarely, as in *Ceroxylon* and some species of *Geonoma*, they reach elevations as high as 10,000 feet in the Andes. There are exceptions, however, for palms are also found in swamps or poorly drained areas (*Manicaria*, *Mauritia*, *Phoenix paludosa*, *Metroxylon sagu*, *Raphia*) or brackish estuaries and lagoons (*Nypa fruticans*) or areas subject to alternate flooding and drying (*Copernicia alba*). They also occur in deserts or on seashores when underground water is present (*Hyphaene*, *Washingtonia*, *Cocos nucifera*), or in open savanna, grassland, or gallery forest, or restricted to such special habitats as limestone outcrops (*Maxburretia rupicola*), serpentine soils (*Gulubia hombronii*), or river margins (*Astrocaryum jauari*, *Leopoldinia pulchra*) where competition is limited.

Habitats occupied by palms

The mode of germination of palm seeds appears to have some ecological significance. Forest palms mostly have adjacent ligular germination and seldom succeed when exposed to full sun. Palms of drier or more open areas tend to have a form of germination wherein the seed leaf petiole elongates from an inch to two feet or more and buries the stem tip well below the surface of the soil. Delayed germination, in which some seeds germinate later than others, ensures replacements for early seedlings that may be lost through adverse climatic effects or other causes.

Trailing underground stems may be assumed to be adaptations to fire, as in *Serenoa repens* of pine woods and sand dunes in the southeastern United States, or to tidal fluctuations (*Nypa*).

FORM AND FUNCTION

Stem and root. Most palm stems are erect and solitary or clustered, but some grow horizontally, becoming procumbent, or trailing, at or below the surface of the soil and producing the crown at ground level, while others are high-climbing vines. Rare instances of regular branching (in *Allagoptera*, *Chamaedorea*, *Hyphaene*, *Nannorrhops*, *Nypa*, *Vonitra*) appear to involve equal or subequal division at the apex that results in a forking habit. The two newly formed branches may continue equally or one may be overtopped by the other (*Nannorhops*). When thickening occurs, as in the royal palms (*Roystonea*) or in the few that produce conspicuous swellings or "bellies" as in *Colpothrinax*, it is due to an increase in number or size of internal cells and not to new cell production at a cambium, or growing layer. The cortex, or "bark," may be smooth or rough, and is sometimes fiercely armed with spines or is covered with old leaf bases. Nodes, intermittently spaced thickenings of the stem, are often prominent and ringlike.

A typical palm stem, unlike the woody cylinder of a hardwood tree, is composed of hundreds to thousands of conducting strands scattered in softer ground tissue. There are usually two unequal regions, however. The central region is larger and softer than the outer region, which often is composed of densely packed fibres so hard that they quickly dull any cutting instrument. Thus palms are often left in place when forests are cleared for cultivation. Such a construction is a natural example of the best principles of engineering, permitting maximum strength with economic use of strengthening tissues, and it accounts for the ability of palms to withstand winds that break ordinary trees.

The first root of palms, unable to increase in diameter as do roots of broad-leaved plants, is eventually replaced by adventitious roots from the basal nodes of the stem. Most roots penetrate the ground, but in some palms, adventitious roots may form a mound above ground or appear at intervals along the stem. In *Cryosophila* and *Mauritia*, roots along the stem are transformed into spines. Stout prop roots forming a dense or open cone are found at successive nodes along the stem of certain varieties of palms.

The unusual nature of roots in the palms

The leaf. Leaves in the Arecales order are diverse in size, shape, and division. Three distinct but continuous regions are apparent in most: the sheath, petiole or leaf stalk, and blade. The sheath, often fibrous, is that lowermost portion of the leaf that partially or completely surrounds the stem. Sheaths sometimes are elongate or tubular, and when they appear to form a continuation of the stem, they are referred to as a crownshaft. A fibrous ligule, a sort of membrane or appendage, is sometimes evident on the upper margin of the sheath on one side or in front of the petiole. The petiole is discernible above the sheath as a supporting axis devoid of leaflets.

The terminal portion or blade of the palm leaf may represent one of two categories for which the terms pinnate (or feather-like) and palmate (or fanlike) are broadly used. The first category embraces those that have distinct pinnae or leaflets or are bipinnate (*Caryota*) or merely pinnately veined. The second includes blades with relatively elongate central veins or ribs (costapalmate) or those in which segments radiate from a central region at the tip of the petiole. Many palmate and costapalmate leaves have a prominent, often spear-shaped, ridge of tissue at the apex of the petiole known as the hastula.

Leaves are sometimes armed with spines or marginal teeth on sheath, petiole, or blade, or on all parts. Pinnae may be modified into recurved hooks (*Desmoncus*) or the tip of the central axis may be produced into a long slender whiplike strand armed with recurved spines in climbing palms such as *Calamus*.

The inflorescence. The inflorescence unit, the structure with a particular arrangement of flowers on its axis, may be a large and complex panicle (a loose, many-branched cluster) with numerous small leaflike bracts and branches to the fifth order, or it may be reduced to a head or to a spikelike axis with a single bract, or, as in *Calamus*, it may become a slender clawed climbing organ or cirrus (tendril). The usual flower unit has a supporting peduncle, or main axis, on which is borne a two-keeled first bract, a large appendage, and usually one or more other sterile bracts below the flowering portion. The flowering portion of the inflorescence is composed of a central axis and branched flowering axes of one or more orders having other bracts nearer each branching.

A small bract (bracteole) usually subtends the flower, each of which is borne on the surface of the final flowering branch of the inflorescence. More rarely, the flower is sunken in pits formed by the union of prominent bracts or by differential growth of the axis. The arrangement of flowers is sometimes complex. Flowers may be arranged singly, in pairs of a staminate (male) or a neuter flower and a functionally pistillate (female) flower; in groups of two staminate flowers and a pistillate or of two staminate flowers only, or of three to many perfect (both sexes in one flower) or staminate flowers.

The arrangement of palm flowers

The flower. Flowers of the Arecales order are attached directly to the flowering branch without "stems" or more rarely may have supporting stems. In either case they exhibit much diversity in structure, size, and colour and are based usually on a three-part plan. Most are less than an inch in length, but the female flowers of *Phytelephas* may be 10 inches long. The least specialized flowers are those with a calyx of three sepals, a corolla of three petals, a male complement of six stamens with slender to fleshy filaments and linear anthers that open by longitudinal slits, and a female structure of three distinct carpels, each with one ovule. Unisexual flowers are often markedly dimorphic (*i.e.*, the male and female forms are very different). The staminate (male) flowers have poorly developed female parts or none at all; the pistillate (female) flowers may have evident stamens or none.

The sepals in all flowers are either overlapping or joined in a lobed cupule and are three, rarely two or up to 10 in number. Petals, similarly, may be two or up to 10 but these are usually thicker than sepals, fibrous, and whitish or coloured. They may be distinct and overlapped or they may meet without overlap, especially in male flowers. Petals also are sometimes partially to completely united. In a few palms the sepals and petals are reduced and united into a thin lobed structure.

Stamens, though most often six in number, may rarely be three (*Areca triandra, Geonoma triandra, Nypa fruticans*) or more numerous, ranging from 6 to 36 in *Heterospathe*, to more than 200 in such groups as *Caryota, Phytelephas*, and *Veitchia*. The filaments are erect or inflexed at the apex, are often thick and fleshy, and they vary in shape. Sometimes they are united with each other or are fused to the petals or female structures. Sterile stamens may differ only slightly from fertile stamens, or they may consist of a filament alone without an anther, or be united in a cup about the base of the female structure, or in a tube joined to the petals, or be reduced to minute flaps of tissue. Pollen grains are globose to irregular in shape but most often they are ellipsoidal. The surface of the grains, however, may be smooth or ridged to warty or spiny.

Carpels, the female parts of the flower, are usually three in number, and they arise above the region where sepals, petals, and stamens are joined together. Though sometimes quite distinct, the carpels are most often united in a three-chambered female structure in which each chamber may contain an ovule or in which one or two chambers may be reduced or apparently lacking, the pistil then appearing to be two-chambered or single-chambered. The female part of the flower occasionally, though rarely, consists of one or two distinct carpels or it may even have five to ten chambers.

The fruit. Palm fruits, whether large or small, usually do not split at maturity and thus seeds are not shed separately. The fruit coat (pericarp) is three layered: a smooth, prickly, warty, or scaly outer layer (exocarp); a fleshy, fleshy-fibrous, or fibrous middle layer (mesocarp); and a thin and membranous to thick and bony inner layer (endocarp) surrounding the seed. The mesocarp is sometimes filled with stinging chemical crystals or with "stone cells," thick-walled dense cells. The seed has a thin, often highly vascularized outer layer (testa) over copious, hard, sometimes hollow endosperm in which the pluglike embryo is embedded at either end or on one side.

EVOLUTION

Palms stand by themselves in terms of evolution. They do not appear to be very closely related to any other order of flowering plants unless it be the order Cyclanthales, nor can they be derived clearly from any present or past group of plants. They appear essentially in their present form in fossils from the Upper Cretaceous Period, 65,000,000 to 100,000,000 years ago; indeed, *Nypa* pollen is known from this age in northwest Borneo where the genus still occurs. By Eocene times, about 60,000,000 years ago, palms were widespread, and are known in some abundance and diversity from such locations as the southern United States and the London clay of England. Pollen of the African oil palm is reported from Miocene deposits (about 30,000,000 years old) in the Niger Delta and fossil fruits from the Eocene Period of India and the Pliocene Period (about 10,000,000 years ago) of New Zealand have been referred to the genus *Cocos*.

The fossil record

CLASSIFICATION

Distinguishing taxonomic features. The order Arecales, with a single family named Arecaceae (or Palmae), is divided into subfamilies on the basis of leaf characteristics, nature of the inflorescence, and structure of flowers and fruit. The order is set apart from other members of the flowering plant class Monocotyledoneae by the combination of woody habit, plicate (pleated) leaves, mostly three-parted flowers that are largely unisexual and usually have bracts and superior ovaries with one ovule per chamber, and seeds with copious endosperm.

Annotated classification

ORDER ARECALES (Principes)
Woody Monocotyledoneae with plicate, paripinnate, imparipinnate, pinnately nerved, costapalmate, or palmate leaf blades, the pinnae or segments induplicate or reduplicate in bud. Inflorescence paniculate to spicate, always with a prophyll, often with sterile bracts on the peduncle, usually with bracts subtending all branches, bearing usually bracteate flowers arranged singly, in pairs, or in cincinni of 3 or more.

Flowers perfect or unisexual, calyx and corolla each of usually 3 but sometimes only 2 or as many as 10 sepals and petals or rarely the perianth uniseriate and lobed, androecium of 3 (rarely) or 6 to many stamens, gynoecium of 1 to 3 or rarely up to 10 distinct or united carpels, locules uniovulate, seed with copious endosperm and small embryo.

Family Arecaceae (Palmae)

The family Arecaceae is the only one in the order and therefore has all the characters of the order. There are about 200 genera and 2,600 species in the tropics and subtropics of both hemispheres.

Subfamily Coryphoideae

Leaves palmate or costapalmate with a hastula at apex of petiole on upper side, segments induplicate in bud. Inflorescences developing acropetally, with prophyll only or usually with additional peduncular bracts, rachillae slender. Plants hermaphrodite, monoecious, polygamous, or dioecious. Flowers borne singly, in sympodial pairs, or in cincinni, perfect or unisexual, then not markedly dimorphic, carpels 1–3, distinct or variously united. Fruit lacking scales, sometimes warty. There are about 31 genera and 330 species distributed in both hemispheres.

Subfamily Phoenicoideae

Leaves imparipinnate with lower pinnae modified into spines, pinnae induplicate in bud. Inflorescences developing acropetally, with prophyll only on flattened peduncle, rachillae slender, often fascicled. Plants dioecious; flowers borne singly, unisexual, dimorphic, carpels 3, distinct. Fruit lacking scales. One genus and about 12 species distributed in Africa and Asia.

Subfamily Borassoideae

Leaves costapalmate with a hastula at apex of petiole, segments induplicate in bud. Inflorescences developing acropetally, with prophyll and peduncular bracts, rachillae thick, densely bracteate and catkin-like. Plants dioecious. Flowers borne singly, or in cincinni of 3 to many, unisexual, dimorphic, staminate and sometimes pistillate sunken in pits formed by union of bracts, carpels 3, united. Fruit lacking scales. Seven genera and about 42 species distributed in Africa and Asia.

Subfamily Caryotoideae

Leaves pinnately nerved, pinnate, or bipinnate, pinnae induplicate in bud. Inflorescences developing basipetally or very rarely acropetally, solitary or several in a leaf axil, with prophyll and peduncular bracts, rachillae slender. Plants monoecious. Flowers borne singly or in cincinni of 3 or 2, carpels united, having 1 to 3 locules. Fruit lacking scales. Three genera and about 38 species distributed in Asia and Malaysia to the British Solomon Islands.

Subfamily Lepidocaryoideae

Leaves pinnate, costapalmate, or palmate without a hastula at apex of petiole, segments or pinnae reduplicate in bud. Inflorescences developing acropetally, with prophyll alone or usually with peduncular bracts, rachillae slender to stout, usually with tubular bracts. Plants hermaphrodite, monoecious or dioecious; flowers borne singly, in pairs, or in monopodial clusters, perfect or unisexual and slightly to markedly dimorphic, carpels 3, united, scaly. Fruit covered with overlapping scales. About 22 genera and 500 species distributed in both hemispheres, most abundant in Asia.

Subfamily Nypoideae

Leaves pinnate, pinnae reduplicate in bud. Inflorescences developing acropetally, with prophyll and numerous peduncular bracts, staminate rachillae catkin-like on bracteate branches below the bracteate terminal head of pistillate flowers. Plants monoecious. Flowers unisexual, dimorphic, perianth segments similar, distinct, strap-shaped, much exceeded by 3 or sometimes 4 distinct carpels at anthesis, stamens 3 with united filaments. Fruit not scaly but fibrous. One genus and species, *Nypa fruticans*, in estuaries and lagoons from India to the Ryukyu Islands and British Solomon Islands.

Subfamily Cocosoideae

Leaves pinnate or pinnately nerved, pinnae reduplicate in bud. Inflorescences developing acropetally, sometimes unisexual, with prophyll and 1 peduncular bract, rachillae slender or rarely stout with flowers in pits. Plants monoecious. Flowers in sympodial triads or pairs or sometimes borne singly, unisexual, dimorphic, carpels 3 or rarely up to 7, united. Fruit not scaly, endocarp thick, bony, with 3, or rarely up to 7, pores at base, middle, or apex. About 28 genera and 610 species distributed in America and Africa, 1 species pantropic in cultivation and probably of Old World origin.

Subfamily Arecoideae

Leaves pinnate or pinnately nerved, pinnae reduplicate in bud. Inflorescences developing acropetally, rarely more than 1 per leaf axil, with prophyll and usually 1 to several peduncular bracts, paniculate to spicate, rachillae slender. Plants hermaphrodite (*Pseudophoenix*), monoecious, or dioecious. Flowers borne singly or in sympodial triads or pairs or lines, unisexual and slightly to markedly dimorphic except in *Pseudophoenix* where flowers are perfect, carpels 3, united, often the pistil unilocular and uniovulate by abortion. Fruit not scaly, sometimes warty, endocarp thin or if thick then lacking 3 pores. About 100 to 110 genera and 1,100 species distributed in both hemispheres.

Subfamily Phytelephantoideae

Leaves pinnate, pinnae reduplicate in bud. Inflorescences developing acropetally, with a prophyll and 1 peduncular bract, staminate spikes thick and densely flowered or with knoblike branches, pistillate flowers in heads. Plants dioecious. Flowers borne singly but densely, unisexual, dimorphic, perianth segments connate in a lobed cupule in staminate flowers, distinct, numerous (to 10 sepals, 10 petals) and very large (to 10 inches) in pistillate flowers, stamens numerous, carpels 7 to 10, united, with elongate style and stigmas. Fruit warty. Three or perhaps 4 genera and about 8 species distributed in Panama and South America.

Critical appraisal. The major groups of palms, whether accepted as tribes or subfamilies, have been recognized in large part since about 1883, when six tribes were recognized corresponding to six of the subfamilies presented above. The subfamilies Caryotoideae, Nypoideae, and Phytelephantoideae were included within the tribe Areceae at that time.

Status of palm classification

Largely because of their size, palms have not been as well studied as the majority of plant families. Evidence from many disciplines, which is necessary for an understanding of evolution in the Arecales order, is still being amassed and assessed. Many biological problems, such as detailed studies of pollination, have scarcely been touched. The Arecoideae subfamily in particular is still in need of revision to bring natural groups of genera together at appropriate levels in a taxonomic hierarchy. Four to six distinctive subunits, for example, may eventually be recognized, one of them further divisible into possibly 16 to 17 distinctive groupings.

BIBLIOGRAPHY. O. BECCARI, *Asiatic Palms*, vol. 11–13 (1908–33), the most elaborate and useful series of monographs on palms published, with analytical and large photographic plates, *Palme della tribù Borasseae* (1924), a monographic account of the Borassoideae in Italian, and with R.E.G. PICHI-SERMOLLI, "Subfamiliae Arecoidearum Palmae Gerontogeae tribuum et generum conspectus," *Webbia* 11:1–187 (1955), a conspectus of the Old World genera of Arecoideae and a summary of subfamilies with text in Italian, descriptions in Latin, and many line drawings of floral and fruit details (the basis for subfamilies as currently accepted); E.J.H. CORNER, *The Natural History of Palms* (1966), a popular account copiously illustrated with line drawings and photographs but the text is sometimes unreliable; J.D. HOOKER, "Palmae," in G. BENTHAM and J.D. HOOKER, *Genera Plantarum*, 3:870–948 (1883), a critical study of palm classification with unrivalled detailed descriptions of genera in Latin; R.B. LEDIN (ed.), "Cultivated Palms," *Am. Hort. Mag.*, 40:1–189 (1961), a popular and principally horticultural account; C.F.P. VON MARTIUS, *Historia naturalis palmarum*, 3 vol. (1823–53), a classic and still fundamental work in Latin handsomely illustrated with hand coloured folio plates; E. POTZTAL, "Reihe Principes," in H. MELCHIOR (ed.), *Engler's Syllabus der Pflanzenfamilien*, 12th ed., 2:579–588 (1964), the most recent synoptic account, in German, of the order Arecales (as Principes) with descriptions of subfamilies and tribes but not of genera.

(H.E.Mo.)

Argentina

Argentina (República Argentina) occupies most of the southern part of the continent of South America. In shape it resembles a wedge with its base on the Tropic of Capricorn and its point aimed at Antarctica. It is bounded by Chile on the south and west; Bolivia and Paraguay on the north; and Brazil, Uruguay, and the Atlantic ocean on the east. The length of Argentine territory from north to south is 2,170 miles, and its greatest width is about 868 miles. It has an Atlantic coastline of more than 2,500 miles. Its 973-mile frontier with Paraguay is formed by the Pilcomayo, Paraguay, and Paraná rivers; its 669-mile frontier with Brazil runs along the Iguazú, San Antonio, Pepiri Guazú, and Uru-

guay rivers; and its 490-mile frontier with Uruguay is on the Uruguay and La Plata rivers.

Argentina has a total area of 1,068,301 square miles (2,776,888 square kilometres) in South America. It also claims the Falkland Islands, which are under British rule, various other islands in far southern waters, and a 49° sector of Antarctica, which together cover an area of 481,777 square miles (1,247,803 square kilometres).

Formerly a colony of Spain, the country attained its independence in 1816. It is a federation of autonomous states with a republican system of government and a constitution modelled on that of the United States. Since the latter part of the 19th century, Argentina has undergone considerable economic and social development. By the 1970s it was among the top 10 or 15 countries in the world with respect to education, literacy, medical services, dietary standards, and the use of such goods as telephones, radio and television sets, and automobiles.

THE LANDSCAPE

The natural environment. Argentina has a varied topography, including high mountains, prairies, deserts, and seacoasts. These differences are accompanied by great contrasts in climate, vegetation, and animal life.

Relief. Stretching eastward from the base of the Andes mountains are great alluvial plains less than 650 feet (200 metres) above sea level. In the north, above the 31st parallel, is the Chaqueña prairie, or Chaco, covered by forests, woods, and savannas (grassland with scattered trees and drought-resistant undergrowth). The Pilcomayo, Bermejo, and Dulce rivers meander through it. South of the Chaqueña are the Pampas, characterized by sandy hills, fertile prairies, marshy lowlands, and rambling rivers, such as the Salado, the Carcarañá, and the Quinto. In the south there are small mountain ridges such as Tandil (1,781 feet), Ventana (4,078 feet), Lihuel Calel, Chata, and Choique Mahuida. To the northeast are the ridges of San Luis and Córdoba (9,456 feet [2,882 metres]). In the region north of Buenos Aires known as Mesopotamia, bordering Uruguay and Brazil, the topography consists predominantly of small hills and marshy lowlands. To the south, in the delta area of the Plata, there is a broad, marshy zone containing many streams that flood periodically.

Plateaus and deserts

Dry and inhospitable plateaus are found in the southern and western parts of the country. In Patagonia, from the Río Colorado south to the Strait of Magellan, the land has been eroded by rivers and glaciers. Strong and blustery winds pick up quantities of dust. Altitudes vary from 300 to more than 3,000 feet. In the west central area, between the Río Desaguadero and the Andes, there are sandy deserts with salt pits and dry shrub vegetation. Altitudes are above 1,600 feet. In the northeastern part of the country, where the province of Misiones projects like a tongue between Brazil and Paraguay, is a plateau of sandstone and basalt covered by subtropical forest. It has a rough relief with mountains, the highest of which is Victoria (3,375 feet [1,029 metres]). In the valleys are rapid-running rivers.

The western borderlands are dominated by the Andes mountains. In the northwest, above the 27th parallel, are ranges up to 22,000 feet (7,000 metres) in height, and consist of high basins separated by short but steep-sided ranges. There are two major cordilleras (parallel mountain ranges). To the west are the Salto-Jujeña, cut by canyons called *quebradas* through which run small rivers that extend back into the high desert plateau (average altitude, 12,000 feet). The peaks here reach 16,000–22,000 feet and many have snowcaps (Chañi, Cachi, Palermo). East of them are the forest-covered Sierras Subandinas, and to the southeast are the Sierras Pampeanas. South of the 27th parallel, the Andes become a single dominating range with peaks up to 22,000 feet or more. From the 39th parallel southward to Tierra del Fuego are the low-lying Patagonian Andes, divided by transverse valleys with lakes and meadows. Glacial and volcanic formations dominate the landscape. There are forests on the slopes and snow on the summits.

The great river mouth of the Río de La Plata, the entrance for Argentina's commerce with the world, is lined with swampy coasts where fields of tall coarse grass enclose the shore. Much of this has given way to harbour facilities and urban areas. The Atlantic coast near Buenos Aires has dunes and large beaches catering to summer residents. From there south to the Río Colorado, the coasts are less accessible, although there are a number of ports, bathing resorts, and villages. The Patagonian coast from the Río Colorado to Tierra del Fuego is wild and irregular, with many cliffs.

The Andes contain a cold, dry desert area known as the Puna, inhabited chiefly by miners and shepherds. This area extends northward from Catamarca to Bolivia and includes part of northern Chile. Another desert is in the wastelands and plains of the Sierras Pampeanas in the eastern Andes. The climate is dry, and the occasional rain disappears into the sandy soil. The sparse population is clustered around water sources, at crossroads, and in mining areas, petroleum fields, and at service centres.

Major rivers

Drainage and vegetation regions. The largest river basin in Argentina is that of the Río de La Plata. It drains an area of 1,600,000 square miles, including the whole of Paraguay, eastern Bolivia, most of Uruguay, and a large part of Brazil. Its tributaries include such major rivers as the Uruguay, Paraná, Iguazú, Paraguay, Pilcomayo, Bermejo, Salado, and Carcarañá. In the Province of Buenos Aires, small courses born in the western mountains cross the Pampas to the sea. It is an area of temperate climate and fertile chernozem (dark, deep, rich humus) soils.

The rivers of Patagonia, in the south, are formed in the Andes and flow to the east across the plateau in low, wide valleys. The Río Colorado is 744 miles long. The Río Negro, formed by the junction of the Neuquén and Limay rivers, is 527 miles long. It supplies water for irrigation. The Río Chubut, 248 miles in length, has also been dammed. Other rivers farther south are the Deseado, the Chico, the Santa Cruz, and the Gallegos.

There are small river basins in the Andes that gather waters from Argentine sources and drain westward into Chile. These mountain basins have podsol soils (light gray topsoil, coffee-brown subsoil) with a forest cover.

The Mar Chiquita Basin of the Pampas is watered by the Dulce, Primero, and Segundo rivers. To the south lies the Pampas proper, a temperate, fertile region with cattle and grain that is the economic heart of Argentina. The Río Desaguadero and its tributaries in the Andes mountains to the west (27°–36° south latitude) water the sandy deserts of Mendoza province. The principal tributaries are the Jachal, San Juan, Mendoza, Tunuyán, Diamente, and Atuel. All of them are used for irrigation and electric power in the Andean piedmont.

Climate. Argentina lies almost entirely within the Southern Hemisphere's temperate zone. Its climate is also influenced by the proximity of oceans on both sides and of high mountains along its western border.

The northeastern part of the country, between the Uruguay and Paraná rivers, is subtropical and humid. The Chaco prairie in the northwest is tropical but has mild, dry winters. The eastern Andes in the northern provinces of Salta, Jujuy, and Tucumán have a mountain tropical climate. The Pampas are a temperate region, varying according to latitude and with the proximity of mountains or ocean. In the southern part of the country temperatures are generally colder, although they are moderated because the continent tapers sharply, bringing the oceans closer. Temperatures in Tierra del Fuego at the southern extremity range from about 36° F (2° C) in the months of May through August to about 48° F (9° C) in January and February. In the northern part of Argentina, in the Chaco, temperatures range from about 64° F (18° C) in June or July to about 82° F (28° C) in January or February. Winter frosts occur in all parts of the country, although they are unusual in the far north.

Rainfall

Rainfall is most abundant in the northeastern area between the Uruguay and Paraná rivers known as Mesopotamia. Much of Argentina is arid or semi-arid, particularly the western parts of the country near the Andes.

At the southern tip in Tierra del Fuego there is abundant year-round rainfall, as also through much of the Patagonian Andes. In the Humid Pampa (the Province of Buenos Aires and parts of the Provinces of Santa Fe, Córdoba, and La Pampa), the rainfall varies from 39 inches in the east to 20 inches in the areas near the Andes —about the minimum for nonirrigated crops.

Animal and plant life. Argentina's fauna and flora are as varied as its soils and climate. In heavily settled regions, the natural animal and plant life have been profoundly modified by man.

The subtropical forests of the far northeast, in the Provinces of Misiones, Corrientes, and eastern Chaco, are rich in large trees, bamboo, palm, and rattan. Animals include the tapir; the pampas cat; the jaguar; the ocelot; the peccary, or mountain pig; monkeys; reptiles, including the boa constrictor, lance-headed viper, and alligators. Large fish abound, including the magnificent gamefish, the dorado. The mountain forests in the northwestern Provinces of Salta and Tucumán flourish in the heavy summer rains; the trees grow large, especially rattan and ferns, and there are jungle cats, monkeys, birds, and reptiles. Between these two areas are the plains of the Chaco, covered by dense growths of the red quebracho tree and the quaiacum, or lignum vitae; there are palm trees, carob trees, and cacti and, in some places, open plains. The fauna of the Chaco are rich and varied, including deer, peccaries, monkeys, tapir, jaguars, anteaters, armadillos, aquatic capybaras, and otter.

In the Mesopotamia region the humid, temperate climate sustains low forests of talas, carob, palm, and mimosa. Animals include foxes, skunks, wildcats, snakes, hares, and tortoises. Many birds inhabit the forests, pastures, ponds, and rivers.

MAP INDEX

Political subdivisions

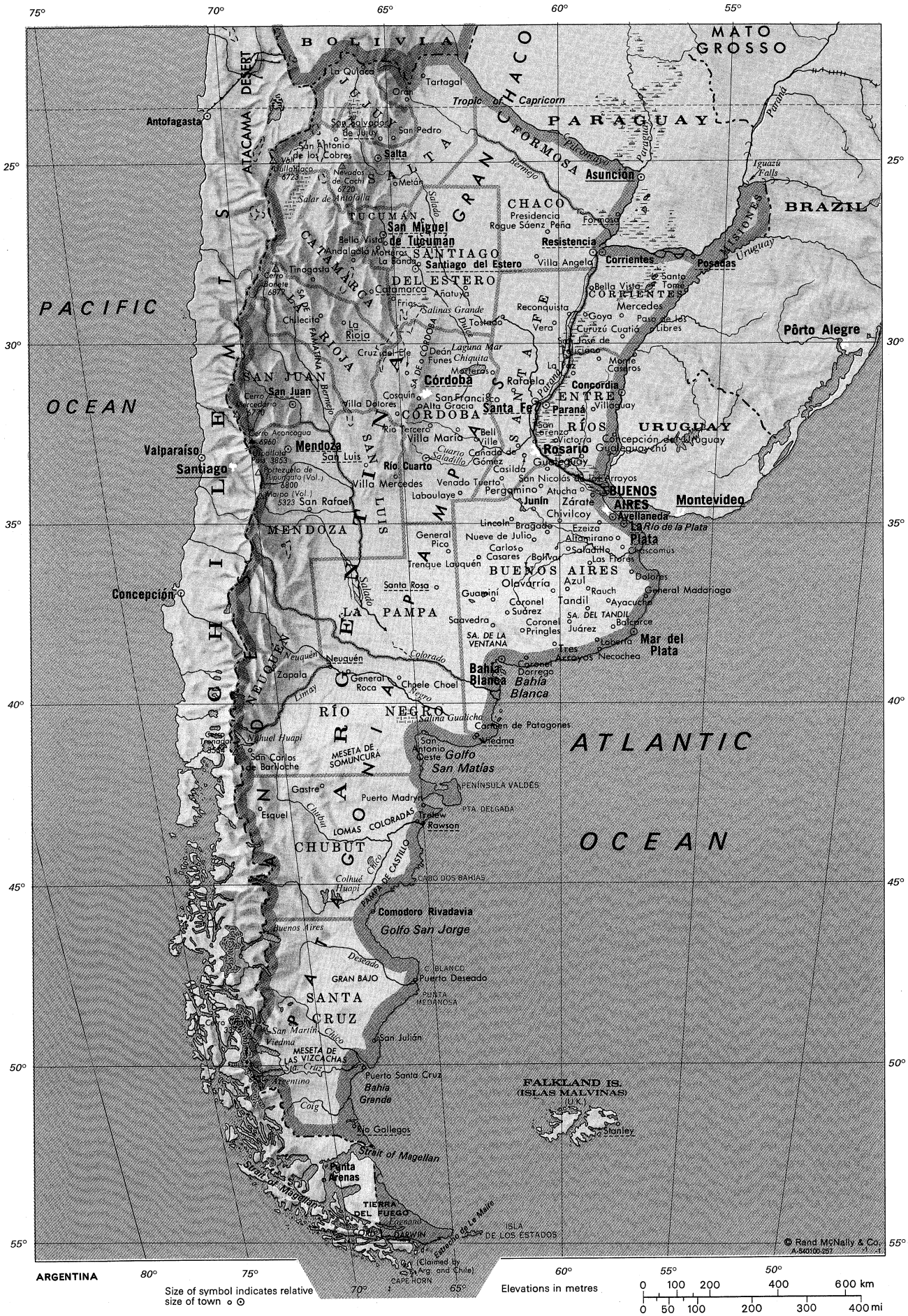
Buenos Aires....36·00s 60·00w
Catamarca.....27·15s 67·15w
Chaco.....26·25s 60·30w
Chubut.....43·30s 69·00w
Córdoba.....32·00s 64·00w
Corrientes.....29·00s 58·00w
Entre Ríos.....32·00s 59·20w
Formosa.....25·00s 60·00w
Jujuy.....23·00s 66·00w
La Pampa.....37·00s 66·00w
La Rioja.....29·30s 67·00w
Mendoza.....34·40s 68·30w
Misiones.....27·00s 54·40w
Neuquén.....39·00s 70·00w
Río Negro.....40·00s 67·00w
Salta.....25·00s 64·20w
San Juan.....30·50s 69·00w
San Luis.....34·00s 66·00w
Santa Cruz.....48·00s 70·00w
Santa Fe.....31·15s 61·00w
Santiago del Estero.....27·40s 63·15w
Tucumán.....27·00s 65·30w

Cities and towns

Alta Gracia.....31·40s 64·26w
Altamirano.....35·26s 58·12w
Añatuya.....28·30s 62·50w
Andalgalá.....27·36s 66·20w
Atucha.....33·58s 59·18w
Avellaneda.....34·40s 58·20w
Ayacucho.....37·10s 58·30w
Azul.....36·45s 59·50w
Bahía Blanca....38·44s 62·16w
Balcarce.....37·52s 58·15w
Bella Vista.....27·02s 65·19w
Bella Vista.....28·30s 59·00w
Bell Ville.....32·40s 62·40w
Bolívar.....36·14s 61·07w
Bragado.....35·10s 60·30w
Buenos Aires....34·36s 58·27w
Cañada de Gómez.....32·49s 61·25w
Carlos Casares...35·37s 61·22w
Carmen de Patagones.....40·48s 63·00w
Casilda.....33·03s 61·10w
Catamarca.....28·30s 65·45w
Chascomús.....35·35s 58·00w
Chilecito.....29·10s 67·30w
Chivilcoy.....34·52s 60·02w
Choele-Choel....39·15s 65·30w
Comodoro Rivadavia.....45·50s 67·30w
Concepción del Uruguay.....32·30s 58·14w
Concordia.....31·24s 58·02w
Córdoba.....31·25s 64·10w
Coronel Dorrego.38·45s 61·17w
Coronel Pringles.38·00s 61·20w
Coronel Suárez..37·25s 61·56w
Corrientes.....27·30s 58·50w
Cosquín.....31·15s 64·29w
Cruz del Eje.....30·44s 64·49w
Curuzú-Cuatiá...29·50s 58·05w
Darwin.....38·45s 65·49w
Deán Funes.....30·25s 64·20w
Dolores.....36·19s 57·40w
Esquel.....42·55s 71·20w
Ezeiza.....34·51s 58·32w
Formosa.....26·10s 58·11w
Frías.....28·40s 65·10w
Gastre.....42·17s 69·15w
General Madariaga.....36·59s 57·14w
General Pico.....35·38s 63·46w
General Roca....39·02s 67·33w
Goya.....29·10s 59·20w
Gualeguay.....33·10s 59·20w
Gualeguaychú...33·00s 58·30w
Guaminí.....37·02s 62·36w
Juárez.....37·41s 59·49w
Jujuy, see San Salvador de Jujuy
Junín.....34·35s 60·58w
La Banda.....27·44s 64·15w
Laboulaye.....34·05s 63·25w
La Paz.....30·45s 59·38w
La Plata.....34·55s 57·57w
La Quiaca.....22·05s 65·36w
La Rioja.....29·25s 66·50w
Las Flores.....36·02s 59·07w
Lincoln.....34·55s 61·30w
Lobería.....38·07s 58·47w
Mar del Plata....38·01s 57·35w
Mendoza.....32·54s 68·50w
Mercedes.....29·10s 58·02w
Metán.....25·30s 65·00w
Monte Caseros...30·15s 57·38w
Monteros.....27·10s 65·30w
Morteros.....30·40s 62·00w
Necochea.....38·34s 58·45w
Neuquén.....39·00s 68·05w
Nueve de Julio...35·30s 60·50w
Olavarría.....36·53s 60·20w
Oran.....23·13s 64·17w
Paraná.....31·45s 60·30w
Paso de los Libres.....29·45s 57·05w
Pergamino.....33·53s 60·36w
Posadas.....27·25s 55·50w
Presidencia Roque Sáenz Pēna.....26·50s 60·30w
Puerto Deseado..47·45s 65·55w
Puerto Madryn...42·46s 65·02w
Puerto Santa Cruz.....50·04s 68·32w
Rafaela.....31·17s 61·30w
Rauch.....36·47s 59·05w
Rawson.....34·40s 60·02w
Reconquista.....29·10s 59·40w
Resistencia.....27·30s 58·59w
Rió Cuarto.....33·08s 64·20w
Río Gallegos.....51·37s 69·10w
Río Tercero.....32·10s 64·05w
Rosario.....32·57s 60·40w
Saavedra.....37·45s 62·23w
Saladillo.....35·38s 59·48w
Salta.....24·47s 65·24w
San Antonio de los Cobres.....24·15s 66·20w
San Antonio Oeste.....40·44s 64·57w
San Carlos de Bariloche.....41·08s 71·15w
San Francisco...31·27s 62·05w
San José de Feliciano.....30·25s 58·45w
San Juan.....31·30s 68·30w
San Julián.....49·19s 67·40w
San Lorenzo.....32·45s 60·44w
San Luis.....33·20s 66·20w
San Miguel de Tucumán.....26·49s 65·13w
San Nicolás de los Arroyos.....33·20s 60·14w
San Pedro.....24·14s 64·50w
San Rafael.....34·40s 68·21w
San Salvador de Jujuy.....24·10s 65·20w
Santa Fe.....31·40s 60·40w
Santa Rosa.....36·40s 64·15w
Santiago del Estero.....27·50s 64·15w
Santo Tomé.....28·35s 56·05w
Tandil.....37·20s 59·05w
Tartagal.....22·32s 63·50w
Tinogasta.....28·05s 67·34w
Tostado.....29·15s 61·45w
Trelew.....43·15s 65·20w
Trenque Lauquen.....35·58s 62·44w
Tres Arroyos.....38·22s 60·15w
Tucumán, see San Miguel de Tucumán
Ushuaia.....54·47s 68·20w
Venado Tuerto...33·46s 61·58w
Vera.....29·22s 60·09w
Victoria.....32·40s 60·10w
Viedma.....40·50s 63·00w
Villa Ángela.....27·35s 60·43w
Villa Dolores.....31·58s 65·12w
Villaguay.....31·55s 59·00w
Villa María.....32·25s 63·15w
Villa Mercedes...30·08s 68·41w
Zapala.....38·55s 70·05w
Zárate.....34·05s 59·02w

Physical features and points of interest

Aconcagua, Cerro, *mountain*.32·39s 70·00w
Andes Mountains.....35·00s 70·00w
Antofalla, Salar de, *salt flat*.....25·40s 67·45w
Antuca, *volcano*..37·35s 71·45w
Argentino, *lake*...50·15s 72·45w
Atlantic Ocean...45·00s 55·00w
Bermejo, *river*....26·52s 58·23w
Blanca, Bahia, *bay*.....38·55s 62·10w
Blanco, Cabo, *cape*.....47·15s 65·45w
Bonete, Cerro, *mountain*.....27·51s 68·48w
Buenos Aires, *lake*.....46·35s 72·00w
Cachi, Nevados de, *mountains*...25·05s 66·40w
Castillo, Pampa del, *plain*.....45·58s 68·24w
Chaltel, Cerro, *mountain*.....49·17s 73·05w
Chico, *river*.....43·50s 66·25w
Chubut, *river*....43·20s 65·05w
Coig, *river*.....51·00s 69·10w
Colhué Huapi, *lake*.....45·30s 68·45w
Coloradas, Lomas, *hills*.....43·15s 67·20w
Colorado, *river*...38·30s 66·00w
Córdoba, Sierra de, *mountains*...31·15s 64·30w
Cuarto Saladillo, *river*.....33·00s 63·25w
Darwin, Cordillera, *mountains*.....54·45s 67·00w
Delgada, Punta, *point*.....42·45s 63·40w
Deseado, *river*...47·45s 65·50w
Dos Bahías, Cabo, *cape*.....44·55s 65·32w
Dulce, *river*.....30·31s 62·32w
Fagnano, *lake*....54·35s 68·00w
Famatina, Sierra de, *mountains*...29·00s 67·50w
Grand Bajo, *plain*.47·35s 68·45w
Gran Chaco, *plain*.....25·00s 62·00w
Grande, Bahía, *bay*.....51·00s 68·40w
Grande, Salinas, *salt flats*.....29·37s 64·56w
Gualicho, Salina, *salt flat*.....40·25s 65·20w
Iguazú, *river*.....25·36s 54·36w
Iguazú Falls, *waterfalls*.....25·41s 54·26w
Limay, *river*.....39·00s 68·00w
Llullaillaco Vokán, *volcano*..24·43s 68·33w
Magellan, Strait of.....52·30s 69·00w
Magallanes, Estrecho de, see Magellan, Strait of
Maipó, *volcano*...34·10s 69·50w
Mar Chiquita, Laguna, *lake*....30·42s 62·36w
Medanosa, Punta, *point*....48·08s 65·58w
Mercedario, Cerro, *mountain*.31·58s 70·07w
Nahuel Huapi, *lake*.....41·00s 71·30w
Negro, *river*.....41·02s 62·47w
Neuquén, *river*...38·59s 68·00w
Pampas, *physical region*.....35·00s 63·00w
Paraguay, *river*...27·18s 58·38w
Paraná, *river*.....33·43s 59·15w
Patagonia, *physical region*..44·00s 68·00w
Pilcomayo, *river*..25·21s 57·42w
Plata, Río de la, *estuary*.....34·45s 57·30w
Salado, *river*.....37·48s 66·10w
San Jorge, Golfo, *gulf*.....46·00s 66·50w
San Martín, *lake*..49·00s 72·40w
San Matías, Golfo, *gulf*.....41·30s 64·20w
Santa Cruz, *river*.....50·10s 68·20w
Somuncurá, Meseta de, *plateau*.....41·30s 67·00w
Tandil, Sierra del, *mountains*..38·40s 59·40w
Tierra del Fuego, *island*.....54·00s 69·00w
Tronador, Cerro, *mountain*.....41·10s 71·54w
Tupungato, Portezuelo de, *volcano*.....33·22s 69·47w
Uruguay, *river*...34·12s 58·18w
Uspallata Pass...32·50s 70·05w
Valdés, Península.....42·30s 64·00w
Ventana, Sierra de la, *mountains*...38·00s 63·00w
Viedma, *lake*.....49·40s 72·35w
Vizcachas, Meseta de las, *plateau*.....50·35s 71·55w


BOLIVIA
PARAGUAY
BRAZIL
PACIFIC OCEAN
ATLANTIC OCEAN
MATO GROSSO
CHACO
FORMOSA
GRAN CHACO
ANDES MTS
ATACAMA DESERT
PUNA DE ATACAMA
SALTA
TUCUMÁN
CATAMARCA
LA RIOJA
SAN JUAN
MENDOZA
SAN LUIS
CÓRDOBA
SANTIAGO DEL ESTERO
CHACO
CORRIENTES
ENTRE RÍOS
SANTA FE
BUENOS AIRES
LA PAMPA
NEUQUÉN
RÍO NEGRO
CHUBUT
SANTA CRUZ
TIERRA DEL FUEGO
MISIONES
Tropic of Capricorn
Antofagasta
Valparaíso
Santiago
Concepción
Mendoza
San Juan
Salta
San Miguel de Tucumán
Santiago del Estero
Catamarca
La Rioja
Córdoba
San Luis
Río Cuarto
Santa Fe
Paraná
Rosario
Asunción
Resistencia
Corrientes
Posadas
Pôrto Alegre
Montevideo
Buenos Aires
Avellaneda
La Plata
Río de la Plata
Junín
Bahía Blanca
Neuquén
Viedma
Comodoro Rivadavia
Golfo San Jorge
Golfo San Matías
Península Valdés
Puerto Deseado
Puerto Santa Cruz
Río Gallegos
Punta Arenas
Strait of Magellan
Falkland Is. (Islas Malvinas) (U.K.)
Stanley
Isla de los Estados
Cape Horn
Meseta de Somuncurá
Meseta de Las Vizcachas
Gran Bajo
Península Valdés
Golfo San Matías
Golfo San Jorge
Bahía Grande
Tierra del Fuego
Darwin
Ushuaia (Claimed by Arg. and Chile)
Aconcagua 6960
Tupungato (Vol.) 6800
Maipo (Vol.) 5323
Nevados de Cachi 6720
Cerro Bonete 6872
Nevados de Aconquija 5550
Cerro Mercedario 6770
Cerro Bonete 6759
Cerro Domuyo 4709
Chillán (Vol.) 3853
Copahue (Vol.) 3953
Tronador (Vol.) 3554
Lanín (Vol.) 3776
Payún (Vol.) 3715
Iguazú Falls
Paraná
Uruguay
Pilcomayo
Bermejo
Salado
Colorado
Negro
Chubut
Chico
Deseado
Coig
Gallegos
Santa Cruz
Viedma
Limay
Neuquén
Chimehuin
Huahum
Nahuel Huapi
Meliquina
Traful
Ñorquinco
Epuyén
Lepá
Pico Truncado
Puerto Madryn
Trelew
Rawson
Esquel
Gastre
Comodoro Rivadavia
Caleta Olivia
Las Heras
Río Mayo
Río Pico
Río Senguer
Río Chico
Río Fenix
Río Grande
Río Ewan
Río Cullen
Río Turbio
Río Gallegos
Río Coyle
Río Chico
Río Santa Cruz
Río Deseado
Río Chubut
Río Negro
Río Colorado
Río Salado
Río Bermejo
Río Pilcomayo
Río Paraguay
Río Paraná
Río Uruguay
Río de la Plata
Río Grande

In the middle latitudes of Argentina there are several distinct zones of flora and fauna. In the east, in the densely populated Humid Pampa where man has done most to change nature, the plains have been converted into farms and ranches; European plants and animals have been added to the indigenous species. To the west of the Humid Pampa is a dry, wind-eroded area with a sandy soil where hardy grasses and thistles grow. Encircling the Pampa on the north, west, and south is a region of scrubby, woody plants known as *monte*. Animals that flourish there include wildcats, pumas, foxes, weasels, skunks, ferrets, reptiles, deer, wild boars, the rhea, or American ostrich, and eagles. Farther west the flora and fauna decrease; desert plants prevail. In the Andean mountains the climate becomes cold and dry. Except for a few Alpine meadows, the plants are mostly tough grass and spiny shrubs; animals include the guanaco (the wild cameloid ancestor of the llama), the vicuña, the alpaca, and the llama (all members of the camel family), pumas, foxes, chinchillas, eagles, and condors.

To the south, on the cold, arid plains of Patagonia, the surface is covered with low shrubs, grasses, and carob trees. The native animals include guanacos, rhea birds, tinamous, wild turkeys, and parrots. The chief use of the land is for sheep grazing.

The humid mountains of the southern Andes nourish deciduous forests. There are both native and imported deer, which are hunted for sport, as well as wild boars, ferrets, pumas, and wildcats. The abundant birds include cormorants, wild turkeys, flamingos, ducks, black-necked swans, vultures, and eagles. The lakes and streams are stocked with trout and salmon for recreational fishing.

Traditional regions. Argentinians think of their country as composed of a number of distinct regions: the Pampas, Mesopotamia, the Chaco, the Northwest, the Pampean Sierras, the Cuyo, and Patagonia.

The prosperous east

The Pampas. The heart of Argentina lies in the fertile plains of the Provinces of Buenos Aires, southern Santa Fe, southeastern Córdoba, and eastern La Pampa. Until the middle of the 19th century this was a poor region, given over to range cattle and mule breeding. The development of the Pampas began with the building of railroads, progress in ocean transport, and the introduction of British beef cattle. The raising of livestock then became the main source of the country's wealth; the export of meat brought foreign capital to finance the further expansion of commerce and industry. The Pampas today have 60 percent of the country's highways and railroads, 60 percent of its population, and over 80 percent of its industrial establishments. In the 1970s there were a number of cities of more than 100,000 population, including Buenos Aires (3,000,000); Rosario (800,000), an industrial city and port on the Río Paraná; Córdoba (800,000), the rail and highway hub of central Argentina; Mar del Plata (300,000), a fishing centre and tourist resort on the Atlantic; and Bahía Blanca (162,000), an important harbour and industrial town. In the rural areas of the Pampas, the standard of living is high, and along with the wealthy ranchers there is a comfortable rural middle class.

Mesopotamia. The northern lowlands between the Paraná and Uruguay rivers have had a separate history because of their relative isolation. The rolling grasslands in the south provide pasture for much of the country's horses and cattle and are the leading sheep-raising area. The northern extension in Misiones province is known for its yerba mate (Paraguayan tea), its subtropical fruits, and its rain forests. Until the 20th century the Mesopotamian lands were sparsely settled; the population was largely Indian and mestizo (mixture of Indian and white), who in Corrientes province still largely own the cattle ranches. After World War II, however, heavy immigration of Germans and east Europeans settled on the land, along with Paraguayans, Brazilians, Swiss, Italians, French, Japanese, and Negroes. Each ethnic group maintains its own customs and traditions. On the Brazilian border in the north are the famous Cataratas del Iguazú (Iguazú Falls), where the broad Río Iguazú drops about 270 feet over a series of islands.

The provinces of the north and west

The Chaco. The Provinces of Santiago del Estero, Chaco, Formosa, and the northern parts of Córdoba and Santa Fe belong to South America's Gran Chaco ("hunting ground"), a great plain that extends into eastern Bolivia, western Brazil, and Paraguay. The slow development of transportation has held back its settlement, although the eastern part has colonies of Europeans who engage in ranching, lumbering, and the cultivation of cotton, wheat, and sorghum. The settlers were originally drawn by the quebracho trees, important as a source of tanning extract for the leather industry. Elsewhere the population consists largely of indigenous Indian tribes—the Tobas, Matacos, Chiriguanos, and others—who follow their traditional agricultural ways. The cities are situated on rivers or have grown up along the railroads and highways that followed the early lumberers inland.

Northwest. The Northwest of Argentina was the first part to be settled by the Spanish, who came across the Andes from Chile into what are now the Provinces of Jujuy, Salta, Tucumán, and northern Catamarca. The Inca culture still survives among the Indians and mestizo shepherds of the arid tablelands, who live meagrely with their goats, sheep, and llamas. In the valleys and piedmonts are cattle ranches, sugarcane plantations, citrus groves, and tobacco and vegetable farms, the owners of which are of Spanish descent. The provincial capitals are the old colonial cities of San Salvador de Jujuy (population 77,000), Salta (176,000), and San Miguel de Tucumán (306,000).

The Pampean Sierras. In the desert area of the western mountains are small oases where people of Spanish and European stock have pushed out the Indians, replacing them with isolated communities of small farmers and shepherds. In the second half of the 20th century, the dry mountain climate has drawn numbers of tourists, and many hotels and summer cottages have sprung up.

Cuyo. Also in the Andes are three provinces traditionally known as Cuyo—San Juan, San Luis, and Mendoza. In colonial times the area was closer both economically and politically to the central valley of Chile on the other side of the mountain passes than to the Atlantic coast. In 1884, with the completion of the first railway, its isolation was broken. Extensive irrigation has produced a flourishing agriculture specializing in vineyards and fruit trees, the crops going to eastern markets. Oil wells and copper, lead, and uranium mines contribute to the economy. The largest city is Mendoza, destroyed by an earthquake in 1861 and rebuilt with broad tree-lined streets and numerous parks and squares.

The south

Patagonia. The region south of the Río Colorado makes up more than a quarter of Argentina's territory, but it was for long regarded as a desolate place inhabited only by warlike Indians. After these were exterminated in the Indian campaigns of 1879–83, white settlers entered to breed sheep on the plateaus. The first settlers were mainly Welsh, Scots, English, and Chilean. The modern sheep ranches are very large and require relatively few workers except at shearing time. In the valleys of northern Patagonia, small farmers grow vegetables and fruit on irrigated lands. Halfway down the Atlantic coast is the oil port of Comodoro Rivadavia, the focus for the exploitation of natural gas and oil deposits on the adjacent Gulf of San Jorge. In western Patagonia the lakes and mountains near San Carlos de Bariloche are one of South America's most popular tourist resort areas.

THE PEOPLE

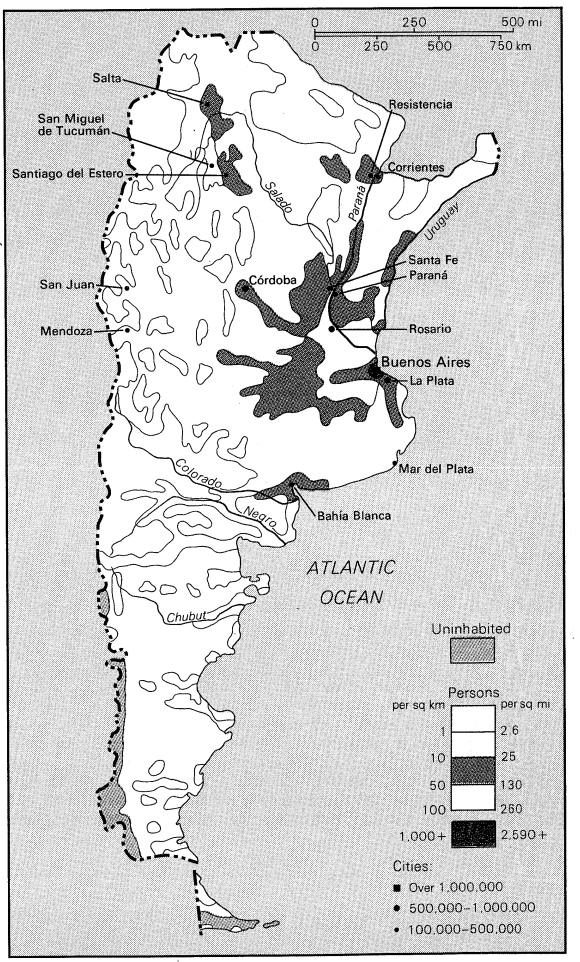
Four and a half centuries of European immigration have produced in Argentina a people who are predominantly white and European. The Indians and mestizos (mixture of Indian and white), who are so numerous elsewhere in Latin America, have been pushed aside or absorbed. The population estimates of the colonial period suggest that in 1800 Argentina had somewhat more than 400,000 people. Of these perhaps 30 percent were Indian and 10 percent were Negro and mulatto. There was a large element of mestizos, and whites were in a minority.

Many mestizos found their livelihood in working with

cattle, and in the mid-19th century they produced that nearly legendary figure, sometimes called half man and half horse, the gaucho. In his heyday the gaucho was scorned as a drinker and a rough vagabond. The colonization of the Pampas supplanted him with foreigners. After the gaucho—like the cowboy of the western United States—had virtually ceased to exist, he became a romantic figure in literature. This humble figure of the 19th century is now world famous as a symbol of nobility, selflessness, and courage.

Immigration in the late 19th century

The great wave of immigration that began after 1870 made Argentina a country of white people. The population grew from roughly 500,000 in 1816 to 1,800,000 at the time of the first census in 1869 to about 4,000,000 in 1895, 8,000,000 in 1914, 16,000,000 in 1947, 20,000,000 in 1960, and to over 23,000,000 in 1970. Indians are now thought to number no more than 100,000, and considerably fewer maintain their traditional ways of life. The Negro and the mulatto have disappeared, apparently absorbed into the dominant population. Mestizos are numerous in the provinces bordering Chile, Bolivia, and Paraguay, but the urban areas are almost entirely white.



Population density of Argentina.

The population growth of the late 19th century was the result of European immigration, a net total of 3,500,000 being added between 1857 and 1939. Most of the Europeans were Italians (44 percent) and Spaniards (31 percent), but substantial numbers came from France, Poland, Russia, and Germany. In 1914 the foreign-born population was 30 percent of the total, and in large cities the foreigners outnumbered the natives by as much as two to one.

Languages and dialects

Spanish remains the national language, although it is spoken in many accents. Numerous foreign languages and dialects are to be heard, from Basque and Sicilian to Welsh and Gaelic. Mexicans, Bolivians, Chileans, and Paraguayans speak varying dialects of Spanish. An underworld language called Lunfardo has developed in Buenos Aires, composed of words from many languages, among them Italian, Portuguese, Spanish, French, German, and African. This argot has found its way into popular songs and the theatre.

Some Indian languages are still spoken in certain areas. Guaraní is often spoken as a second language in Mesopotamia and is widely used in the coastal folklore. Toba, Chiriguano, Mataco, Mocoví, and Chorotí are spoken in the Chaco; in the northwest, Lule, Tocontés, Atacama, and Diaguita; and in the south, Araucanoan, Ranquel, Moleche, and Tehuelche. In Tierra del Fuego there are a few speakers of Yámana, Ona, and Selknan.

Argentina is now an Atlantic and urbanized country, increasingly middle class in outlook. Greater Buenos Aires had 8,350,000 inhabitants in 1970, over a third of the country's population. Over 70 percent of the people live in towns with more than 2,000 inhabitants. To the flow of immigrants from overseas, since mid-20th century, has been added another flow from the countryside to the city, drawing with it many peasants from Chile, Bolivia, Paraguay, Brazil, and Uruguay.

Argentina, Area and Population

	area		population	
	sq mi	sq km	1960 census	1970 census*
Federal district				
Distrito Federal	77	200	2,967,000	2,972,000
National territory				
Tierra del Fuego†	7,873	20,392	8,000	16,000
Provinces				
Buenos Aires	118,843	307,804	6,766,000	8,775,000
Catamarca	38,540	99,818	168,000	172,000
Chaco	38,468	99,633	543,000	567,000
Chubut	86,751	224,686	142,000	190,000
Córdoba	65,161	168,766	1,754,000	2,060,000
Corrientes	34,054	88,199	533,000	564,000
Entre Ríos	29,427	76,218	805,000	812,000
Formosa	27,825	72,066	179,000	234,000
Jujuy	20,548	53,219	241,000	302,000
La Pampa	55,382	143,440	159,000	172,000
La Rioja	35,649	92,331	128,000	136,000
Mendoza	58,239	150,839	824,000	973,000
Misiones	11,506	29,801	361,000	443,000
Neuquén	36,324	94,078	110,000	155,000
Río Negro	78,383	203,013	193,000	263,000
Salta	59,759	154,775	413,000	510,000
San Juan	33,257	86,137	353,000	384,000
San Luis	29,632	76,748	174,000	183,000
Santa Cruz	94,186	243,943	53,000	84,000
Santa Fe	51,354	133,007	1,885,000	2,136,000
Santiago del Estero	52,222	135,254	477,000	495,000
Tucumán	8,697	22,524	744,000	766,000
Total Argentina	1,072,157	2,776,889	20,011,000‡§	23,364,000

*Provisional. †Area and population are for the "sector continental" of the 1960 census: Bahia Thetis, San Sebastian, and Ushuaia. They exclude the area (1,247,803 sq km) and population (3,000) of those areas not under direct Argentine administration: Antartida Argentina, Islas Malvinas (Falkland Islands) and the south Atlantic Islands claimed by Argentina. ‡Census result excludes adjustment for underenumeration. Adjusted total is 20,941,000. §Figures do not add to total given because of rounding.
Source: Official government figures.

As the country has become more urban, its growth in numbers has tended to become slower. Births declined from an average of over 35 per 1,000 of the population in 1914–15 to about 22 in the late 1960s. Death rates also declined, but since 1950 they have fallen less rapidly than birth rates. Argentina's population grew at an average rate of 1.5 percent a year in the late 1960s.

Most of the people are nominally Roman Catholics. In the early 1970s about 95 percent of the population claimed membership in this official religion of the country; there were also about 400,000 Protestants, 90,000 adherents of the Eastern Orthodox Church, and 80,000 Jews. In practice, religious freedom prevails, and the various faiths have their own churches, cemeteries, hospitals, and social centres. Education is largely the province of the state, although there are a number of religious schools. Various religious sects, including spiritualists, thrive in the cities and suburbs. (A.Si.)

THE ECONOMY

Argentina is one of the most prosperous countries in Latin America. Output per capita was over $800 in the early 1970s, the second highest within the region after Venezuela. On a worldwide basis, however, Argentina's economy has lost ground steadily since World War II, although it was still among the first 30 in output per capita in 1970. It had the basic essentials for rapid economic growth, but political difficulties and policy mistakes caused wide swings between boom and recession, and overall expansion was slow. During the 1960s real growth averaged 3.5 percent a year, while output per capita increased by just under 2 percent a year.

Post-war inflation

The increase in government spending during the Perón years touched off a vast inflation that continued through the 1960s and into the 1970s. The Argentine peso, which stood at 5 to the dollar in 1955, underwent a series of devaluations until in 1970 its parity was 400 to the dollar. The cost of living in Buenos Aires rose from a base of 100 in 1943 to 680 in 1955. Later estimates using 1960 as the base year placed the level at nearly 700 in 1970 (with corresponding index numbers of 3 for 1943 and 20 for 1955). This means that 700 pesos had approximately the same purchasing power in 1970 that three pesos had in 1943.

Unlike its neighbours, Argentina is not, strictly speaking, a developing country, although agricultural products make up the bulk of its exports. These are not, however, tropical raw materials for the developed nations' industries, but products of a temperate-zone economy, such as beef, mutton, and wheat. Argentina is the world's third biggest producer of beef; it is fourth in the production of wool and wine, and about twelfth in the output of wheat.

Foreign trade plays a comparatively small part in the creation of Argentina's national income. In terms of imports and exports per capita, Argentina's position as a trading nation is near the world median; three Latin American nations, Venezuela, Chile, and Paraguay, lead it. Within the Latin American Free Trade Area, of which Argentina is a member, however, it plays a predominant role as an exporter of manufactured goods. When the country was isolated from its traditional sources of supply of manufactures—western Europe and North America—by the distribution of shipping in World Wars I and II, there was a rapid growth of domestic manufacturing. As a result, Argentina became self-sufficient in most consumer goods and in an increasing range of capital goods. It is the second most important oil-producing country in Latin America, after Venezuela. Its manufacturing sets Argentina apart from other developing nations by giving it a balanced structure of agriculture and industry. This balance is reflected in the high standard of living enjoyed by the average Argentinian and in the absence of the large-scale poverty prevalent elsewhere in Latin America.

Natural resources. *Minerals.* Much of Argentina's mineral wealth still remains to be explored. With the exception of oil and natural gas, the mineral reserves being exploited are generally small and widely scattered. Argentina imports most of its iron ore, coke, and coal. Fairly rich reserves of lead, zinc, silver, gold, copper, tin, cobalt, bismuth, vanadium, beryllium, tantalite, manganese, tungsten, and uranium are known to exist, but only some of these are being worked. A wide range of nonmetallic minerals are also found.

The most important coal deposits are those of Río Turbio, in the Province or Santa Cruz, but these are of low calorific value and far away from industrial users. Other deposits exist in the Provinces of Jujuy, La Rioja, Mendoza, and San Juan.

The iron-ore deposits of Sierra Zapla, in the northwestern Province of Jujuy, are low in iron content but provide the raw material for the Zapla steelworks. The biggest known reserves of iron ore are situated in the Sierra Grande, in Río Negro province; their development was planned for the 1970s.

The exploitation of zinc, tin, gold, silver, beryllium, sulfur, and asphalt is on a small scale. Salt and limestone are produced mainly for local consumption.

Oil and natural-gas resources have been intensively explored and developed. Argentina is almost self-sufficient in crude petroleum and petroleum products. All petroleum reserves belong to the state, which controls the industry through the autonomous corporation Yacimientos Petroliferos Fiscales (YPF). Foreign companies have been allowed to participate in the industry as agents of the YPF. The main producing oilfields are in the Patagonian Provinces of Chubut and Santa Cruz, around Comodoro Rivadavia, and in Mendoza, Neuquén, Río Negro, and Salta. Production increased rapidly during the 1960s, with a parallel development of pipelines and refining capacity. By the early 1970s some 2,000 miles of trunk pipelines were in use. The natural-gas industry is the most highly developed in South America. The main gasfields are in the northern, western, and southern provinces; an extensive pipeline system links them with the industrial centres. Some 3,720 miles of gas pipeline were in use by the early 1970s, including a long line from Comodoro Rivadavia to Buenos Aires.

Importance of the Pampas

Agricultural resources. Argentina's long north–south expanse contains a wide variety of climates and soils. Cereals are the main crop, particularly wheat and corn (maize), which are grown principally in the Pampas, the fertile plain extending over a radius of 600 miles from the city of Buenos Aires. This region also leads in livestock. Wheat, corn, and beef from the Pampas account for the bulk of Argentina's export earnings. Crops and cattle in other parts of the country are produced mainly for domestic consumption. Sugarcane is grown in the Province of Tucumán; vines, fruit trees, and olive trees are grown in Jujuy, Catamarca, La Rioja, San Juan, Mendoza, and Río Negro. The big citrus plantations are in the northwestern and northeastern provinces, which also produce tobacco. Tea, tung (an oil-producing tree), and mate (Paraguay tea) are grown in Misiones; cotton in western Chaco and northern Corrientes.

Livestock has played a major part in Argentina's export economy. By the early 1970s the number of cattle stood at about 50,000,000. Meat cattle are mostly the British breeds, Aberdeen Angus, Hereford, and Shorthorn, and the French Charolais. Patagonia, in the south, is a land of sheep breeders; in the early 1970s the flocks numbered close to 45,000,000, making Argentina fourth among the world's wool producers after Australia, the Soviet Union and New Zealand.

There are large forest reserves, mostly in the subtropical northeastern provinces and the south. The Chaco area produces most of the hardwood, particularly quebracho, which provides an extract used in tanning. The timber industry is very small.

Hydroelectric and other power resources. Supplies of power have not increased fast enough to meet the country's industrial needs. The vast hydroelectric potential in the rivers of the western mountains and the northeast was yet to be exploited in the 1970s. By the end of the 1960s, installed hydroelectric capacity amounted to about 600,000 kilowatts, mostly in the Provinces of Córdoba and Mendoza. Another 1,650,000 kilowatts were expected to become available in the 1970s at El Chocon-Cerros Colorados, in northern Patagonia. Most of the electricity is supplied by thermoelectric plants burning oil, coal, and natural gas. Total capacity at the beginning of the 1970s amounted to 6,300,000 kilowatts. Many private firms built their own generators to overcome power shortages, particularly during the early 1960s.

Sources of national income. *Agriculture and fishing.* Agriculture and fishing jointly contributed 15 percent of the gross domestic product in 1969. During the 1960s these sources grew less rapidly than domestic output as a whole, and their share of the total declined. By the early 1970s they were also providing proportionately less employment than they did in 1960, the agricultural labour force having fallen from 18 to 15 percent of total employment. The small part played by agriculture in the formation of domestic product and in providing jobs sets Argentina apart from the rest of Latin America and from the developing countries in general. A characteristic feature of Argentina's agriculture is the important part

played by very large farms, particularly in grain and cattle breeding. Although holdings covering more than 2,470 acres accounted for less than 10 percent of the total number of farms, they occupied 75 percent of the total area in use. Half of Argentina's total land area is devoted to pastures, 23 percent to forests, and only 12 percent to crops. Of the total crop area about one-quarter is occupied by wheat and 15 percent by corn. In the early 1970s about 16,000,000 tons of cereals were produced annually, along with 15,000,000 tons of industrial crops (mainly sugarcane), 10,000,000 tons of forage, 4,000,000 tons of vegetables, 3,250,000 tons of fruit, and 2,000,000 tons of oilseed.

Grain and cattle

In the 1960s agricultural production rose at an average annual rate of 2 percent, a very slow rate in comparison with the trend in neighbouring Brazil and Bolivia. After World War II, government policies favoured industrial rather than agricultural development, particularly during the regime of President Perón. As a result, productivity is lower than it might be, as demonstrated by the performance of the more efficient farms. Argentina is nevertheless well ahead of its neighbours in the use of farm machinery and fertilizers. Tractors are produced domestically, but a large proportion of the fertilizers have to be imported.

Beef production amounted to 1,400,000 tons annually in the early 1970s, of which 735,200 tons were exported. As domestic consumption grew, the proportion available for export declined. In the early 1970s the average Argentine consumed more meat than anyone else in the world: over 200 pounds (90 kilograms) per year. The government has attempted to increase exports through various incentive schemes. Sheep are reared mostly for their wool, although mutton and lamb production is also important. There is a large export trade in cattle hides and sheepskins.

The fishing industry expanded rapidly during the 1960s. The catch of sea fish rose from 92,000 tons in 1960 to 182,339 in 1970, mostly hake and anchovies from the Mar del Plata area. A rise in demand for fish, both for human consumption and for the pig and poultry industries, encouraged new investment in fishing. Argentina is well behind other Latin American countries, particularly Chile and Peru, in the development of its abundant fishing resources.

Mining and quarrying. The contribution of the mining and quarrying industries to gross domestic product is small: it rose from 1.1 to 1.7 percent in the 1960s. Annual crude petroleum output, however, in the early 1970s reached 24,000,000 tons, and Argentina was self-sufficient in oil by 1972. Natural-gas production in the early 1970s reached 10,000,000 cubic yards annually, but potential demand far outstripped supply in the early 1970s and considerable imports were needed.

Oil and natural gas

Production of coal in the early 1970s amounted annually to about 615,000 tons. Output satisfied only a small part of domestic requirements. Argentina continued to depend on coal and coke imports because its own supplies are both inadequate and unsuitable for such major uses as steelmaking.

Manufacturing. Manufacturing grew rapidly after World War II. It accounted for 35.1 percent of the annual gross domestic product in the late 1960s and employed 30 percent of the labour force. The initial impetus to its expansion came from World Wars I and II, when many imported products were scarce. By the early 1970s Argentina was producing practically all the consumer goods, including cars, refrigerators, and television sets, that it once imported. It also was manufacturing a wide range of investment goods, among them iron and steel products and industrial machinery and equipment.

During the 1960s manufacturing output as a whole increased at an average annual rate of 5 percent. The automobile, metal, and chemical industries expanded by an average 8 or 10 percent a year. The increase would have been more rapid but for a chronic shortage of capital that delayed modernization in the older industries and held back investment in the new industries. Argentina's manufacturing also experienced difficulty in meeting international competition; many firms in the older industries were too small to be economically viable and owed their continued existence to high tariff walls.

Energy consumption rose rapidly during the 1960s, outpacing supply so that there were recurrent shortages. The energy bottleneck was partly overcome through oil and natural-gas production. About a quarter of the natural gas marketed through the state corporation Gas del Estado is sold to thermal power plants; the rest is divided between the cement, metallurgical, and petrochemical industries. The annual output of electricity from public utilities amounted to 16,900,000,000 kilowatt hours in the early 1970s and was rising at an average annual rate of 9 percent. Hydroelectric resources accounted for less than 10 percent of the electricity generated, but that proportion was expected to rise. Coal plays a secondary part in energy production: the main users are the railways, the Buenos Aires power stations, and the state-owned steelworks at San Nicolas. The discovery of uranium encouraged the development of atomic energy, and a 318,000-kilowatt nuclear power plant was being built at Atucha in Buenos Aires province in the early 1970s.

Energy consumption

Financial services. Argentina's financial and banking system is highly developed. The banking system includes banks owned by the national government, by provincial and municipal authorities, and by private Argentine and foreign companies. The central bank, which acts as the government's banker, is the sole currency-issuing agency, as well as the custodian of the gold and foreign-exchange reserves, and regulates the volume and uses of bank credit. The regulation of bank credit has been employed particularly in limiting credit for import financing operations. Private financial institutions known as *financieras* also play an important part in providing short-term finance. Investment capital became so scarce during the 1960s that the government-owned National Development Bank was virtually the only source of medium- and long-term credit. Economic reforms and a slowdown in the rate of inflation during the late 1960s brought a rise in foreign investment.

The Buenos Aires stock exchange is the second largest in Latin America after that of Mexico. There is some trading on seven other exchanges, mainly in Córdoba, Mendoza, and Rosario.

Foreign trade. Exports grew at an average annual rate of 4.5 percent in the 1960s, reaching an annual peak of $1,740,000,000 in 1971. Imports grew at a rate of 2.6 percent, reaching $1,888,000,000 in 1971. These figures conceal an actual decline in the share of exports in the gross national product, which had to be offset by import restrictions. The declining share of foreign trade was related to the fact that at least 90 percent of Argentina's exports consisted of agricultural products. Some new manufactured products entered the export trade after 1960, accounting for 13 percent of annual earnings. They consisted mainly of meat and steel products, calculating machines, motor vehicles, and books. The principal export products, however, continued to be meat and cereals, each accounting for about one-quarter of total annual export earnings at the beginning of the 1970s. The government's policy of encouraging domestic substitutes for imported goods was clearly reflected in the composition of imports: those products that could be made locally were being gradually phased out of the import list. By the end of the 1960s raw materials and semiprocessed goods comprised 60 percent of total imports, and machinery and equipment about 30 percent.

The crisis in exports

Argentina's main export market is in the countries of the European Common Market, which accounted for 48 percent of annual foreign sales at the beginning of the 1970s. The Latin American Free Trade Area, mainly Brazil and Chile, was the second-largest buyer, followed by the United States. Argentina tended to buy where it sold. The main suppliers were the Common Market countries, LAFTA, and the U.S.

Management of the economy. The Argentine people have traditionally favoured a system of free enterprise, private capital, and investment by foreigners, but Argen-

tine governments have sometimes taken a different attitude. The regime of President Juan Perón (1946–55) encouraged state participation in many areas of the economy, most of which were later restored to private control. The government maintains a predominant share in the oil and natural-gas industries, the coal and steel industries, the railways and other transport services, telecommunications, electric and nuclear power, and banking. It also seeks to influence economic activity elsewhere through export and industrialization incentives, import controls and tariffs, price controls, subsidies, support prices, and minimum-wage legislation. Occasional outbursts of ultranationalism and distrust of foreigners among the population have made difficulties for foreign investors, notably in the oil industry.

Taxation. Large budget deficits were a major cause of the high rates of inflation during the 1960s. After 1966 efforts were made to cut down on public expenditure, and the budget deficit was reduced from 1,400,000,000 new pesos in 1966 to an estimated 428,000,000, in 1970.

Kinds of taxes

Tax-collecting methods were improved; a complete overhaul of the tax system was being planned for the 1970s. The overall tax revenues were estimated at 9 percent of the gross national product in 1969.

Argentina levies taxes at both the federal and the provincial level. The main federal taxes early in the 1970s were the personal income tax, with rates ranging from 12 to 46 percent, a corporate profits tax of 33 percent, a capital gains tax, a substitute inheritance tax, a sales tax, a stamp tax, and customs and excise duties. The main provincial taxes are turnover taxes on commercial transactions, stamp taxes, and land taxes. The income and corporate profits taxes yielded 26 percent of total tax revenues annually in the late 1960s, and the sales tax 24 percent.

Trade unions. A large proportion of the labour force is unionized, the unions having acquired much of their strength during the Perón administration. Their political influence was somewhat reduced during the second half of the 1960s, but it remained considerable. Most unions are affiliated with the General Confederation of Labour, which negotiates directly with the government on such matters as minimum wages. Labour conditions generally are regulated by a combination of legislation and collective agreements. The collective agreements are usually negotiated between unions and employers' associations on an industry-by-industry basis, rather than between individual unions and firms. The number of working days lost through strikes reached a peak in the politically disrupted mid-1960s but fell sharply afterward.

Policies and prospects. The various governments of the late 1960s pursued three main economic objectives: higher growth rates for the economy as a whole, a reduction in the rate of inflation, and improvements in the external payments position. These objectives also formed the basis for the 1970–74 development plan. An acceleration of growth rates had already occurred in the late 1960s, so that the plan's target of a 5.5 percent annual increase in real domestic product seemed well within reach.

By the beginning of the 1970s private investment spending and exports had overtaken consumer expenditure and public investment as the most dynamic sectors of the economy. The runaway inflation of the early 1960s had also been slowed down through the stringent controls over prices and minimum wages and a reduction in budget deficits. The government relied on export incentives, import controls, and new legislation to attract and protect foreign investment. These measures were backed by a further devaluation of the peso in June 1970.

Shortage of capital

At the beginning of the 1970s, large areas of Argentina's economy remained undeveloped or underdeveloped. Progress had been made in creating a favourable climate for foreign capital, both private and public, but the bulk of the country's financial needs would still have to be met from domestic resources. The soaring cost of living during the 1960s, when increases averaged over 20 percent a year, made saving unattractive to both individuals and corporations. While the rate of inflation was reduced after 1968, prices were still rising by 10 percent a year early in the 1970s. The lack of confidence this condition engendered kept total savings well below the country's requirements. Another problem was the low level of productivity, in both agriculture and industry.

Despite these difficulties, the economy was basically strong. The Argentine people enjoyed an average standard of living well above that achieved anywhere else in Latin America. (E.I.U.)

TRANSPORTATION

Dominance of the railroads

Argentina's transportation system developed out of the railroad boom of the late 19th century, when British and Argentine investors built lines radiating from Buenos Aires across the Pampas. The railroads were not designed as a unitary system but were meant primarily to carry the produce of the hinterland to the Atlantic ports of Buenos Aires, Rosario, and Bahía Blanca. Communication between the towns of the interior remained difficult, and their isolation was enhanced by the lack of an adequate highway system. In 1971 the country had only 20,753 miles of paved highway as compared to 29,000 miles of railroad lines—mainly in the Pampas.

An important adjunct to the railroads and highways are the rivers of the Atlantic coast, with a navigable length of over 2,000 miles. These include the Río de la Plata and its tributaries, the Paraná, the Paraguay, and the Uruguay, that run northward from Buenos Aires. There is also an important coastal shipping trade on the Atlantic.

Argentina had 28 international airports in 1970, the most important being Ezeiza near Buenos Aires and Aeroparque in the city itself. There were 39 domestic airports handling commercial traffic.

ADMINISTRATION AND SOCIAL CONDITIONS

The structure of government. Argentina is a federal union of 22 provinces, one territory, and a federal capital (coterminous with the city of Buenos Aires). The country arrived at its federalism after a long struggle between supporters of central government and defenders of local interests. The constitution of 1853 was modelled after that of the United States, and—although it was modified in 1860, 1866, and 1898, replaced by a new constitution in 1949–57, and again modified in 1972—the Argentine system has continued to be republican, representative, and federal.

The national government is composed of an executive, a legislative, and a judicial branch.

Executive power is vested in a president and vice president who are elected for a term of four years and may succeed themselves to a second term. Since 1972 they are elected by direct popular vote. The president is commander in chief of the armed forces and appoints all civil, military, and judicial officers. The vice president serves as president of the senate.

The national congress consists of two houses: the senate, composed of three representatives from each province and the federal capital; and the chamber of deputies in which the number of representatives from each province is proportional to the population. The senators are chosen by direct vote to terms of four years and may be re-elected indefinitely.

The judges of the federal supreme court, the federal courts of appeal, and the federal district courts are appointed for life by the president with the approval of the congress.

Each province has its own government, with an executive, a legislature, and a judicial system corresponding to those of the federal government. The executive of the federal capital and the governor of the territory of Tierra del Fuego are appointed by the president. The provinces are divided into districts or departments.

Voting is compulsory for all citizens over 18 years of age except clergymen, members of the regular army, and those deprived of the vote for legal reasons. The ballot is secret. Argentina has a number of political parties. The Conservative Party has traditionally been supported by the landowners. The Radical Party has a middle class

Political parties

following. Others include the Christian Democrat Party, the Progressive Democrat Party, the Socialist Party, the Worker's Party, and the Peronist Party with a mixed following of nationalists and trade unionists. All the parties compete for popular support.

From 1916 to 1971 the country had 19 presidents, 11 of them military men and eight of them civilians. Of the civilian presidents only two completed their six-year terms. Of the military presidents, nine were deposed by revolutionary coups. The civilian presidents governed for 25 years and the military presidents for 30 years from 1916 to 1971. At various times the constitution has been suspended, and the political parties have been suppressed. In 1973 the civilian candidate of the Peronist Party, Hector Cámpora, was elected on a program endorsed by former President Juan Perón. After Cámpora's resignation in his favour, Perón was elected to the presidency in September 1973; his third wife, Isabel, became the first woman vice president in Latin America.

The army, navy, and air force play an important part in Argentine life. Along with their functions of defense they also help to maintain roads, airports, and waterways. Despite the political role of many military leaders, the defense budgets since World War II have been rather small in proportion to the gross national product.

Education and social services. In numbers of schools and the supply of social services, Argentina is among the leading countries of Latin America.

Education. Primary education is compulsory and free for all children from six to 14 years of age. Secondary and higher education is also available in free public schools and in private schools subsidized by the state. In the early 1970s there were more than 3,000 preprimary schools, over 20,000 primary schools, more than 4,000 secondary schools of various kinds, and 495 universities, teacher-training schools, and other institutions of higher education. The oldest university is the University of Córdoba, founded in 1613. Other major universities are at Buenos Aires, La Plata, Santa Fe, Tucumán, and Cuyo.

Welfare and housing. Argentina's social welfare services were developed on a large scale during the presidency of Juan Perón (1946–55). A system of social security was set up providing pensions upon retirement for all workers, health and maternity benefits, compensation for workers injured on the job, severance pay for those losing their jobs, family allowances for the unemployed, and financial assistance to the aged and indigent.

There is an extensive system of hospitals and clinics run by national, provincial, and local authorities as well as by private organizations. In the early 1970s there was one hospital bed for every 160 persons and one doctor for every 504 persons. Public-health and sanitation standards were high, particularly in urban areas and developed regions. Endemic diseases, such as smallpox, cholera, yellow fever, and tuberculosis, that had been great scourges a century earlier were brought under control. Vaccination against poliomyelitis is common.

Urban slums

Housing became a problem in the cities as the influx of workers from rural areas outran the pace at which new living quarters were being built. Large ramshackle communities built from corrugated iron, known as *villas de miserías*, sprang up on the outskirts of the cities, and the efforts of various agencies—national and municipal housing authorities, unions, cooperatives, and private individuals—were not sufficient to keep the situation from growing worse in the 1960s.

Social divisions. Under Argentina's constitution all its citizens are equal without regard to race or ancestry. As in other countries, however, there are various distinctions among social classes and groups: between city people and peasants, or *campesinos;* between inhabitants of Buenos Aires (*porteños*) and provincials; between rich and poor, well-to-do and needy. There are ranchers, or *estancieros*, who own great expanses of land and vast herds of cattle but reside in the cities. There is also a rural middle class consisting of those who own or rent land and live on it, working it with their families and a few employees.

The landed aristocracy has gradually been giving up its power to the industrial and commercial class, to high-ranking officers, and to eminent professional men. At the same time considerable power has fallen into the hands of the working class with its strong unions, able to negotiate with employer associations and political leaders. There is also an urban middle class made up of professionals, minor officials, and small businessmen.

CULTURAL LIFE AND INSTITUTIONS

Argentina's European outlook

Argentina's cultural life has always had a European orientation. The basic Spanish heritage has been adopted by the immigrants of all nationalities. As in other Latin American countries, many aspects of art and culture follow French models. Buenos Aires, with its 60 art galleries and numerous museums and concert halls, is the centre of cultural life; but the larger cities of the country are also well endowed with libraries, museums, theatres, and concert halls. Argentines read about 150 domestic newspapers.

The traditions of the various nationalities remain strong, being kept alive in fairs and fiestas. There is also much regional folklore, for which there are a number of museums. At various times of the year local festivals are held, such as the Fiesta del Poncho in Catamarca, where woven ponchos of vicuña, alpaca, and llama are displayed; the Fiesta de la Vírgen del Valle in the northwest, which attracts pilgrims from other parts of the country; the Quebradeños carnival, characterized by marching, dancing, and singing; the Misachichos Valleros, a religious and folkloric festival which combines elements of Christianity and paganism; and the National Folklore Festival in Cosquín.

BIBLIOGRAPHY. General works containing useful guides to further reading include GEORGE PENDLE, *Argentina*, 3rd ed. (1963); YSABEL F. RENNIE, *The Argentine Republic* (1945); and JAMES R. SCOBIE, *Argentina: A City and a Nation* (1964). Introductions to Argentina's geography and economics include PIERRE DENIS, *The Argentine Republic: Its Development and Progress* (1922); and PRESTON E. JAMES, *Latin America*, 4th ed. (1969). More detailed material may be found in FRANCISCO DE APARICIO and HORACIO A. DIFRIERI (eds.), *La Argentina: Suma de geografía*, 9 vol. (1958–63); and FEDERICO A. DAUS, *Geografía de la Republica Argentina*, 2 vol. (1954) and *Geografía y Unidad Argentina* (1957). The UNITED NATIONS ECONOMIC COMMISSION FOR LATIN AMERICA has produced useful studies of contemporary economic problems. Material on cultural life and institutions may be found in ISIDRO J.F. CARLEVARI, *La Argentina*, 3rd ed. (1967).

(A.Si.)

Argentina, History of

It has been estimated that the population living in the land area of what is now called Argentina before the arrival of the Europeans totalled some 300,000. It was comprised of Indians believed to have been descendants of those Asians who, in prehistoric times, are thought to have migrated across the Bering Strait to the North American continent and who gradually spread across North, South, and Central America. The state of civilization of the Argentine Indian did not match those of the Aztec and the Maya and the empire-building Incas. Down to the late 18th century, the European population was concentrated in the northern part of the country, which had been settled chiefly by the overflow from the neighbouring Spanish colonies of Chile, Peru, and Paraguay (Asunción). A simple but vigorous society developed on the basis of Indian labour and the horses, cattle, and sheep imported by the Spaniards, as well as native products, such as corn and potatoes. Some of the Indians worked as virtual serfs (not slaves) under the encomienda system, others in free villages, and others in missions established by the Roman Catholic Church. (See LATIN AMERICA AND THE CARIBBEAN, COLONIAL.)

Politically, Argentina was a divided and subordinate part of the viceroyalty of Peru until 1776, when the territories in the Plata Basin (modern Argentina, Uruguay, Paraguay, and southern Bolivia) were detached and made into the new viceroyalty of La Plata, with Buenos Aires as its capital. Spain created the new viceroyalty chiefly to strengthen the territories against a potential threat from the Portuguese colony of Brazil.

In Argentina, as in most of Spanish America, the independence movement was precipitated mainly by Napoleon's intervention in Spain, beginning in 1808. On May 25, 1810, an autonomous government was set up in Buenos Aires to administer the viceroyalty of La Plata in the name of Ferdinand VII, pending his restoration. After his restoration, in 1814, an assembly representing most of the country met at San Miguel de Tucumán and, on July 9, 1816, declared the country independent under the name of the United Provinces of the Río de la Plata.

Several years of hard fighting followed before the Spanish royalists were defeated in northern Argentina. But they remained a threat from their base in Peru, until it was liberated by Simón Bolívar in 1824. The Buenos Aires government tried to maintain the integrity of the old viceroyalty of the Río de la Plata, but the outlying portions, never effectively controlled, soon were lost: Paraguay in 1814, Bolivia in 1825, and Uruguay in 1828. The remaining territory, modern Argentina, was frequently disunited until 1860. The root cause of the trouble, the question of the relation of Buenos Aires to the rest of the country, was not settled until 1880, and, even after that, it continued to cause dissatisfaction, though never again to the point of disrupting the union.

After 1820 there was general agreement that the government should be republican but violent disagreement as to whether it should be a centralized or federal form of rule. There was also disagreement over whether the revolution should stop with the winning of independence or be continued within the country with a view to reforming Argentine society.

EFFORTS TOWARD RECONSTRUCTION, 1820–29

In 1820 there were only two political organizations that could claim more than strictly local and provincial followings: the revolutionary government in Buenos Aires and the Liga de los Pueblos Libres (League of Free Peoples) that had grown up in the lowlands along the Río de la Plata and its tributaries under the leadership of José Gervasio Artigas. But both organizations collapsed simultaneously in that year, and there remained the 13 provinces. Buenos Aires seemed on the verge of losing its position as the seat of national government along with the general collapse of centralized power, but, as the city regained its function as an indispensable intermediary between the nation and foreign governments, it also regained its former position of national prominence.

The collapse of central government

Dominance of Buenos Aires. In practically all the provinces, military leaders had assumed power. Portions of the old national army took over in the provinces of Córdoba and San Juan; garrison troops did likewise in the Indian frontier regions of Santa Fe and Santiago del Estero. Local militia with experience in the wars against the royalists took power in Salta, and others with no significant experience at all took power in Mendoza; and the remains of the old dissident federal army rose up in Entre Ríos. Military power proved to be an insufficient base for the new political regimes, however, and their foundations had to be expanded to include other elements. Thus, each provincial political regime soon acquired its own character, according to the relative power held by military strong men and by local political interests.

This differentiation was not, however, cause for friction among the provinces; their cleavages were more directly attributable to economic and geographical differences. In the central interior region, the dominant attitude of all the provinces was one of suspicion toward Córdoba, the most important among those provinces and the one aspiring to national leadership. In the north, the rivalries were even more complex; while, in the Litoral (the lowlands along the tributaries of the Río de la Plata), the charismatic caudillo (chief), Estanislao López, guaranteed the supremacy of Santa Fe, in league with leaders from Corrientes. The Province of Entre Ríos was exhausted from war and immobilized by internal tensions; its unpopular governor, a career officer, Lucio Mansilla, was satisfied with retaining his own power.

Buenos Aires was able to make significant advances toward national leadership by prudently taking advantage of these interprovincial rivalries. Its leaders supported the Andean provinces from Mendoza to Salta against the dominance of Córdoba; in the Litoral they counted on the friendship of Santa Fe, which welcomed a peaceful arrangement with its powerful neighbour.

Within the Province of Buenos Aires itself, the regime of the so-called Partido del Orden (Party of Order) put into effect reform measures that were viewed with favour both locally and internationally. Principal among them was the dismantling of the military apparatus that had remained from the war and the dedication of what remained to the defense of the frontier areas and plains (recognized as the country's greatest source of potential wealth) against the Indian threat. This prudence on the part of the government in power won the support of the rural landowners as well as the urban businessmen, whose backing allowed the regime to confront the opposition in the city of a popular party loyal to the memories of revolutionary war. Victory at the polls was assured in the rural areas by the loyalty of the landowners who could always deliver the local electorate; in the city, it was won by electoral manipulation.

But the political order that seemed to be taking hold in Buenos Aires and in the country as a whole was more fragile than its successes would indicate. It was achieved by setting aside, rather than resolving, some fundamental difficulties. In particular, the institutional organization of the country was not carried out, and nothing was done about the Banda Oriental (the east bank of the Uruguay River), which was occupied first by Portuguese and then by Brazilian troops. By 1824, both problems were becoming urgent. England was willing to recognize Argentine independence if a government was established that could act for the whole country. As for the Banda Oriental, some action became necessary when a small group of eastern patriots took over large sectors of the rural areas and agitated for their reincorporation into the Provincias Unidas del Río de la Plata (United Provinces of the Río de la Plata), forcing the Buenos Aires government to face the possibility of war with the Brazilian Empire.

Another problem of the 1820s was that of achieving financial stability in order to obtain capital from investors in Europe, particularly in Great Britain. After a brief period of high optimism on the part of foreign investors, Buenos Aires was left with a considerable debt from a loan that had been arranged in London, the proceeds of which were consumed in the war of 1825–27 with Brazil.

Presidency of Rivadavia. In the meantime an attempt was made to establish a national government through a constituent assembly that met in December 1824. This assembly saw an attempt by the more sophisticated of the provincial political factions to escape from the controls of the rural or military caudillos (chiefs) with whom they had formerly collaborated. Some sought to unseat the very governments they were representing, and these ambitions only served to sharpen provincial opposition to the assembly. In February 1826, overstepping its legal authority, the constituent assembly created the office of president of the republic and installed the *norteño* (northerner) Bernardino Rivadavia as its first occupant. The new president placed the entire Province of Buenos Aires under his direct authority. Civil war flared up in the interior provinces, soon dominated by the military influence of Juan Facundo Quiroga—the caudillo from La Rioja who led the dissidents (Federalists) who opposed Rivadavia's centralization of government. When the assembly finally drafted a national constitution, the major portion of the country rejected it.

Constituent assembly of 1824

War against Brazil had begun in 1825. While the Argentine forces were able to defeat the Brazilians on the pampas of Uruguay, the Brazilian navy blockaded the Río de la Plata and succeeded in crippling Argentine commerce. Unable to end the war on favourable terms, Rivadavia resigned in July 1827, and the national government dissolved. Leadership of the Province of Buenos

Aires was given to a Federalist, Col. Manuel Dorrego. Behind Dorrego there were local interest groups whose political spokesman was the great landowner Juan Manuel de Rosas, who had been named commander of the rural militia. Dorrego was prevailed upon to make peace with Brazil. In 1828 the disputed eastern province was constituted as the independent state of Uruguay; the territory that Rivadavia had considered indispensable to the "national integrity" of Argentina was never to be recovered.

The armed forces, which had been reorganized for the war with Brazil, reacted adversely to these developments. They also found distasteful the prospect of losing power in a country that had dismantled its central government and was preparing, with no great zeal, to replace it with a loose confederation of provinces. In December 1828, troops returning from the war overthrew Dorrego and installed Gen. Juan Lavalle in his place, and Dorrego was executed.

While there was little resistance to the new governor in the city of Buenos Aires, uprisings began quite promptly in the outlying areas of the province. In Santa Fe, a convention of representatives from the provinces dominated by the Federalists, under the leadership of Rosas, called upon the Governor of Santa Fe to take steps against the Lavalle regime. Lavalle's authority was, in fact, eroded by the frequent uprisings in the plains; he and his Unitarian advisers (who favoured national unity with Buenos Aires at the head) failed to agree among themselves. Lavalle finally was disposed to come to terms with Rosas. They agreed to hold elections in Buenos Aires for a new provincial legislature and to throw their mutual support behind a slate of candidates composed principally of opponents of unitarism. Lavalle's ministers successfully thwarted the intent of this agreement by backing a rival slate of Unitarian candidates who won in the city, while in the plains Rosas prevented the elections from taking place altogether. Under a new compromise, Rosas and Lavalle appointed a moderate Federalist as governor of Buenos Aires pending new provincial elections. But political tensions were too great for this attempt at reconciliation. Rosas reconvened the old legislature that Lavalle had disbanded when he came to power—a triumph for the most intransigent forces of federalism. The legislature unanimously elected Rosas governor on December 5, 1829.

CONFEDERATION UNDER ROSAS, 1829–52

The regime of Rosas in Buenos Aires enjoyed far broader support than any of its predecessors. Special interest groups, landholders, and export–import merchants (along with the British diplomatic contingent that was identified with these interests) all fell behind the new governor. Practically all the influential sectors in the province identified Rosas' triumph with their own best interests.

Hegemony of Buenos Aires

Domestic politics, 1829–35. The new governor, however, saw clearly the ambiguities and dangers of such widespread support. If the urban and rural popular sectors were counting so heavily on him, it was because a long process of political mobilization had made them less passive than their counterparts in the rest of Latin America. He sought a political formula that would enable him to harness their energies to his own purposes. He found this formula in factionalism. By imposing an orthodox federalism, the exact nature of which he alone would determine according to his own political objectives, Rosas would be assured of control over the masses and would be able to use this influence as a restraint on the economic and social elites, with whose interests Rosas fundamentally identified. This blueprint assigned to Rosas the role of permanent arbiter of a delicate and constantly threatened balance; it reflected rather well the peculiar political experiences through which the Province of Buenos Aires as a whole had passed. In the other provinces where the political mobilization had been only partial and superficial, federalism was just another name for the return of Buenos Aires to its dominant position. The only instrument by which Rosas could impose some discipline on those provincial leaders accustomed to a far more intricate way of handling political matters was their fear of the superior power of the Province of Buenos Aires. This was a radical departure from the system of the 1820s, when Buenos Aires had endeavoured to control the other provinces by playing them off against each other. The experiences of 1824–27 had shown how difficult it was to prevent this from leading to a general conflagration. By comparison, Rosas' uncontested rule in Buenos Aires and the equally uncontested supremacy of Buenos Aires over the rest of the provinces lasted for more than a decade.

By 1832, the opposition to federalism had disappeared throughout the country, and Rosas turned over the reins of the government of Buenos Aires to his legal successor, Gen. Juan Ramón Balcarce. Balcarce's assumption of the office fanned sparks of dissidence among those who had pledged to uphold the principles of federalism. This unrest was particularly intense in the ranks of the regular army and threatened to destroy Rosas' continuing military and political prominence. The contest of strength between Rosas and Balcarce was unequal from the start, and Balcarce was overthrown. His successor took office with a Cabinet composed of Rosas' friends and adopted policies that were designed to lead to normality. But it was normality that Rosas feared more than anything, since it would have entailed the demobilization of his mass political following. When Quiroga was assassinated in 1835, having been sent to the interior as an emissary of Rosas to mediate an interprovincial dispute, his death was ascribed to antifederal forces. The legislature in Buenos Aires was induced to designate Rosas as governor of the province, but under conditions that Rosas himself successfully imposed: he was granted extraordinary resources, absolute public authority, and an extension of the governor's term of office from three to five years. The prosecution of Quiroga's assassins and their accomplices (among whom was the Governor of Córdoba) was carried out by Rosas under powers delegated by all the provinces under the Pacto Federal of 1831.

Foreign policies. Also in the field of international relations, Rosas' policies left no room for anything other than total success or total failure. International difficulties arose as extensions of domestic turmoil. At that time it was not clear to everyone that the neighbouring countries of Bolivia, Paraguay, and Uruguay were destined to be independent states and not part of the federal grouping controlled by Buenos Aires. Gen. Andrés Santa Cruz, who had established a confederation of Peru and Bolivia, supported opponents of Rosas in Argentina. Rosas in turn supported the influential governor of the northern Province of Tucumán, when that governor decided to go to war against the forces of General Santa Cruz. In 1839 the war ended in victory for the northern Argentine forces, in part thanks to an alliance with Chile. But the victory was not without its cost: the north had suffered economic hardships, and Rosas was singled out (mistakenly, in fact) as bearing the primary responsibility for having initiated the war.

Wars with Argentina's neighbours

His involvement in civil struggles in Uruguay proved even more costly. It led to the first open friction with France, which sent warships to blockade Buenos Aires in 1838. This created dissension in the coastal region, which had always depended heavily on export trade. Argentine political exiles in Montevideo received French backing in their efforts to overthrow Rosas, and in the north a league of dissident provinces was formed.

This formidable coalition of adversaries soon fell apart. Faced with other problems, France abandoned its adventure in the Río de la Plata area and left its local allies to fend for themselves against Rosas. At the same time, an army organized by Buenos Aires and commanded by the Uruguayan leader Manuel Oribe, gained control of most of the Argentine interior. For the first time since 1820, troops from Buenos Aires had advanced as far as the Bolivian and Chilean frontiers. The hegemony of Buenos Aires under Rosas' system of federalism was not to be challenged again. Oribe went

on to conquer most of Uruguay, and his predominantly Argentine army began the siege of Montevideo in February 1843. The city was saved only through the intervention of British warships, and in 1845 an Anglo-French fleet blockaded Buenos Aires, and a British fleet sailed up the Río Paraná. Eventually the British and French withdrew their aid to Montevideo and made peace with Rosas.

Having brought a halt to European interventions on two occasions, Rosas' power came to be respected throughout Europe and America. But he showed considerable restraint in the way he used his success vis-à-vis the great European powers, perceiving that in order to prevent further intervention he would have to respect the rights of foreign citizens and avoid threatening the positions gained by foreigners in the nation's economy. His attitude toward his local enemies was less moderate, though political terror diminished after the massacres of October 1840 and April 1842. Even in this moment of political triumph, however, the forces that were to overthrow him were coalescing.

That fact that Rosas was able to conduct a vigorous foreign policy for so many years was partly because of the weakness of Brazil, Argentina's natural rival in the Río de la Plata area, which had been involved in a civil war (1835–45) in Río Grande do Sul. Once the rebellion had been put down, it was only a question of time until Brazil's influence was again a factor in the Río de la Plata region. This influence was destined to be exerted against Rosas. It worked in support of a rebellion by Gen. Justo José de Urquiza, governor of the Province of Entre Ríos, who had been Rosas' mainstay in the wars in the eastern provinces. On May 1, 1851, Urquiza proclaimed his revolt against a "despot" who had violated a long-standing promise to establish constitutional government in Argentina. He then formed an alliance with Brazil and Uruguay. The allies first forced Rosas' troops to abandon the siege of Montevideo and then defeated his main army in the Battle of Caseros (February 3, 1852), just outside Buenos Aires. Rosas, abandoned by most of his troops as well as his political supporters, sought protection from the British. He was taken to England and retired to a farm near Southampton. He died there in 1877.

Economic development, 1820–50. In the 30 years after 1820, Argentina's society and economy underwent considerable changes. Buenos Aires was the province best adapted to the new era of free trade based upon the export of cattle products in return for consumer goods from overseas. The interior provinces adjusted slowly, replacing their traditional markets in highland Peru with new ones in Chile, where a great expansion of the mining industry was taking place. The coastal provinces fared better, although their livestock industry suffered from the effects of the civil war. For Santa Fe, the decade of the 1830s brought a return to moderate prosperity, and a similar trend began in Entre Ríos and Corrientes provinces in the 1840s.

Growth of a cattle economy

The cycle of wars that had accompanied Rosas' rise to power could have only a negative effect on this slow return to stability, but the direction of the country's recovery proved better than might have been hoped for. The establishment of a paper currency gave the government of Buenos Aires independence from local money lenders (*agiotistas*), while abuse of this power was avoided through the financial discipline that Rosas had imposed. And the achievement of an orderly domestic society through the slow conquest of the entire country by Rosas' armies had favourable effects on commercial activity.

These advances did not satisfy Rosas' political adversaries, above all the younger generation, which, from 1837 on, had opposed both unitarism and federalism. To them the chief sin of the Rosas regime was not its authoritarianism but its lack of a long-range program. They wanted a policy of national development based on European investment and immigration. But in the turbulent period between 1825 and 1850, no government would have been able to attract capital investment to Latin America. Moreover, the opponents of Rosas exaggerated the stagnation under his regime. The circumstances of his fall showed other limitations of his political system and the weakness of the unity that had been achieved. The supremacy of Buenos Aires was lost once more, as in 1820; once more the organizing of the country would have to wait upon the settlement of the power struggle in the country's most important province. The fall of Rosas marked the beginning of a new cycle of warfare.

NATIONAL CONSOLIDATION, 1852–80

General Urquiza, who was to help in Rosas' overthrow, had been one of his staunchest supporters, and the governors of formerly rosista provinces in the interior hastened to throw their support behind him. In many of these provinces, old or new opposition groups had planned to use Rosas' downfall as the signal to overthrow their own local governments, and those groups were not satisfied to let things continue as before. An opposing faction began to take shape. The Province of Buenos Aires was hostile to Urquiza and made league with his opponents in other provinces. They failed to obtain much public support, however.

Urquiza called a constitutional convention that met in Santa Fe in 1852. Buenos Aires refused to participate, but the convention adopted a constitution for the whole country that went into effect on May 25, 1853. Buenos Aires refused to join the new confederation, the first elected president of which was General Urquiza and the first capital of which was Paraná. The dissidence was a serious financial handicap to the state, since Buenos Aires kept for itself all the revenues from customs duties on imports. In 1859 Urquiza incorporated Buenos Aires by armed force, but he had to agree to a constitutional revision that underscored the federal character of the government.

Constitutional convention of 1852

Before the unification took effect, however, Urquiza was succeeded in the presidency by Santiago Derqui. Another civil war broke out; this time Buenos Aires won, defeating the federal forces led by Urquiza. The latter came to an understanding with Gen. Bartolomé Mitre, governor of Buenos Aires. They agreed that Mitre would lead the country but that Urquiza would exercise authority over the provinces of Entre Ríos and Corrientes. Derqui resigned, and Mitre was elected president in 1862. Buenos Aires became the seat of government.

This political accommodation was followed by the overthrow of local governments in the provinces of the interior. At the same time, however, the authority of the new president (1862–68) was progressively weakened by opposition within his own province of Buenos Aires, arising over what were thought to be Mitre's centralist leanings and his supposed forbearance toward Urquiza and the vanquished Federalists. The pressures of this opposition forced Mitre definitely toward intervention in the political struggles of Uruguay, where the Blanco Party, the ally of the Federalists, was in power, and thence into a war with Paraguay. From 1865 to 1870, a triple alliance of Argentina, Brazil, and Uruguay carried on a war to the death against Paraguay, employing modern weapons and tens of thousands of troops.

The war with Paraguay did not disrupt Argentina's commerce, as other wars had. In the 1860s and 1870s foreign capital and waves of European immigrants poured into the country. Railroads were built; alfalfa, barbed wire, better breeds of cattle and sheep, and finally the refrigeration of meat were introduced. The war itself helped to consolidate a national armed force, which was to become one of the cornerstones of the new centralized state.

But if the national state seemed to be taking shape with surprising vigour, the process never resulted in consolidation of the hegemony of the so-called liberal faction that had backed Mitre in 1861–62. The splits in the liberal ranks seemed irreparable. The opponents of Mitre in Buenos Aires and Córdoba welcomed the aid of old Federalist supporters of Rosas, and in the provinces there were open signs of a new Federalist upsurge. Moreover,

the army refused to uphold the polices of the President. One of Rosas' nephews rallied the support of the military behind the presidential candidacy of Domingo Faustino Sarmiento.

A native of San Juan, Sarmiento was an author and journalist, one of the intellectuals who had been sent into exile by Rosas. His emergence as an important political figure had been entirely unforeseen. Sarmiento's victory was guaranteed by the influence of the military combined with the support of the liberal faction in Buenos Aires that opposed Mitre. The new president (1868–74) held office without a political party of his own. Credit from abroad brought a continuing prosperity that allowed Sarmiento to engage in a costly civil war to put down an uprising in Entre Ríos that had begun with the assassination of Urquiza in 1870. At the same time the "schoolmaster president" was able to carry out an ambitious program for the development of public education as well as to continue the expansion of Argentina's economic capacities.

In 1874 Sarmiento was succeeded by his minister of justice, public education and worship, Nicolás Avellaneda (1874–80), a native of Tucumán; his victory was a new success of the reaction against the hegemony of Buenos Aires. The new president faced serious financial difficulties engendered by the European economic crisis of 1873. When the external sources of capital dried up, the country once again felt the consequences of Buenos Aires' financial superiority; through the power of its bank, the province was able to dictate its conditions to the national authorities. Avellaneda adopted a policy of austerity that required the cooperation of all political factions, and he undertook to reach an agreement with the declining forces of liberal nationalism led by Mitre. Not surprisingly he encountered strong opposition among the *autonomistas* in Buenos Aires, as well as from the provincial governments, which were to be replaced by new pro-Mitre administrations.

The next president was Gen. Julio Argentino Roca (1880–86), a native of Tucumán who was related by marriage to highly influential families in Córdoba. Roca had led a brilliant military career, during which he performed not only military but also political service for several successive presidents. In 1879 he had brought the long cycle of Indian wars to a victorious close. This achievement, opening up the pampas to settlement, made Roca a political hero. His campaign for the presidency provoked a new rebellion in Buenos Aires, but the uprising was quickly put down by the national army. The perennial question of the city's status was now settled by making it a federal territory and converting it into the national capital; a new capital for the Province of Buenos Aires was established at La Plata.

THE CONSERVATIVE REGIME, 1880–1916

The entire country was now dominated by the National Autonomist Party (Partido Autonomista Nacional), which was composed of an alliance of the various groups supporting Roca. These included many of the big ranchers, as well as commercial and business interests who were more than happy with Roca's formula of "peace and efficient administration." The new holders of power had technical and administrative rather than political goals. In that sense their aims were in accord with the national interest and commanded the spontaneous and unanimous backing of popular opinion. With such a broad base of support, the president was able to assert himself as never before. Roca's policies propelled the country onward in the direction it had taken since 1852 (and in certain respects even since 1810): economic development based on an export economy; modernization of the machinery of political administration, with increased professionalism in technical positions and in the military services; and the general secularization of public life. A period of prosperity contributed to the success of this program; exports, which had increased by about ten times between 1810 and 1850, multiplied another seven times between 1850 and 1880.

Economic expansion

Argentina's economic expansion in this period owed much to an inflow of British capital under government guarantees. This made possible the building of an extensive rail network linking the upriver provinces to Buenos Aires and the sea. Along with the growth of agriculture and ranching, there was also some expansion in other industries; but the impetus of the boom came from large-scale foreign investment that was channelled mainly into the transportation system.

This was also an era of rapid growth in population, largely from immigration. Argentina's population grew from fewer than 2,000,000 in 1869 to 3,900,000 in 1895 and to nearly 8,000,000 in 1914.

The crisis of 1890. The economic expansion led ultimately to inflation, accompanied by the issuance of too much paper currency and a crisis of confidence in the London capital market (known as the Baring crisis after the Baring Brothers, Argentina's financial agent in Britain). The financial crisis was followed by a political one. The government of Roca's successor, Miguel Juárez Celman (1886–90), had been wary to launch a necessarily unpopular anti-inflationary program. Discontent was growing, both within and outside the official party ranks. Brought together in active opposition to the government and participating in an effective program of agitation in the capital during all of 1890 were such groups as former president Mitre and his few followers, Catholics opposed to secularization of the state, and leaders of importance in the ruling Partido Autonomista Nacional, while General Roca followed a line of malevolent neutrality toward his too independent successor. In the spring of 1890 there were mass demonstrations in the streets of Buenos Aires. In July a revolt erupted that had strong support from within the army, but it was defeated by loyal elements. Even so, Juárez Celman was forced to step down in favour of the vice president, Carlos Pellegrini, a brilliant Buenos Aires politician and a solid ally of Roca.

The legacy of the twofold crisis was grave. A moratorium of payments had to be negotiated with foreign creditors, and Argentina had to stabilize its paper currency and adopt the gold standard. Economic expansion in the urban, commercial, and financial sectors came to a standstill, but the rural sector witnessed an impressive and rapid recovery. The export of refrigerated beef continued to expand, as did the cultivation of cereal grains. In Santa Fe, southern Córdoba, and western Buenos Aires, the cultivation of grain increased mostly through the efforts of immigrant tenant farmers and sharecroppers; by the beginning of the 20th century, exports of cereals equalled those of meat.

The rise of radicalism. The difficulties of the 1890s brought the rise of a new party, the Unión Cívica Radical, strongly opposed to the ruling regime and the compromise candidate, Luis Sáenz Peña, accepted by Mitre and the more moderate opponents of the Roca-Juárez Celman regime. As president (1892–95), Sáenz Peña attempted to bring the radical party into the political system. He failed and was replaced by the vice president, José Evaristo Uriburu, who was under the political guidance of the former president Roca. In 1898 Roca was able to return to the presidency for a second term (1898–1904), and he began an attempt to bring the more moderate radicals effectively back into the loose alliance of local political groups, which after 1890 controlled the national government. The most intransigent radical factions, headed by Hipólito Irigoyen, who was later to become president twice, remained in opposition.

Broadening of political life

While political opposition was on the decrease, social unrest was becoming more widespread. The Socialist Party was founded in 1896 by Juan B. Justo. The anarchist movement became highly influential in the labour unions. An even more serious threat to stability was the growing disarray within the government itself, which was aggravated by the continuing influence of Roca, whose principal tactic for conserving his own supremacy lay in playing off his followers against each other. Roca's break with Pellegrini was a final blow for the National Autonomist Party, and it was only through wasteful and con-

fused manoeuvres that Roca was able to avoid being succeeded in office by Pellegrini in 1904. The candidate he finally put into the presidency, Manuel Quintana (1904–06), was far from being one of Roca's staunchest supporters. Quintana was forced to quell a radical revolution in 1905, and his death opened the way to the presidency for the Cordoban José Figueroa Alcorta (1906–10), who turned immediately to the task of destroying Roca's political machine. In 1910, Alcorta installed as his successor Roque Sáenz Peña (1910–14), a brilliant politician who had entered the diplomatic service as a refuge from Roca's enmity, but who was fully prepared to construct a governing coalition on new foundations.

The course of Argentine politics in the final stages of Roca's career had convinced many of his most influential and militant followers that the country needed electoral reform. The loose coalition of conservative political cliques that governed the country was kept in office by patronage and electoral manipulation. Reform was not seen as excessively dangerous since the Radical opposition seemed to have limited support. In 1912, Sáenz Peña had the Congress pass an electoral-reform law that called for a compulsory, secret ballot for all male citizens. His death in 1914 deprived the national leadership of its guiding force, and the electoral law he had authored opened the gates of power to the radicals. The interim presidency of Victorino da la Plaza (1914–16) was followed by that of the Radical leader Hipólito Irigoyen (1916–22). He was the first Argentine president who owed his victory to the popular vote rather than to selection by the incumbent president among the members of a ruling oligarchy.

The democratization of the political system came during a long upswing in prosperity that was only momentarily interrupted by World War I. It benefitted principally foreign investors and the large landowners. At the same time, an urban and even a rural middle class were growing, comprised in part of prosperous renters of productive land. A significant industrial expansion began at the turn of the century in food processing, textiles, and light metallurgy. The expansion of the economy depended, as before, upon export trade: between 1880 and 1913, the value of exports increased another seven times.

THE RADICAL REGIME, 1916–30

The Radical front was a coalition of heterogeneous social groups, which made it difficult to put through economic and social reforms without upsetting the front's equilibrium. Not surprisingly, Irigoyen preferred to concentrate on curing the political ills he had inherited from the conservative regime. The most urgent measure had to do with political patronage, which had been used by the conservatives to keep their candidates in office within a political milieu in which the monopoly of power by provincial political elites was taken for granted. Patronage had to be not only put to the service of the new party in power but also adapted to the new conditions implicit in the expansion of the electoral system. If as a result the bureaucracy became more expensive and not necessarily more efficient, the political efficacy of this was not to be denied: the Radical machine made itself virtually unbeatable at the polls in almost every province.

Political patronage and university reform

In other fields the Radical administration also showed concern for expanding its political base. Irigoyen achieved substantial rapport with the more moderate labour unions—a rapport expressed in a generally prolabour policy. That policy had to be tempered after violent clashes occurred in the capital city during the general strike of January 1919, alarming conservative opinion and bringing the intervention of the army. Henceforth the President was obliged to be more cautious in his dealings with organized labour, because the military had aligned itself with conservative interest groups. His administration supported organizations and movements among tenant ranchers in the grain-growing regions and also put through a university-reform plan that the students had advocated and granted them partial control of the institution, thus weakening considerably the President's conservative foes in the university.

Irigoyen's influence was a deciding factor in the election of his successor, Marcelo T. de Alvear (1922–28). The son of an aristocratic family, with a record of staunch, but passive, loyalty to radicalism, Alvear represented a safe choice. But he was not content with the restrictions imposed upon him by his mentor Irigoyen. He became the reluctant leader of a conservative wing, the Antipersonalists, hostile to Irigoyen. In the elections of 1928, Irigoyen ran for a second term and was elected by a margin of two to one in the popular vote. This was a great personal triumph and made him the master of his party.

The Radical political machine, many of whose representatives had scarcely been heard of outside their own districts, won control of the legislature in the election. If Irigoyen was scarcely a revolutionist, his victory over the economic, social, and political elites of the country had nonetheless earned him their strong enmity. His political machine, while an excellent mechanism for securing power, proved to be incapable of governing; its inadequacies were starkly revealed in the crisis of 1929, precipitated by the army, which expelled Irigoyen from office on September 6, 1930. This marked the end of the constitutional continuity that had lasted for 68 years and also the end of the long period of economic expansion based on the export of raw materials.

Between 1913 and 1928, Argentina's export figures had doubled again. Behind the continuing upward trend lay a new development—a shift in economic power to foreign merchants and processors at the expense of the Argentine landowning class. Whereas, before 1914, these foreign interests had been concentrated mainly in the grain-growing sector, after 1920 they moved into the cattle-raising industry. Another development was linked with the commercial and financial relationship with Great Britain and the United States. Private investment still came mainly from Great Britain, which was also the main market for Argentine exports. The United States provided industrial and transportation equipment and was the government's main source of credit, but it had erected tariff and other barriers to the importation of Argentine goods (including sanitary restrictions on imported meat). Irigoyen had adopted a strongly anti-North American (and pro-British) line. The growing economic weakness of Great Britain made the situation dangerous, and the economic crisis of the 1930s forced drastic readjustments.

THE CONSERVATIVE RESTORATION, 1930–43

The military coup that expelled Irigoyen installed Gen. José Félix Uriburu in the presidency (1930–32). A descendent of an old northern family with more than a century of involvement in Argentine politics, Uriburu, along with his basic conservatism, showed a leaning toward the Fascist ideas of the 1930s. His influence within the ranks of the army, however, was not as great as that of Gen. Augustín Pedro Justo, a former minister of war under Alvear who favoured a conservative reorientation of the country within constitutional limitations. Events soon obliged Uriburu to accept the restoration of representative institutions, provided the Radical faction could be kept out of power. The Radicals had been reorganized under the leadership of Alvear and had won an unexpected victory in trial elections held in the Province of Buenos Aires in April 1931. The problem of keeping the Radicals out was solved by imposing such stringent political restrictions on them (and persecuting them as well) that they finally boycotted the national election of 1931. General Justo, with the backing of a coalition of conservatives, anti-Alvear Radicals, and independent Socialists, and with only the limited use of electoral fraud, was elected by a large majority.

The new president (1932–38) faced a difficult economic situation. The Roca-Runciman Agreement with Great Britain (1933) guaranteed Argentina a fixed share in the British market for meat and ruled out tariffs on British cereal imports. In return, Argentina agreed to certain stipulations with regard to trade, currency exchange, and the preservation of Britain's commercial interests in the

The Great Depression

country. The treaty sharpened doubts of Argentines concerning the economic relationship between Argentina and Britain. Other unpopular reforms included a restructuring of the monetary system and the establishment of agencies to control exports. After 1935 the economic climate improved. Limitations on imports of manufactured goods resulted in a new industrial expansion, growth in the urban population, and increases in the membership of labour unions.

Radical movement gains

In 1934 the Radical movement returned to the electoral battles. Its victory in the provincial elections in Córdoba heightened political tensions, and electoral violence increased. As conservative hegemony was threatened, the temptation to abandon constitutional government drew strength from the belief that a dictatorial regime would be less constrained by the interests that earlier had imposed unpopular domestic policies and brought about dependence on Britain in external affairs. A series of political and financial scandals eroded the prestige of the semi-constitutional regime.

The election of 1937, in which the government retained its power, was marked by fraud and violence. The new president, Roberto M. Ortiz (1938–40), returned to more proper electoral procedures, calling for federal intervention in the province of Buenos Aires, where a corrupt conservative machine had been in control; this intervention made it possible for the Radicals to win. Attempts by Ortiz' enemies to implicate him in financial scandals failed, but poor health obliged him to resign. His successor, Ramón S. Castillo (1940–43), restored the conservative coalition to power, gaining the military and political support of General Justo.

At the outbreak of World War II, Argentina had declared its neutrality. After the United States entered the conflict in 1941, Castillo's government alarmed Washington by remaining neutral. This issue united all the opposition groups—from the small Communist Party to the conservatives who disliked Castillo's authoritarian style. To disarm his opponents, Castillo imposed a state of siege and relied increasingly on the support of military officers, who gradually took control. In January 1943 a stabilizing influence disappeared with the death of General Justo. President Castillo fell in June 1943 at the hands of officers led by his own minister of war, Gen. Pedro P. Ramírez.

THE PERÓN ERA, 1943–55

Transition period. The military government faced several urgent and difficult problems. Neutrality was advocated by some, and U.S.–Argentine tensions over the emergence of Brazil as the privileged ally of the U.S. in South America prevented a reversal of this policy. A choice between the restoration of a representative system and the installation of a long-term military dictatorship was also required. Gen. Arturo Rawson, who had been made president, was forced to resign after two days when his desire for a radical break with the conservative past and his advocacy of the emerging United Nations won no military support.

Rawson, Ramírez, and Farrell

General Ramírez replaced Rawson as president (1943–44); neutrality was maintained, and the opposition from all organized political groups (except the nationalist right wing and the Fascist sympathizers) increased. Press censorship, the dissolution of political parties, and the introduction of religious instruction in the public schools were manifestations of an emergent authoritarianism. Under pressure from the U.S., the regime broke off diplomatic relations with Germany. This deed was not favoured by the military officers, and Ramírez turned the presidency over to Gen. Edelmiro J. Farrell (1944–46). Argentina found itself in virtual international quarantine, and to avert ruin it had to face the return to representative democracy.

The search for a solution ended in the rise of Col. Juan Perón to the office of president. Since October 1943, when he had gotten the minor job of running the labour department, Perón had been building a political empire based in the labour unions. He helped the unions win favourable settlements from employers; he pushed through a welfare program that provided vacations, retirement benefits, and severance pay. By 1945 he was also vice president and minister of war, and he sought to orient government policy with national and international conditions. His changes included giving autonomy to universities, reconstructing political parties (including the Communist Party, prohibited since 1936), and declaring war on Germany, thereby making it possible for Argentina to enter the United Nations. But with the return of political freedom came renewed opposition, culminating in a mass demonstration in Buenos Aires in September 1945. Emergency measures were enacted, Perón's military support gave way, and on October 9 he was removed from office and arrested. Neither the political nor the military opposition, however, could agree on what to do. Perón's adherents in the unions (Confederación General del Trabajo; CGT) organized a strike that found enthusiastic support among the people. He was released on October 17, and his foes were forced to resign.

Perón in power. Perón now campaigned for the presidency in the elections of 1946. He organized a Labour Party that drew members from the unions and dissidents from traditional parties. He was resisted by all the old parties and by the major vested-interest groups. His victory, though narrow, gave him control of both houses of Congress and all the provincial governorships.

In power, the Peronista movement consolidated its majority, increasing it to nearly 65 percent in the election of 1951. Perón's political strategy and tactics were authoritarian and personalistic. Education and the courts were politically purified; a state of internal war was declared, to allow an expansion of executive authority; revenues were redistributed in favour of the workers; public services were nationalized; and urban and industrial areas were given preferential treatment over their rural counterparts.

Until 1949 Perón's economic policies were successful, largely because of the prosperity of exporters during and just after the war. As inflation increased and terms of trade worsened, however, it became difficult to finance imports of vital raw materials. The constitutional reform of 1949 allowed Perón to be re-elected for the term 1952–58; his government took on a more conservative hue in such matters as providing credit to rural interests and imposing wage controls. This new conservatism was hastened by the death of the President's wife, Eva Perón, in July 1952. She had been a powerful political figure in her own right, burnishing the regime's image of popular democracy. After 1952 Perón incurred the increasing hostility of the church and the students; economic difficulties led him to turn to foreign industrial firms for capital and technical help. His efforts to eliminate the influence of the church provoked disaffection in the officer corps, and in September 1955 he was overthrown in an uprising led by Gen. Eduardo Lonardi and supported by part of the Córdoba garrison.

Influence of Eva Perón

ATTEMPTS TO RESTORE CONSTITUTIONALISM, 1955–66

The leaders of the coup were agreed on the need for radical changes in economic policies—in essence, a program of austerity for the country as a whole and particularly for the urban centres. It was soon evident, however, that concessions to the latter were necessary. The new policies, followed with only brief interruptions until 1966, stressed exports and a quest for foreign credit and investment, but decreased the income of wage earners.

General Lonardi, acting as provisional president, aligned himself with an authoritarian Catholic group that supported accommodation with Perón's former followers, but he soon gave way to Gen. Pedro Eugenio Aramburu, who became president in November 1955. The new administration (1955–58) was a military dictatorship seeking to restore constitutional government. Taking a fiercely anti-Peronista stance, it dissolved Perón's old party and placed the CGT under state administration. A Peronista military uprising in June 1956 was easily crushed. The Peronistas searched for ways to infiltrate the new political structure, and in the election of 1957 one-fourth of the voters heeded their call to cast blank ballots. The Peron-

istas wielded considerable influence on the factions that were competing for power and in 1958 supported Arturo Frondizi, a radical leader who promised to readmit them to political life in return for their support. Frondizi won the presidency and majorities in both houses of Congress.

The new president (1958–62), known for his commitment to economic nationalism, showed a keen interest in reviving the flow of foreign investment. A currency devaluation that favoured exporters and foreign investors had adverse effects on the middle and lower classes. The campaign against inflation brought restrictions on credit, increasing the difficulties of the industrial sector, already faced with shrinking markets. Frondizi found himself dependent on military support for the pursuit of his unpopular policies.

In March 1962 the reorganized Peronistas gained control of important districts, among them the province of Buenos Aires. The armed forces withdrew support from Frondizi, dissolved Congress, and set up a government in the name of José María Guido, president pro tempore of the Senate, whose constitutional claim on the presidency was doubtful. Guido's 18-month administration was one of confusion as two military factions fought for control. The Colorados (Reds) sought a dictatorship that would deal strongly with the Peronistas and extreme leftists, who were regarded, after the success of Fidel Castro in Cuba, with alarm. The Azules (Blues), who prevailed, favoured a new constitutional government by a coalition including the Peronistas, though the latter were to be confined to a role weaker than that indicated by their strength among the voters.

Arturo Illia

The elections of July 1963 resulted in an unexpected victory for Arturo Illia, the candidate of the People's Radicals. President Illia (1963–66) inherited Frondizi's economic problems, although the drastic reorientation of the economy had begun to show signs of success, improving the balance of trade. Illia tried without success to split the resurgent Peronistas, who were now in control of the labour unions, from their exiled leader. Antagonized, the Peronistas supported a coup in June 1966 that brought to power Gen. Juan Carlos Onganía, a former leader of the Azules and commander in chief of the army.

GOVERNMENT BY THE ARMED FORCES

Generals Ongania and Levingston. It took the new government six months to find an economic policy. Adalbert Krieger Vasena was named minister of economy and labour at the end of 1966, and he strove for economic stability through yet another currency devaluation —but one that would benefit the state rather than the exporters of primary commodities. He also undertook programs in electric power, steel, roads, and housing.

On the political front, after some initial repression—the purging of the Supreme Court and intervention in the universities—the regime came to show a certain toleration of opposition groups. In May 1969 disturbances and riots in the cities of Corrientes, Rosario, and particularly Córdoba rose out of student and labour conflicts; these incidents, later known as the Cordobazo, were seen as indicating resentment at Krieger Vasena's authoritarian management of the economy. Krieger Vasena was removed, but the Onganía administration was unable to agree on an alternative economic policy, and the Cordobazo decisively affected the political climate. Underground activities were organized by a Trotskyite group, the Ejército Revolucionario del Pueblo (People's Revolutionary Army; ERP), and by Peronista groups, which typically recruited their followers among former Catholic rightists who were undergoing radicalization to the left. In 1970 one of these Peronista organizations, the Montoneros, besides decimating the moderate Peronista union leadership, captured and killed former president Aramburu, who, with the aid of moderate Peronistas, was organizing a movement for a return to constitutional rule. The armed forces, disturbed by the assassination, overthrew the Onganía government in June 1970. Gen. Roberto Marcelo Levingston, who replaced Onganía, tried to generate political support by adopting a moderately nationalistic and populist position. The return of inflation and his political ambitions caused his overthrow; he was replaced in March 1971 by Gen. Alejandro Agustín Lanusse.

General Lanusse. The third military government since 1966 was committed to the restoration of representative democracy. Although the political parties had agreed on that policy, they were wary of identifying with an unpopular military dictatorship. Preoccupied with the political problem, the Lanusse government gave free rein to inflation, which was advantageous to the rural exporters and satisfactory to unionized labour, whose wages rose. Resentment arose, however, among the middle classes, from which recruits for the underground organizations were drawn.

Perón's new bid for power

While supporting the Peronista underground and adopting some of its socialist slogans, Perón also used other means in a new bid for power. He maintained a formal alliance with the Frondizi followers and other minor political groups, but the cornerstone of his strategy was an understanding with the largest non-Peronista party, the Radicals. Mindful of the influence of vested interests, he purged his economic proposals of any motives that could alarm the propertied classes.

The military government could not rival this virtuoso performance; it maintained the legal restrictions that prevented Perón's own candidacy but was forced to witness his triumphal visit to Argentina in November 1972 and the electoral victory of the Peronista coalition (Frente Justicialista de Liberación) in March 1973.

THE RETURN OF PERONISM

The newly elected president, Héctor J. Cámpora, a lacklustre Peronista, took office in May 1973. Under his administration the Peronista left wing, without dissolving its underground organizations, consolidated its hold on the Peronista youth and acquired considerable influence. The final return of Perón in June was the occasion of a battle between right and left. The union leadership and Perón's private secretary, José López Rega (appointed minister of social welfare by Cámpora), launched a violent anti-leftist campaign that had the discreet support of Perón himself. In July Cámpora resigned, and new elections were called. An interim president, Raúl Lastiri, the son-in-law of López Rega, began a purge of leftist influences in the government.

Perón's second presidency. Perón was elected president on a ticket that included his wife, María Estela (called Isabel) Martínez de Perón, as vice president and took office in October 1973. He continued the campaign against the left, which lost its last strongholds in some provincial administrations and universities, and in May 1974 the victims of the purge acknowledged the break with their leader and passed into (still legal) opposition. Underground activities, somewhat muted after May 1973, again became evident, and a mysterious right-wing organization, the Alianza Anticomunista Argentina (Argentine Anti-Communist Alliance; AAA), suspected by many to be close to the police and intelligence branches of the administration, began to take a heavy toll of political, student, and union leaders.

Peronista break with Perón

Perón tried to limit the conflict to the Peronista movement; he maintained his understanding with the non-Peronista parties and publicly sought advice from the leader of the Radicals, Ricardo Balbín. His economic policies after May 1973 combined monetary stabilization and rigid control of prices and wages with a gradual redistribution of income in favour of wage earners and a limitation on profits of the agrarian exporters. A rapid expansion of government expenditures, financed through new currency emission, put great strain on the controlled prices, and rural producers reacted by limiting production (especially of grain). By 1974 the indirect consequences of the world petroleum crisis were felt; Europe could not afford to maintain its substantial imports of Argentine meat, and after more than 10 years of expansion of exports the balance of payments again imposed constraints.

Perón's legacy. Perón expended his last energies in defense of his economic policies against the growing impatience of labour. When he died on July 1, 1974, he left

to his widow and successor a deeply compromised inheritance. The transition of power was smooth, however, and Mrs. Perón, in close alliance with López Rega, became even more inflexibly oriented toward the right. Violence reached new heights. López Rega, who used the rightist crusade to consolidate his power base, favoured labour and army leaders who were committed to support him. This course created hostility among union, political, and military leaders. In the autumn of 1975 a drastic devaluation and a steep drop in real wages were proposed. The union leadership renewed its hold on its followers, the armed forces refused to repress legitimate popular protest, and Mrs. Perón was persuaded to dismiss López Rega and his friends. The unrest deepened, however, and in December there was an uprising by the air force. On March 24, 1976, military officers deposed Mrs. Perón and took over the government.

BIBLIOGRAPHY. The best general history of Argentina continues to be that edited by the ACADEMIA NACIONAL DE LA HISTORIA, under the direction of RICARDO LEVENE, *Historia de la Nación Argentina*, 3rd ed. (1961–); another work published by the Academy is *Historia Argentina Contemporánea, 1862–1930*, 4 vol. (1963). Among numerous studies of specific problems are JOSE LUIS ROMERO, *Las ideas políticas en Argentina*, 3rd ed. (1959; Eng. trans., *A History of Argentine Political Thought*, 1963), a presentation of the broad lines of political development and thought; ALDO FERRER, *La economía argentina; las etapas de su desarrollo y problemas actuales* (1963; Eng. trans., *The Argentine Economy*, 1967); and H.S. FERNS, *Britain and Argentina in the Nineteenth Century* (1960). A detailed study of the evolution of Argentina in the first half of the 19th century is WOODBINE PARISH, *Buenos Ayres and the Provinces of the Rio de la Plata*, 2nd ed. (1852). Financial politics is treated in MIRON BURGIN, *The Economic Aspects of Argentine Federalism, 1820–1852* (1946). A synthesizing presentation of economic expansion after 1870, with attention to its social and political implications, is ROBERTO CORTÉS CONDE and EZEQUIEL GALLO, *La formación de la Argentina moderna* (1967). THOMAS F. MCGANN, in the first part of his *Argentina, the United States, and the Inter-American System, 1880–1914* (1957), has painted a fairly happy picture of the ruling class at the time of the culmination of the exporting, expansionistic era. A different view is that of RODOLFO PUIGGROS, *Historia crítica de los partidos políticos argentinos* (1956).

(T.H.-D.)

Ariosto, Ludovico

An important and influential poet of the 16th century, Ludovico Ariosto is the author of *Orlando furioso*, a poem generally regarded as the most perfect expression —both in substance and form—of the artistic tendencies and spiritual attitudes of the Italian Renaissance. He also contributed to the development of the drama: his plays, which are themselves minor works, were the first of those imitations of Latin comedy in the vernacular that have characterized European domestic comedy.

Ariosto was born on September 8, 1474, at Reggio Emilia, where his father, Count Niccolò, was commander of the citadel. When Ludovico was 10, the family moved to his father's native Ferrara, and the poet always considered himself a Ferrarese. He showed an inclination toward poetry from an early age, but his father intended him for a legal career, and so he studied law, unwillingly, at Ferrara from 1489 to 1494. Afterward he devoted himself to literary studies until 1499. Count Niccolò died in 1500, and Ludovico, as the eldest son, had to give up his dream of a peaceful life devoted to humanistic studies in order to provide for his four brothers and five sisters. In 1502 he became commander of the citadel of Canossa and in 1503 entered the service of Cardinal Ippolito d'Este, son of Duke Ercole I.

Ariosto and the Este family

His duties as a courtier were sharply at odds with his own simple tastes. He was expected to be in constant attendance on the Cardinal and to accompany him on dangerous expeditions as well as travel on diplomatic missions. In 1509 he followed the Cardinal in Ferrara's campaign against Venice. In 1512 he went to Rome with the Cardinal's brother Alfonso, who had succeeded Ercole as duke in 1505 and had sided with France in the Holy League war in an attempt to placate Pope Julius II.



Ariosto, woodcut after a drawing by Titian from the third edition of *Orlando furioso*, 1532.
By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.

In this they were totally unsuccessful and were forced to flee over the Apennines to avoid the Pope's wrath. In the following year, after the election of Leo X, hoping to find a situation that would allow him more time to pursue his literary ambitions, Ariosto again went to the Roman court. But his journey was in vain, and he returned to the Cardinal in Ferrara.

So far he had produced a number of Latin verses inspired by the Roman poets Tibullus and Horace. They do not compare in technical skill with those by Pietro Bembo, a contemporary poet and outstanding scholar, but they are much more genuine in feeling. Since about 1505, however, Ariosto had been working on *Orlando furioso*, and, indeed, he continued to revise and refine it for the rest of his life. The first edition was published in Venice in 1516 at the Cardinal's expense. This version and the second (Ferrara, 1521) consisted of 40 cantos written in the metrical form of the ottava rima (an eight-line stanza, keeping to a tradition that had been followed since Boccaccio in the 14th century through such 15th-century poets as Politian and Matteo Maria Boiardo). The second edition shows signs of Bembo's influence in matters of language and style that is still more evident in the third edition.

Orlando furioso

Orlando furioso is an original continuation of Boiardo's *Orlando innamorato*. Its hero is Orlando, the Italian form of the French Roland, hero in the so-called French Charlemagne epics. In the poem there are three principal nuclei around which a variety of episodes and themes, mostly taken from the literature of the Middle Ages, is grouped: Orlando's passion for Angelica, the war between Christians and pagans near Paris, and the secondary love story of Ruggiero and Bradamante. The first is the most important, particularly in the first part of the poem; the second represents the epic background of the whole narrative; the third is introduced partly as a literary courtesy to the Este family, who are supposed to owe their origin to the union of the two lovers.

In 1517, one year after the first publication of *Orlando furioso*, the Cardinal was created bishop of Buda. Ariosto refused to follow him to Hungary, however, and in the following year he entered the personal service of Duke Alfonso, the Cardinal's brother. He was thus able

to remain in Ferrara near his mistress, Alessandra Benucci, whom he had met in 1513. But, in 1522, financial necessity compelled him to accept the post of governor of the Garfagnana, a province in the wildest part of the Apennines. It was torn by rival political factions and overrun by brigands, but Ariosto showed great administrative ability in maintaining order there.

During this period, from 1517 to 1525, he composed his seven satires (titled *Satire*), modelled after the *Sermones* (satires) of Horace. The first (written in 1517 when he had refused to follow the Cardinal to Buda) is a noble assertion of the dignity and independence of the writer; the second criticizes ecclesiastical corruption; the third moralizes on the need to refrain from ambition; the fourth deals with marriage; the fifth and sixth describe his personal feelings at being kept away from his family by his masters' selfishness; the seventh (addressed to Pietro Bembo) points out the vices of humanists and reveals his sorrow at not having been allowed to complete his literary education in his youth.

Retirement and the final edition of *Furioso*

By 1525 Ariosto had managed to save enough money to return to Ferrara, where he bought a little house with a garden. Probably between 1528–30 he married Alessandra Benucci (though secretly, so as not to forego certain ecclesiastical benefices to which he was entitled). Thus his dreams of a peaceful life devoted to study and poetry came true, and he spent the last years of his life with his wife, his crippled brother Gabriele, and his son Virginio, cultivating his garden and revising the *Orlando furioso*. The 3rd edition of his masterpiece (Ferrara, 1532) contained 46 cantos (a *giunta*, or appendix, known as the *Cinque canti*, or "Five Cantos," was published posthumously in 1545). In this final version the poet's style and poetic personality, which confers on his characters a refined spirituality, at last achieved its perfection. It was published a few months before his death on July 6, 1533.

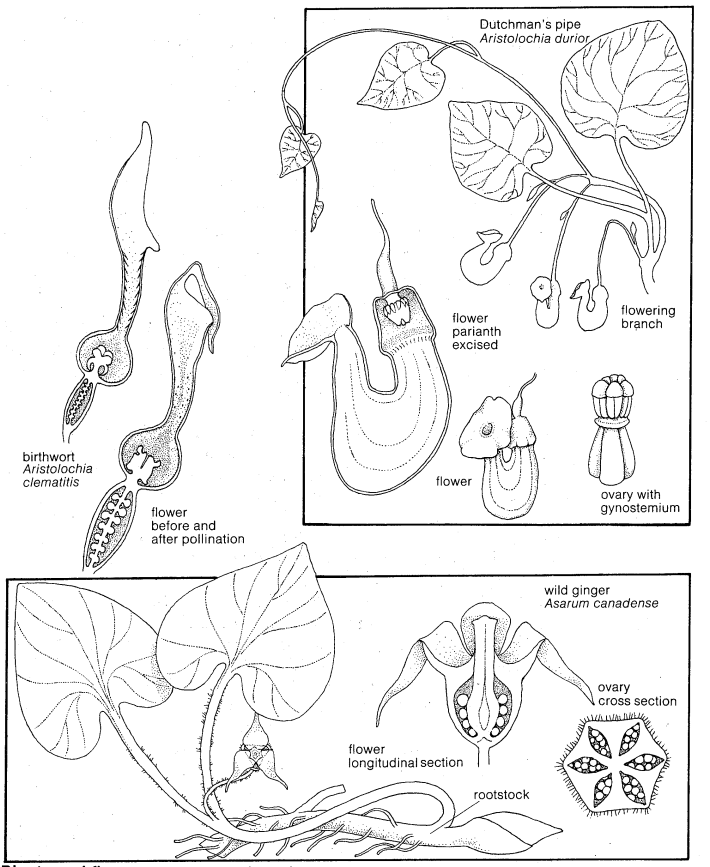
MAJOR WORKS

POETRY: *Orlando furioso* (1st ed., 1516; 2nd rev. ed. 1521; 3rd rev. ed., 1532); *Satire* (1517–25).

PLAYS: *Cassaria* (1508); *I suppositi* (1509); *Il negromante* (1520); *La lena* (1529); *I studenti* (completed by Gabriele Ariosto, published posthumously as *La scholastica*).

BIBLIOGRAPHY. Modern editions of Ariosto's works are: LANFRANCO CARETTI, *Orlando furioso* (1954); CESARE SERGE, *Orlando furioso* (1964), a critical edition; and *Opere Minori* (1954); ANTONIO CAPPELLI, *Lettere*, 3rd ed. (1887); ANGELO STELLA, *Lettere* (1965); A. SALZA, *Gli Studenti, con le continuazioni di Gabriele e Virginio Ariosto* (1915). See also GIUSEPPI AGNELLI and GIUSEPPI RAVEGNANI, *Annali delle editione ariostee* (1933), a detailed catalog of editions of Ariosto's work; GIUESEPPE FATINI, *Bibliografia della critica ariostea, 1510–1956* (1958), a fundamental bibliographical list; PIO RAJNA, *Le Fonti dell'Orlando furioso*, 2nd ed. (1900), on the sources of the poem; MICHELE CATALANO, *Vita di Ludovico Ariosto riconstruita su nuovi documenti*, 2 vol. (1930–31), the standard biography of the poet; BENEDETTO CROCE, *Ariosto, Shakespeare, e Corneille* (1920; Eng. trans., 1966), and ATTILIO MOMIGLIANO, *Saggio su l' "Orlando furioso,"* 5th ed. (1959), both fundamental for an aesthetical appreciation of the poem.

(G.A.)



Plants and flowers representative of the family Aristolochiaceae.

Drawing by M. Moran based on (wild ginger) N.L. Britton and A. Brown, *Illustrated Flora of the Northern States and Canada*, vol. 13 (1913); New York Botanical Garden, (birthwort, Dutchman's pipe, ovary with gynostemium) after Baillon, Eichler, Oliver, Sachs, Solereder, Bot. Magazine in A. Engler, *Syllabus der Pflanzenfamilien* (1964); Gebruder Borntraeger, (Dutchman's pipe, except ovary with gynostemium) reprinted with permission of The Macmillan Company from *Taxonomy of Vascular Plants* by G.H.M. Lawrence. Copyright 1951 by The Macmillan Company

Aristolochiales

Aristolochiales is a small order of flowering plants, comprising one family, the Aristolochiaceae, which consists of ten genera and approximately 600 species. The order is not a particularly important one in terms of size or widespread occurrence, but it does represent an interesting morphological type, showing flowers highly adapted to rather unusual means of insect pollination (mainly by carrion flies and similar organisms). A further feature of interest is the basically trimerous (*i.e.*, the flower parts are present in the number three or multiples thereof) structure of the flowers—this being a much more common feature in the class Monocotyledoneae (monocots) than in Dicotyledoneae (dicots), in which the Aristolochiales order is one of the few examples.

General features. Plants of the order are herbaceous, or more usually, woody, the greater proportion of the woody species being scramblers, climbers, or lianas. Their flowers are often of bizarre shapes and colours, and are sometimes very large; they frequently emit unpleasant smells (usually rather like rotting meat) and so may often be rather conspicuous. They are found most abundantly in the tropics, where many species form lianas in the forests, but a few species occur in the temperate regions (both north and south) of the Old and New Worlds. Very few of the species are of economic importance; a few are cultivated as curiosities in greenhouses in temperate areas, and several contain medicinally active compounds that have been used by various peoples in the treatment of snake and scorpion bites, as well as for less clearly specific purposes.

Natural history. *Life cycle.* Plants of the Aristolochiales order show few special adaptations to the environment, except for those connected with pollination. Many of the herbaceous species are rhizomatous (*i.e.*, they have an underground rootlike structure, the rhizome), and may spread vegetatively by this means; otherwise, vegetative reproduction is not of great importance. It must be stressed, however, that most of the available detailed knowledge of these plants refers to the temperate representatives, which form a very small part of the whole order. As more is learned about the life cycles of the tropical species, more information concerning special adaptations and reproduction may become available.

Pollination. The flower is strongly protogynous, meaning that the female structures (ovary, style, stigma) mature before the male ones (anthers) of the same flower, and the receptive stigmas completely cover the unopened anthers. Insects are attracted to the flowers by the smell and the mottled coloration of the perianth (petals and sepals) limb. They pass down the tubular part of the flower to the swollen base, which contains the gynostemium (a specialized structure, involving both male and female flower parts); their exit from the base is pre-

Role of pollinating insects

vented by the occurrence of stiff, downwardly pointing hairs that line the perianth tube. If pollination occurs (the insect having brought pollen from another flower), the stigmas become erect and expose the dehiscing (splitting open) anthers, thus covering the insect with a further supply of pollen. Also as a result of pollination, the stiff hairs in the perianth tube wither, allowing the insect to escape. As the ovules mature, the perianth also withers, and becomes closed.

Ecology. The Aristolochiales order consists mainly of plants of the tropical forests, where they occur as lianas, climbers, twiners, or as forest-floor herbs. The temperate representatives may also occur in forests but are also found as twiners and herbs in scrub and on drier hillsides and similar places. Little is known of the precise climatic conditions required by the various species, as they have never been widely cultivated for any purpose.

Form and function. *Vegetative characters.* Because these plants are either herbaceous or woody, the growth form varies from creeping, prostrate herbs to woody climbers, some of considerable size; these climbers frequently show anomalous forms of secondary thickening, and are usually characterized by the occurrence of broad medullary (internal wood) rays. The leaves are alternate, stalked, without stipules (small leaflike appendages at the base of the leaf stalk), and are usually simple and often cordate in shape. The tissues contain oil cells.

Flower and fruit characters. The flowers vary from small to very large, and may be solitary, or in inflorescences (clusters) of varying degrees of complexity; in some species the flowers are borne in clusters on the old wood (cauliflory). They are often strikingly, if bizarrely, shaped and coloured, dark reds, greens, and browns predominating, sometimes with leopard-like, yellow mottling. The coloured parts are formed (in all but one of the genera) by the calyx (sepals), which is tubular at the base. The tube may be more or less straight, or may be variously bent, a U-shape being frequent; it is generally swollen at the base. Toward the apex the tube divides into three sepals; these may be similar and actinomorphically (radially symmetrical) arranged, but in the majority of species they are of different shapes and sizes, so that a more or less two-lipped flower is produced; the ends of the segments are often drawn out into long tails or appendages. The mouth of the calyx tube is usually conspicuously hairy or papillose (bearing minute nipple-shaped projections) within. Petals are usually absent, though in the one genus (*Saruma*) in which they occur they are large and conspicuous. The stamens vary in number from six to many and may be free or variously united to the top of the ovary or the stylar (the upper part of the ovary) column; the anthers open by means of longitudinal slits. The ovary is usually inferior (*i.e.*, it is enclosed within the basal portions of the other flower parts), rarely superior. It is composed of occasionally four, but usually six carpels (ovary structures), which are free or loosely united. Each carpel contains numerous ovules. The styles are usually united into a column, and the stamens are often united to this, forming a compound structure known as a gynostemium, which usually occupies the swollen base of the perianth tube. The fruit is usually a capsule with numerous seeds. It opens by the carpels splitting away from each other. The seeds may be hard, immersed in a pulpy endocarp, or dry and often winged.

Evolution. *Fossil record.* Extremely little is known of the Aristolochiales in the fossil record; what is known sheds no light on the evolutionary history of the group.

Possible evolutionary origins

Phylogeny. The Aristolochiales order is a morphologically rather isolated group of uncertain origin. Some authorities suggest an origin in the order Magnoliales, which shares some characters with the Aristolochiales, notably uniaperturate (one germination pore) pollen grains (nonaperturate in some species of Aristolochiales), which are found only in these two plant groups among the dicots, and ethereal oil cells in the tissues. The frequent apocarpy (occurrence of separate carpels) of the Aristolochiales is another link with the Magnoliales, and the basic three-numbered structure of the flowers suggests an origin close to that of the monocots. The genus *Saruma*, which is the only genus of the Aristolochiales order to have petals, has been suggested as a link between the rest of the group and the Magnoliales order.

In some classifications the parasitic families Hydnoraceae, Cytinaceae, and Rafflesiaceae are included in the Aristolochiales. These families show many extreme adaptations to the parasitic habit, and their relationship to the Aristolochiales order is obscure. Some authors have also suggested a relationship between the Aristolochiales and the Cucurbitales order, but this has not been widely accepted.

Classification.

ORDER ARISTOLOCHIALES

Herbs, climbers, scramblers, or lianas, often woody, sometimes with anomalous secondary thickening. Leaves alternate, simple, with petioles (stalks), without stipules, often heart-shaped. Flowers solitary, racemose, or in clusters, bisexual, epigynous ("inferior" ovary) or rarely perigynous ("half-inferior" ovary). Perianth usually uniseriate (petals lacking) but biseriate with 3 sepals and 3 petals in *Saruma*, tubular below, with expanded actinomorphic or more usually zygomorphic (bilaterally symmetric), often strangely shaped limb above. Stamens 6 to many, usually in 1–2 series, or united to the top of the ovary or the stylar column (gynostemium). Anthers opening by slits. Ovary of (4–6) carpels, more or less free or weakly united, usually inferior. Ovules numerous in each carpel. Styles free or more frequently united. Fruit a capsule. Seeds numerous with fleshy endosperm and small embryo. One family with 10 genera and approximately 600 species, pantropical in distribution and with a few species in north and south temperate regions.

Family Aristolochiaceae

The only family in the order, hence it has the characters of the order. The genera, *Aristolochia* (including *Isotrema*, recognized by some authors as a separate genus) and *Asarum*, make up the bulk of the family, *Asarum*, with approximately 100 species, and *Aristolochia*, with nearly 500. A few species of each are grown as horticultural curiosities.

Critical appraisal. The order Aristolochiales appears to be a somewhat isolated group, not very closely related to any other order of dicots. Its position within the taxonomic hierarchy is, however, similar in most modern systems. Some authorities place it between the orders Piperales and Rafflesiales, noting that it is thought to derive directly from the Magnoliales order. Others endorse the presumed origin but place the Aristolochiales between the Piperales and the Nymphaeales orders, all three orders being considered close to the Magnoliales. Still others, mostly in older classification systems, include the parasitic families Hydnoraceae and Rafflesiaceae in the order and place it after the Piperales. Another classification also includes the Hydnoraceae and Rafflesiaceae families as well as the Nepenthaceae in the order and places it between the orders Berberidales and Piperales, suggesting its origin in the order Ranales (Ranunculales) via the family Menispermaceae.

In all these systems the order Aristolochiales is placed close to the Magnoliales/Ranunculales complex, and tends to be associated with the Piperales order and usually with the parasitic order Rafflesiales.

Recently, the theory has been put forward that the expanded perianth limb of *Aristolochia*, the genus in which the largest and most bizarrely shaped perianths occur, is not homologous of the trimerous, more or less actinomorphic perianths found in the other genera but is homologous of a foliage leaf. If this suggestion is accepted by other taxonomists, it will obviously be important in the classification of the group, and may introduce changes, but its confirmation requires the examination of many species of *Aristolochia*, as well as those of the other genera, which has not yet been done.

The group as a whole is in need of further study, particularly in its biology and taxonomy in the tropics, and such research will greatly extend knowledge of these curious plants.

BIBLIOGRAPHY. T. ARAKI, "Systema Generis Asari," *Acta Phytotax. Geobet.*, 15:33–36 (1953), an up-to-date taxonomic account of the genus *Asarum;* M.P. GRERORY, "A Phyletic Rearrangement in the Aristolochiaceae," *Am. J. Bot.*, 43:110–122 (1956), a well-illustrated account of the possible phylogeny

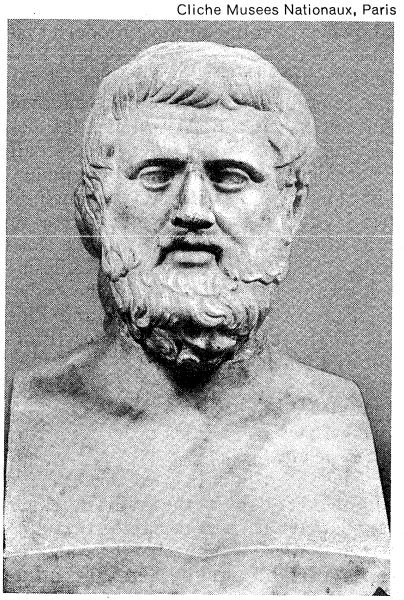
of the family, based on morphology and cytology; J.W. LORCH, "The Perianth of Aristolochia: A New Interpretation," *Evolution*, 13:415–416 (1959), a new suggestion about the morphology of the flower of *Aristolochia*.

(J.Cul.)

Aristophanes

The most famous of all comic dramatists of ancient Greece, Aristophanes is also the one whose works have been preserved in the greatest quantity (though, in fact, more of them have been lost than have survived). He is the only extant representative of the Old Comedy, that is to say, of the phase of comic dramaturgy in which chorus, mime, and burlesque still played a considerable part, reflecting that which they had played in the ritual origins of Greek drama; but he belongs to the end of this phase, and, indeed, his last extant play, which has no choric element at all, may well be regarded as the only extant specimen of the short-lived Middle Comedy, which, before the end of the 4th century BC, was to be superseded in turn by the more realistic productions of the New Comedy. Although the comedies of Aristophanes are often loosely constructed, inconsequential in the development of the plot, and feeble in characterization, they nevertheless have stood the test of time, having been frequently produced on the 20th-century stage or broadcast on the radio in one or another of numerous translations. This success must be attributed to the wit of the dialogue; to the generally good-humoured though occasionally malevolent satire; to the brilliance of the parody, especially when the author mocks the controversial tragedian Euripides; to the ingenuity and inventiveness, not to say the laughable absurdity, of certain scenes, comparable, indeed, to the fancies of the modern cartoonist; to the peculiar charm of the choric songs, the freshness and simplicity of which can still be conveyed effectively enough, even in languages other than Greek; and, at least for audiences of a permissive age, to the licentious frankness of many scenes and allusions.

Cliche Musees Nationaux, Paris



Aristophanes, restored Roman herm probably copied from a Greek original, fourth century BC. In the Louvre, Paris.

Very little is known about the life of Aristophanes, and most of the known facts are derived from references in his own plays. Born *c.* 450 BC, he was an Athenian citizen belonging to the tribe named Pandionis, but his actual birthplace is uncertain. (The fact that he or his father, Philippus, owned property on the island of Aegina may have given rise to an accusation of non-Athenian birth that his fellow citizens may have brought against him; see below.) He began his dramatic career in 427 BC with a play, the *Daitaleis* (*The Banqueters*), which appears, from surviving fragments, to have been a satire on his contemporaries' educational and moral theories. He is thought to have written some 40 plays in all (an average of one for every year of his activity as a dramatist). A large part of his work is concerned with the social, literary, and philosophical life of Athens itself; but an equally important part deals with themes provoked by the great Peloponnesian War (431–404 BC), which was essentially a conflict between Athenian imperialism and conservative Sparta and so was long the dominant issue in Athenian politics. A pacifist, Aristophanes was naturally an opponent of the more or less bellicose statesmen—from Pericles to Cleophon—who controlled the government of Athens throughout the better part of his maturity.

Importance of Peloponnesian War on his work

The *Babylonioi* (*The Babylonians*), another work extant only in fragments, was produced in 426 BC at the festival of the Great Dionysia, which was attended by delegates of the city-states, theoretically "allies" but practically satellite subjects of Athens. Because the play not only attacked Cleon, the demagogue then in power in Athens, but also showed the "allies" as the slaves of the Athenian Demus (a personification of the Athenian citizen electorate), Aristophanes was thereupon impeached; but whether the charge was treason or was based on a technicality arising from insubstantial allegations of non-Athenian birth is not known. In any case, Aristophanes got off lightly.

The *Acharneis* (425 BC; *The Acharnians*) is the earliest of the 11 comedies of Aristophanes that have survived virtually intact. This play is a forthright attack on the folly of war. Its farmer-hero, Dicaeopolis, secures a private peace treaty for himself from the Spartans in spite of the violent opposition of the chorus of embittered and bellicose old charcoal burners of Acharnae. The play contains two harsh comic scenes set in Dicaeopolis' private market: one in which a needy Megarian sells his two daughters, disguised as pigs, rather than let them starve at home; and another, in which a prosperous Boeotian, seeking a truly Athenian item, buys an informer.

The next year's major play, the *Hippeis* (424 BC; Latin *Equites*, English *The Knights*), shows how little Aristophanes was affected by the prosecution he had incurred for *The Babylonians*. It consists of a violent attack on the same demagogue, Cleon, who is depicted as the favourite slave of the stupid and irascible Demus until he is, at last, ousted from his position of influence and authority by one Agoracritus, a sausage seller with even greater claims than himself to scoundrelism and impudence.

The *Nephelai* (423 BC; *The Clouds*), which seems to mark a return to the subject matter of Aristophanes' first comedy, is an attack on "modern" education and morals as imparted and taught by the Sophists. In this play Socrates and his pupils are ridiculed, and, at the end of it, their school, the Phrontisterion (Thinking Shop), is burned to the ground. The author's selection of Socrates as the typical Sophist, although he was one of the Sophists' most hostile critics, has long worried scholars.

Relations with Socrates

The *Sphekes* (422 BC; *The Wasps*) satirized the litigiousness of the Athenians in the person of the old juryman, Philocleon (*i.e.*, Love-Cleon). The latter's son, Bdelycleon (*i.e.*, Loathe-Cleon), arranges for his father to hold a "court" at home; but, since the first "case" to be heard is that of the house dog, accused of the theft of a cheese, Philocleon is finally cured of his passion for the law courts.

The *Eirene* (early 421 BC; *The Peace*) was staged seven months or so after both Cleon and Brasidas, the two main champions of the war policy on the Athenian and Spartan sides, respectively, had been killed in battle and, indeed, only a few weeks before the ratification of the Peace of Nicias (? March 421 BC), which suspended hostilities between Athens and Sparta for six uneasy years. The comedy's main theme is the extraction of Peace from the cavern in which Zeus has allowed War to immure her.

The *Ornithes* (414 BC; *The Birds*), with its Cloudcuckooborough, or Land of Cockaigne, setting, can be regarded as a "comedy of fantasy." Some scholars, however, see it as a political satire on the imperialistic dreams that had led the Athenians to undertake their ill-starred expedition of 415 BC against Syracuse in Sicily.

Plot of *Lysistrata*

The *Lysistrate* (411 BC; *Lysistrata*) was written not long after the catastrophic defeat of Athenian enterprises in Sicily (413 BC) and not long before the revolt of the Four Hundred in Athens, whereby an oligarchic regime ready to make peace with Sparta was set up (411 BC). The play depicts the seizure of the Acropolis and of the treasury of Athens by the women who, at Lysistrata's instigation, have, together with all the women of Greece, declared a sex strike until such time as the men will make peace. It is a strange mixture of humour, indecency, gravity, and farce.

In the same year as *Lysistrata*, *Thesmophoriazousai* ("Women Celebrating the Festival of the Goddess Demeter") appeared—perhaps the funniest play Aristophanes ever wrote. Its plot is as follows: Euripides has discovered that the women of Athens, angered by his constant attacks upon them in his tragedies, mean to discuss during their coming festival the question of contriving his death. Euripides tries to persuade the effeminate Agathon, a tragic poet, to plead his cause. Agathon refusing, Euripides convinces his brother-in-law Mnesilochus to undertake the assignment. Mnesilochus is, in consequence, disguised (with great thoroughness) as a woman and sent on his mission. Unfortunately for him, his true sex is discovered, and he is at once seized by the women. There follow three scenes in which he tries to escape; all three involve parodies of Euripides' tragedies, *Telephus*, *Palamedes*, and (the extant) *Helen*, and all three attempts fail. Finally, Euripides himself arrives and succeeds in rescuing his advocate by promising never again to revile women.

The *Batrachoi* (405 BC; *The Frogs*) is a literary comedy. Dionysus, the god of drama, discontented with the poor quality of present-day tragedy now that his recent favourite, Euripides, is dead, disguises himself as the hero Hercules and goes down to Hades to bring Euripides back to the land of the living. As the result, however, of a competition arranged between Euripides and his great predecessor, Aeschylus, Dionysus is won over to the latter's cause and returns to earth with Aeschylus, instead.

In the *Ekklesiazousai* (391 BC; *Women in Parliament*), the women of Athens are again in power instead of the men. The plot introduces a communism of wealth, property, and sex that strongly recalls that discussed in the fifth book of Plato's *Republic*—so strongly, indeed, that Aristophanes seems to be parodying the philosopher's theories.

The *Ploutos* (388 BC; Latin *Plutus*, English *The God of Wealth*), last of the author's plays to be performed in his lifetime, is a somewhat drab work and does not enhance his reputation—though, as has been suggested above, it may have inaugurated the Middle Comedy. The poor but honest Chremylus, on the advice of an oracle, pursues and takes home with him a blind old man who turns out to be none other than Plutus, god of wealth. Chremylus brings about the cure of Plutus' blindness, whereupon the god, his sight now restored, gives up his old and bad associates and begins to consort with worthy people whom he can now see to be deserving of his company and patronage. Three elucidatory and not very interesting scenes bring the play to an end.

Shortly after producing his *Plutus*, Aristophanes died, leaving two plays (now lost), the *Aiolosikon* and the *Kokolos*, which his son staged *c.* 387 BC; both of them were probably mythological burlesques.

MAJOR WORKS

Acharneis (425 BC; Eng. trans., 1924); *Hippeis* (424 BC; *The Knights*, 1924); *Nephelai* (423 BC; *The Clouds*, 1924); *Sphekes* (422 BC; *The Wasps*, 1924); *Eirene* (421 BC; *The Peace*, 1924); *Ornithes* (414 BC; *The Birds*, 1924, 1950); *Lysistrate* (411 BC; Eng. trans., 1924); *Thesmophoriazousai* (411 BC; *The Thesmophoriazusae*, 1924; *Ladies Day*, 1960); *Batrachoi* (405 BC; *The Frogs*, 1924); *Ekklesiazousai* (391 BC; *The Ecclesiazousae*, 1924; *Women in Parliament*, 1929); *Ploutos* (388 BC; *The Plutus*, 1924). A translation (1924) of Aristophanes' works is available in the "Loeb Series," and of several plays in the "Penguin Classics Series" (1964).

BIBLIOGRAPHY. The best modern collected editions of Aristophanes' extant comedies are those by JOANNES VAN LEEUWEN with notes in Latin (1893–1909); by B.B. ROGERS with translation into English verse (1902–16; Loeb ed., 1924) and by RAFFAELE CANTARELLA with Italian translation (1948). The fragments are edited by J.M. EDMONDS in his *Fragments of Attic Comedy*, vol. 1 (1957). There is an *Index* to Aristophanes compiled by O.J. TODD (1932, reprinted 1962). The English reader, for an introduction, may consult GILBERT MURRAY, *Aristophanes* (1933, reprinted 1964); and VICTOR EHRENBURG, *The People of Aristophanes*, 3rd ed. (1962).

(M.Pl.)

Aristotelianism

The term Aristotelianism can refer: (1) to philosophical activities organized into schools that are devoted to the study, elaboration, and application of Aristotle's methods, inquiries, and theories and their comparison or coordination with other doctrines; (2) to a deliberate adherence to tenets rightly or wrongly considered basic in Aristotle's philosophy; (3) to an extensive adoption of concepts and procedures characteristic of Aristotle but not necessarily recognized as such; (4) to philosophical trends having little to do with Aristotle but emerging from his doctrine after slow and radical transformations; and (5) to mental attitudes, spontaneous interests, and approaches similar to those of Aristotle. As a general term, Aristotelianism also covers (6) philological and historical scholarship concerned with Aristotle's works and life.

NATURE OF ARISTOTELIANISM

Characteristic features

Features that cumulatively or severally characterize Aristotelianism include, in philosophical methodology, a critical approach to previous, contemporary, or hypothetical doctrines; the raising and discussing of doctrinal difficulties; the use of deductive reasoning proceeding from self-evident principles or discovered general truths to conclusions of a more limited import; and syllogistic forms of demonstrative or persuasive arguments. In epistemology, or the theory of knowledge, Aristotelianism includes a concentration on knowledge accessible by natural means or accountable for by reason; an inductive, analytical empiricism, or stress on experience, in the study of nature—including men, their behaviour and organizations—leading from the perception of contingent individual occurrences to the discovery of permanent, universal patterns; and the primacy of the universal, that which is expressed by common or general terms. In metaphysics, or the theory of Being, Aristotelianism involves belief in the primacy of the individual in the realm of existence; in correlated concepts allowing an articulate account of reality (*e.g.*, 10 categories; genus–species–individual, matter–form, potentiality–actuality, essential–accidental; the four material elements and their basic qualities; and the four causes—formal, material, efficient, and final); in the soul as the inseparable form of each living body in the vegetable and animal kingdoms; in activity as the essence of things; and in the primacy of speculative over practical activity. In the philosophy of nature, Aristotelianism denotes a position optimistic concerning nature's aims and its economy; believing in the perfection and eternity of the heavenly, geocentric spheres, driven by intelligent movers, carrying in their circular movements the stars, sun, planets, and moon, and influencing the sublunar world; and holding that light bodies rise naturally away from the centre of the earth, while heavy bodies move naturally toward it with a speed related to their weight. In aesthetics, ethics, and politics, Aristotelian thought holds that poetry is an imitation of what is possible in real life; that tragedy is directed to purification (catharsis) through fear and pity; that virtue is a middle between extremes; and that man is intrinsically political and the state is a self-sufficient society, necessary for men to achieve happiness. Aristotle's early views on the immortality of the soul and the relations between gods and heavens account for some aspects of the more ancient Aristotelianism.

HISTORY OF ARISTOTELIANISM

Continuity of the Aristotelian tradition. Since Aristotle's death (322 BC) there have been, without interruption until the present, schools and individuals who have cultivated the study of his works and fully or partly

adopted and expounded his doctrines and methods. They have interpreted or misinterpreted, approved or condemned, and reshaped or utterly transformed them. The languages in which this interest was most forcibly expressed have changed in turn from Greek to Latin; to Syriac, Arabic, and Hebrew; to Italian, French, English, and German. The main centres in which it appeared have been as far apart as Greece, North Africa, and Rome in the ancient world; Persia and Spain, Sicily and Britain in the Middle Ages; and Germany and North America in more recent times.

Main centres and traditions

The main course of the Aristotelian tradition has been the Greek line, which lasted 2,000 years, mainly in the area on the eastern Mediterranean Sea, and branched off at many stages between the 4th and 15th centuries giving rise to (or strengthening) other traditions. The Latin branch originated in Rome in the 4th century and acquired a first new impulse, probably from Athens, in the early 6th century; from these beginnings it was revived in the 9th century and again in the 12th, at which time a second and stronger Aristotelian wave emerged from Constantinople and then a third, via the western Arabic schools, from Spain; and both spread to Italy, France, and Britain. The final direct contribution from the Greek to the Latin tradition came to Italy, once more from or through Constantinople, in the 15th century.

Shortly after the beginning of Latin Aristotelianism, certain Armenian and Syrian members of the Greek schools of Athens and Alexandria in Egypt transplanted Aristotelian teachings into their schools. The Armenian tradition was still alive in the 19th century in such places as Madras (India) and Venice; and the Syrian, which never completely disappeared, was still powerful in the 14th century, after having given birth, in the 9th and 10th centuries, to an Arabic tradition. Arabic Aristotelianism was the product of Syrians, Persians, Turks, Jews, and Arabs who wrote and taught in their countries and in Africa and Spain until the 12th century. Much of it and of what the Jews produced in Hebrew in the following two centuries, passed into the Latin tradition between 1130 and 1550. Thus, all of the varied heritage that had derived ultimately from the Greek line and had been vastly enriched by other cultures came to be collected, through the Latin, by modern Western (and other) philosophical movements.

The Greek tradition. The Greek tradition is here traced from Aristotle's school through the 15th century AD.

Early development. For nearly three centuries after his death Aristotle's own school called the Peripatos or Lyceum) remained, in a truly Aristotelian spirit, a centre for critical research—not for the dogmatic acceptance of a closed system. Aristotle's immediate successor, Theophrastus, independently elaborated his master's metaphysics and psychology and added to his study of human behaviour, nature (botany and mineralogy), and logic (theory of propositions and hypothetical syllogisms). Other members of the Peripatos coordinated Aristotelian thought with other current schools of philosophy. Thus, Aristoxenus joined Aristotelian and Pythagorean doctrines; Critolaus united Aristotle's theory of the influence of the heavens on the world with the Stoic theory of providence; and Clearchus of Soli combined Plato's views on the human soul with Aristotle's.

The Stoics, Skeptics, and Epicureans

Outside the Peripatos, the Stoic school was partly following Aristotle in their interest in formal logic, the theory of meaning, and use of the categories (*e.g.*, substance, quality, relation). They were Aristotelian also in their empiricism and their use of inductive theory, as well as in their concentration on nature, in several aspects of natural philosophy, and in their belief that man is intrinsically a social being. The Skeptics sometimes relied on Aristotelian forms of argument to prove their systematic doubts. Even Epicurus, who may have fought against Aristotle's early theology and psychology and ignored his mature philosophy, was near him in his doctrine of the will and in his conception of friendship and the pursuit of knowledge as the high aims that give satisfaction and pleasure to man. Although little was known of Aristotle's doctrinal works in the early 1st century BC, his more popular, literary, and Platonizing writings influenced eclectics such as Panaetius and his pupil Poseidonius; and this influence continued, by way of the Roman philosopher and lawyer Cicero and Greek texts, well into the 4th and 5th centuries. Upon it was based the tendency to harmonize Plato and Aristotle—a feature that recurred through the whole history of Aristotelianism—and perhaps the ascription to Aristotle of the *De mundo*, a cosmological treatise of the 1st century BC, which found favour with all of the different traditions until the 16th century.

In the 1st century BC an almost complete set of Aristotle's doctrinal writings was discovered and was organized into a corpus, critically edited by Andronicus of Rhodes and other scholars, and used by Nicolaus of Damascus, a historian and philosopher, in attempts to expound Aristotle's system. Thus began a new era of a scholarly and scholastic Aristotelianism in which Aristotle had to be taken as the basis for the acquisition of true knowledge in many fields. Individual works began to be commented and lectured upon; organized philosophical studies began to have as their introduction Aristotle's works on logic, especially the *Categories*. Thus, the pattern was set for the next 17 centuries. Almost pure Aristotelianism, based on the doctrinal works, lived on until the 4th century. Many scholars—the most eminent of them being Alexander of Aphrodisias, who from AD 195 held the Athenian chair of Aristotelian studies created by the Roman emperor—provided the works on logic, ethics, metaphysics, natural philosophy, and psychology with detailed and penetrating commentaries for the specialist. The interpretation of Aristotle was for many generations molded by these scholars. Others—the greatest being Themistius, a professor in Constantinople about AD 350—practically rewrote many of Aristotle's treatises in a more modern language and more readable style.

This second Aristotelianism had established itself sufficiently as the philosophical and methodical frame of learning for it to be adopted, at least in part, by most men of culture such as Ptolemy, a geographer and astronomer, and Galen, an encyclopaedist, logician, and medical scientist. The latter added new syllogistic forms to those discovered by Aristotle, developed his views on universals in a nominalistic direction (which holds them to be mere names), and argued in favour of the mortality of the soul.

Relationship to Neoplatonism. Aristotle's works were adopted as a methodical basis and source of specialized knowledge by the systematic builders of Neoplatonism in the 3rd century. Plotinus, the school's chief representative, followed Aristotle wherever he found a possibility of agreement or development, as in his theory of the intellect. And Plotinus' pupil Porphyry, the first great harmonizer of Plato and Aristotle, provided the field of logic with a short introduction (*Isagoge*). The *Isagoge*, in fact, is only concerned with a simple, rather mechanical treatment of five concepts (genus, species, differentia, accidental feature, and property), which had been much used by Aristotle. This introduction soon became an integral part of the *Organon* (a modern name for the collection of the logical works of Aristotle) and thus acquired undeserved Aristotelian authority in all schools for over 1500 years. From that time on, Aristotelianism became indissolubly tied up with Neoplatonism.

Methodological basis of Neoplatonic thinkers

Neoplatonism dominated the school of Athens, where, apart from logic, Aristotle's writings were destined to be studied mainly as a basis of debate in favour of Plato. Scholars like Ammonius—a pupil of Proclus, the most accomplished systematizer of Neoplatonism, head of the Athenian school in the mid-5th century, and himself well-versed in Aristotle—found Alexandria a more attractive place for Aristotelian studies, open to many views, liberal and tolerant. There pagans and Christians co-existed and cooperated, and from there they carried Aristotelian learning to other schools: Simplicius, a pupil of Ammonius who was inclined to Platonism, took it back to Athens and—when Justinian closed that pagan school in

529—to Persia; Sergius, a physician and Nestorian priest, carried it to the Christian schools of Syria; and Stephanus of Alexandria took it to Constantinople. The schools of Alexandria and Athens had produced from about AD 475 to 545 the most intensive collection of Aristotelian commentaries, by scholars like Ammonius, philosophers of science like Simplicius, and philosopher-theologians like Philoponus (see also PLATONISM AND NEOPLATONISM).

Relationship with early Christian theology

Before the 5th century, Christian theology had been affected only marginally and indirectly by Aristotle. The elementary study of Aristotelian logic had proved indispensable for a disciplined training of theologians; and some of the concepts from Aristotle's *Physics* and *Metaphysics* that entered into the elaboration of this logic became equally essential for the formulation, in rational terms, of points of dogma. The aforementioned five terms of Porphyry and the 10 categories of Aristotle were used or implied in the mystical theology of Pseudo-Dionysius (an unidentified 5th-century Christian Neoplatonist), which was to become one of the principal components of Christian speculation in the Greek, Oriental, and Latin schools. Descriptions of God and distinctions between the three Persons of the Trinity came to include, in an increasingly technical sense, the Aristotelian terms substance, essence, accident, form and matter, species, and nature, quality, quantity, and property; these terms were not always used in a purely Aristotelian sense, however. In this way as well as through the purely philosophical schools Aristotelianism entered the first Greek Scholasticism of St. John of Damascus, an 8th-century doctor of the church.

From the Byzantine renaissance to the 15th century. The Byzantine renaissance of studies in the 9th century included a revival of interest in Aristotle: the old books were rediscovered and re-edited (the oldest manuscripts still existing today belong to that time). Photius, patriach of Constantinople and leading figure of that renaissance, included in his encyclopaedic works summaries of the elements of Aristotelian logic. More extensive scholarly activity resulted from the re-establishment of the Academy in Constantinople in the 11th and 12th centuries under the successive leadership of such men as Michael Psellus, an encyclopaedic philosopher; his student John Italus; Michael, the archbishop of Ephesus; and Eustratius, the metropolitan of Nicaea. Teaching and exegetic work went hand in hand; debates on the superiority of Plato or Aristotle and attacks on philosophy by the religious schools did not seriously weaken these activities. There was perhaps not much new in the understanding or development of Aristotle's doctrine; but logic was no more the only focal point of Aristotelian studies. They covered, more widely than had been done in Alexandria, practically the whole corpus, including some work on the *Politics* and concentrating on books of ethics and biology. Some of the philosophical debates were similar to those taking place in the Latin schools; they were based on texts of Aristotle and treated such points as logic of language and theories of universals.

In the 13th and 14th centuries, popularization and systematization—in an encyclopaedic or philosophical form—took the upper hand in the work of Nicephorus Blemmydes, George Pachymeres, and Theodore Metochites. At this time when Greek thought was being strongly influenced by the Latin tradition, especially by the thought of Thomas Aquinas, the traditional debate on Plato and Aristotle took new forms. Aristotelianism appeared in the teaching of Barlaam the Calabrian to champion rationalism in faith; this was combated from a Platonic point of view by Nicephorus Gregoras. In the 15th century, when Greeks were becoming part of the Italian philosophical scene, Aristotelian rationalism was strongly defended by such upholders of Christian theology as Gennadius Scholarius against the proponents of a new universal religiosity linked with admiration for Plato and paganism such as George Gemistus Plethon. The victory in this intellectual battle was reserved to the moderates like John Bessarion, Plethon's influential pupil, who, though he preferred Plato, admired Aristotle, translated his *Metaphysics*, and collected manuscripts of his works; he became a cardinal in the Roman Church.

The early Latin tradition. The echoes of Aristotle's early writings in Cicero, a few signs of his indirect influence on other writers, and a more considerable contribution of post-Aristotelian logic in Apuleius, a Platonic philosopher who flourished in the 2nd century AD, are indications of the general cultural intercourse between Latins and Greeks. The presence of Plotinus and Porphyry in Rome in the 3rd and early 4th centuries probably started the more serious interest in Aristotle there, of which the first results were, perhaps, Victorinus' adaptations in Latin of Porphyry's *Isagoge* and Aristotle's *Categories*. Logic was still the only part of Aristotle that had entered Latin culture when Themistius' teaching attracted the attention of Roman pagan circles in the 4th century.

Influence on Boethius

Again, only the logical works of Aristotle, together with some extracts from Greek commentaries on them, seem to have reached the hands of Boethius, a Roman scholar and statesman, in the early 6th century, when he was attempting to transfer to the Latins as much as he could of Greek learning. He translated these works and elaborated on the commentaries and some other later texts of logic partly based on Aristotle. He acted as a transmitter, and thus it is unsafe to ascribe to him interpretations and plans contained in the Latin works that bear his name. In the style of the time, the transmitter and imitator would use the first person when repeating what he had taken over and approved of. Even his plan of commenting on "as much of Aristotle as would come into his hands" and showing that Aristotle and Plato agree was the traditional plan going back at least to Porphyry. Nothing remains to show where he himself stood in judging Aristotle and the several parts of his philosophy. The same observations probably hold true regarding Boethius' theological treatises, in which the Aristotelian concepts that helped to organize the theology of the Trinity were clearly taken over from similar Greek treatises. A disproportionate value was later attached to Boethius' own original contribution in both logic and theology; the fact that his name was connected with these texts made people in the Middle Ages ascribe to him the primary merit for their contents.

The Syriac, Arabic, and Jewish traditions. The increased sense of linguistic and national identity and the religious movements of the 5th and 6th centuries such as Nestorianism (a heresy that so stressed the distinction between the divine and human natures of Jesus Christ as to suggest that they were two persons joined by a moral union) and Monophysitism (a heresy that asserted that there is only one nature in Jesus Christ) led to the foundation of Syriac centres of studies in the Persian and Byzantine empires, especially at Edessa and Antioch. Proba and Sergius of Resaina are among those who contributed, through translations of the basic logical texts and commentaries on them, to the establishment of Aristotelian studies in these centres. At the time of the Arabic invasion of the Byzantine and Sāsānian empires around 640 and for several generations afterward, these centres—especially a great school of Kinnesrin, which was represented by such men as Severus Sebokht, who wrote on Aristotle's syllogisms; Jacob, bishop of Edessa, a theologian, grammarian, and translator; and Georgius, bishop of the Arabs, author of a commentary on the *Organon*—grew in importance and productivity. They remained, however, mainly confined to logic and its application to theology. The Syrian Christians formed the philosophical and scientific intelligentsia when in the 9th century al-Ma'mūn, the seventh ʿAbbāsid caliph, organized the Arabic centre of learning of the new Islāmic empire in Baghdad. By then the Syrian scholars had extended their acquisition and translation of Aristotle to most of his works. They also then translated them into Arabic, both from the Syriac and directly from the Greek, and added many texts of commentators of Aristotle. In this way Ḥunayn ibn Isḥāq, his son Isḥāq, Abū Bishr Mattā ibn Yūnus, Yaḥyā ibn ʿAdī, and many other Syrians provided the basis for a brilliant philosophical activity in Arabic. The Syrians retained their own independent culture; as

late as the 13th century, their language was used by the converted Jew Bar Hebraeus, "Son of the Hebrew" (also known as Gregorius or Abū-al-Faraj), an encyclopaedist, philosopher, and theologian, who expounded all the works of Aristotle in his *Kĕthabhā dhē-ḥēwath ḥekhmĕthā* (*Book of the Cream of Wisdom*), elaborating many sections on the basis of the Greek and Arabic Aristotelians.

In the 9th century the Arab al-Kindī was the first notable scholar to use the Arabic language in a general introduction of mainly Aristotelian philosophy. In the following century the Turkish Muslim al-Fārābī produced a more specialized study in which he commented upon and expounded the books of logic and attempted to establish the relationship between philosophy and Islām. It was through Avicenna and Averroës, however, that Aristotle became an integral part of lay Arabic culture. Early in the 11th century, the Persian Avicenna (Ibn-Sīnā) made Aristotle's philosophy the foundation of his personal, original system, for which he also found inspiration in a group of Plotinian texts that had passed into Arabic under the title "Theology of Aristotle." Aristotle became, in Avicenna's hands, much more systematic and coherent than he had intended to be; problems and solutions that were, at best, hinted at by Aristotle (*e.g.*, the distinction between essence and existence or the relation between possible and necessary existence) were among the distinctive marks of Avicenna's own work. For the Spanish Arab Averroës (Ibn Rushd) in the 12th century, Aristotle was "the measure and model offered by nature to show the ultimate perfection of man." Philosophy, specifically Aristotelian philosophy, is and teaches truth; revelation or revealed religion is a debased philosophy for the simple. Averroës dissected Aristotle's works, analyzing and reconstructing them with a fine scholarly and philosophical sense and an incredible wealth of information derived from previous Greek and Arabic philosophers. He elicited doctrines that are not easily apparent, made them in some cases more compelling than the texts themselves might allow, but rarely forced onto Aristotle his own views without at least finding some support in the texts themselves. The mortality of the individual soul, the eternity of the world, and the unicity of the Mind for the whole human race to the exclusion of individual minds are key doctrines for Averroës; they have some basis in Aristotle.

Contributions of Avicenna and Averroës

Until the 13th century, Jewish Aristotelianism developed within the Arabic culture of North Africa, Mesopotamia, and Spain. This work was done in the Arabic language and distinguished itself for its almost constant concern with the relation between philosophy and Judaism. Many Aristotelian concepts were considered and discussed by Isaac ben Solomon Israeli, a 10th-century Neoplatonist, in his *Kitāb al-ḥudūd* ("Book of Definitions") and *Kitāb al-usṭuqusāt* ("Book on the Elements"). Form and matter are the basis of the metaphysical structure of the Neoplatonic system of Solomon ibn Gabirol, an 11th-century poet and philosopher known as Avicebron. A fully conscious plan of inserting Aristotle—or at least the Aristotle of al-Fārābī and Avicenna—into the intellectual and spiritual life of Judaism was carried out by Abraham ibn Daud of Toledo in the mid-12th century. Moses Maimonides of Cordova found a way of balancing knowledge by experience and reason with theology and revelation, in clear contrast with his contemporary Spaniard Averroës; he provided a Jewish anticipation of Aquinas' Christian compromise. His proofs of God's existence and his acceptance of a theory of creation from eternity are typical of his approach. From the 13th century, philosophical works, particularly those of Averroës on Aristotle, were being translated into Hebrew; a vast Hebrew literature of "super-commentaries" (those on the works of Aristotle as commented on by Averroës) and independent works was produced by Levi ben Gerson (Gersonides), who was faithful to both Maimonides and Averroës, and by others. Soon after, however, the more orthodox tradition based upon the Bible and the Talmud prevailed. Aristotelian works by Jews and Hebrew versions of Averroës, translated into Latin, contributed their share to the Italian philosophical movement of the 16th century (see also JEWISH PHILOSOPHY).

Relation between philosophy and Judaism

The later Latin tradition. The Latin tradition is here traced from the 9th through the early 15th century.

The discovery of Aristotle's works in the Latin West. Before 1115 only the very short *Categories* and *On Interpretation* were known in Latin (from about 800, in Boethius' version); by 1278 practically the whole of the Aristotelian corpus existed in translations from the Greek, and much of it had a wide circulation. Apart from three other works of logic in Boethius' text, which reappeared about 1115, this wholesale discovery was the result of cultural contacts with Constantinople and a few other Greek centres and the personal initiative of a few scholars. Most notable and first of these was James of Venice, who was in Constantinople and translated the *Posterior Analytics*, *Physics*, *On the Soul*, *Metaphysics*, and other minor texts before or around 1150; other scholars translated anew or for the first time works on ethics, natural philosophy, and logic before 1200. With higher standards of linguistic scholarship, Robert Grosseteste, about 1240, revised and completed the translation of *The Nicomachean Ethics* and translated for the first time from the Greek *On the Heavens*. The Flemish translator William of Moerbeke, between about 1255 and 1278, completed the Latin Aristotelian corpus; he was the first to translate the *Politics* and *Poetics*, and to give a full and reliable translation of the books on animals; he also translated anew some books of natural philosophy, and revised several of the older translations. About half of the works were also translated from the Arabic, mainly in Toledo by Gerard of Cremona and Michael Scot, between 1165 and 1230. With two or three exceptions, these translations came after those from the Greek; all had a much more limited circulation and influence. A considerable contribution to the knowledge of Aristotle came from the translations of the ancient commentaries; nearly all of these were made from the Greek.

Latin translations

The view that Aristotle came to be known in Latin by way of the Arabs must be understood as true only in the sense that a number of Aristotelian doctrines—partly transformed in the process—spread in Latin circles from the works of such men as al-Fārābī, Avicenna, and Abū Ma'shar before the texts of Aristotle were accessible or had been properly interpreted. Further, there is little truth in a view that in the Latin Middle Ages Aristotle was seen in a Neoplatonic light because Plotinian and Proclian texts translated from the Arabic—namely *Theologia Aristotelis* ("Theology of Aristotle") and *Liber de causis* ("Book of Causes")—were ascribed to him.

From the 9th through the mid-13th century. The study of Porphyry's *Isagoge*, of Aristotle's *Categories* and *On Interpretation*, and of theological texts containing Aristotelian elements formed the basis, from the 9th century onward, of logical methodology (dialectic) in many fields. When applied to problems concerning the Trinity or the Eucharist or in general to problems concerning individuality and universality of concepts and things, dialectic was seen as a powerful instrument for clarifying faith or —on the opposite side—for endangering it. For Abelard, the first great Aristotelian of the Middle Ages, dialectic was essential for the analysis and discovery of truth. As part of his study of it, he produced an illuminating account of the linguistic, mental, and objective aspects of universals on the basis of Aristotelian doctrines. Soon after, new developments of Aristotle's theory of language and logic took place, partly as a result of the recently acquired knowledge of his work *On Sophistical Fallacies*. New fields were opened up, and logic was no more just Aristotelian; with this, modern logic had its beginning on a limited Aristotelian basis.

At the same time, in the second half of the 12th century, Aristotle's physics, cosmology, and metaphysics began to attract attention through the Latin texts of Arabic works on science and philosophy and of Aristotle's own works, mainly among scientists of the famous medical school at Salerno and the English philosophers. Alfred of Sareshel used the new texts in his treatise *De motu cordis* ("On the Movement of the Heart"); Robert Gros-

seteste produced the first and best detailed Latin commentary on the *Posterior Analytics;* Roger Bacon commented on the *Physics* and part of the *Metaphysics.* It would be wrong, however, to try to find in this scholarship the origin of modern experimental science, which is rather to be found in the study of ancient and more recent mechanics, medicine, and technology or in original inventiveness.

Reactions to Aristotelianism at the University of Paris

The introduction of the new Aristotle met with difficulties in Paris. The impact of non-Christian Aristotelian and Arabian philosophy engendered fears, doubts, and suspicions. Although the masters at Paris were free to teach Aristotle's logic and ethics and although no obstacle at all was put in most universities, in the first part of the 13th century the ecclesiastical authorities at Paris imposed a ban on lectures relating to the natural philosophy and metaphysics of Aristotle and his commentators. This slowed down some activities but quickened reactions and curiosity; the very need for some kind of censorship of the works led to more intimate study of them. The controls soon became loose. Before 1250 Roger Bacon commented on those "dangerous texts" at Paris, and soon Albertus Magnus—committed though he was, as a Dominican priest, to safeguard the purity of faith and dogma—made Aristotle's works an indissoluble part of Latin philosophical and scientific literature. Albertus' vast encyclopaedia of nonrevealed knowledge and wisdom consisted of an analytical exposition of Aristotle combined with all the information and interpretations that Albertus had gathered from other, mainly Arabic, sources and that he had gained as the product of his own research and speculation. Faced with the danger of being accused of following Aristotle against dogma, he asserted: "I expound, I do not endorse, Aristotle."

The approach of Albertus' pupil, Thomas Aquinas, to Aristotle was as a scholar. He analyzed the structure of every section of most works; he tried to discover their organization and to follow the arguments; and he was careful to obtain the best texts and to get from them the genuine meaning. He respected Aristotle's authoritativeness and reasonableness, even when it was not explicitly justified, sometimes drawing inferences that went beyond Aristotle's own conclusions and allowing himself considerable freedom whenever Aristotle had left loose ends in his attempts to solve difficulties. At these points he often went his own way, without ascribing the new steps to Aristotle but without feeling that he was going against him. Compromises followed, sometimes bearing the mark of nearly impudent genius; *e.g.*, he stepped beyond Aristotle when he argued that the individual soul, although remaining essentially and indissolubly the form of the individual body, is separable from it and immortal. Aristotle was stretched almost to the breaking point, but he was not transformed. Beyond that point Aquinas was not a Christian Aristotle but a man of faith and dogma; he divorced himself from Aristotle when necessary and approached closer to St. Augustine, to the Neoplatonists, or to Avicenna.

From the late 13th century through the 15th century. The suspicion that Aristotle might lead to heresies became stronger when the closer study of his texts and of Averroës' interpretations had enhanced the admiration for the philosopher and increased the following of the commentator, as they were known. Siger of Brabant was the most redoutable of many Averroistic Aristotelians. What came to be called Averroism was in fact a tendency toward an integral acceptance of genuine or consistent Aristotelian tenets, particularly concerning the achievement of human happiness in a nonsupernatural life. Ecclesiastical condemnations of a number of propositions considered false or dangerous and threats against the holders of doctrines implied by them gave a more definite status to the Averroists, although many propositions condemned at Paris and Oxford in 1270 and 1277 had nothing to do with Aristotle and little with Averroës. The effect of the condemnations soon became visible: the separation between the teaching of "philosophy" in the faculty of arts and of "truth" in the faculty of theology became rigid, with the ambiguous result that two "truths"—truth of coherence in philosophical contexts and revealed truth—were thought to coexist. At the turn of the century, however, Dante's poetical vision could still merge the Averroists' Aristotle who claimed a self-sufficient nature and Aquinas' Aristotle who endorsed many of the truths of faith. For Dante, as for Averroës, Aristotle is the embodiment of total human knowledge and the master of wisdom. In Aristotle's *Nicomachean Ethics* and *Physics* Dante found moral guidance and the key to understanding the workings of nature. In some aspects of Averroës' theory of a universal human Intellect combined with the Stoic–Aristotelian principle that all men are by nature citizens of one city, he found the basis of the Empire, the one polity (*civilitas*) for the whole human race.

Views of Scotus and Ockham; Averroism

The 14th century was no less Aristotelian than the 13th, despite the claim that Aristotelianism collapsed. Such an assertion does not take into account the non-Aristotelian components of previous philosophies and the permanent acceptance of Aristotelian doctrines in the new ones. Form, matter, causality, and the intrinsic combination of regularity and contingency were the Aristotelian frame of the system of Duns Scotus. The "terminism" of William of Ockham, a British Franciscan, his rejection of "useless entities," his metaphysics of a world of individual self-contained things, and his conceptualism gave neat, though extreme, expression to Aristotle's theory of language, the economy of nature, and the primacy of individuals in existence and of universals in intellectual knowledge. He followed Aristotle closely in his views on the scientific coordination of notions. He was more faithful to Aristotle than either Aquinas or Averroës when he said that Aristotle did not give a clear lead on the question of the immortality of the soul. The various schools of Scholastic philosophy—Thomism, Scotism, Ockhamism—that asserted themselves in this century and lived on had a common Aristotelian basis but different ways of interpreting it (see also SCHOLASTICISM; CHRISTIAN PHILOSOPHY).

Averroistic Aristotelianism flourished in this century in connection with, or independently of, the other trends. The Italian medical faculties at Bologna and Padua were lively centres of logical and philosophical studies; *e.g.*, Peter of Albano, a professor of medicine at Padua who had been trained at Paris, pushed Aristotle's cosmology onto the brink of determinism in human affairs and used his logic to suggest that Christ's death was only apparent. Political science, which had been a field for lofty speculations or restrained exercises in the analysis and exposition of texts, became important for the politicians and people who wanted to satisfy, under the aegis of Aristotle's doctrine, the potentialities of human beings for happiness. John of Paris wanted France to be self-sufficient, self-controlling, and without interference from the pope; John of Jandun, a successor of Siger of Brabant, upheld Aristotle's *Politics* in all its worldliness; and Marsilius of Padua, Jandun's friend in Paris, applied Aristotle in his insistence that priests should be considered in the same way as members of a guild in a city, with no superior privileges.

Perhaps with less attachment to the details of Aristotle's doctrines and with a keen critical sense, the Mertonians at Oxford and encyclopaedists, scientists, and philosophers in France (*e.g.*, Jean Buridan and Nicole d'Oresme) carried on laborious efforts toward the mathematization of science, the quantification of changes in quality, and the determination of the nature of continuity in movement and the acceleration and speed of falling bodies. Their starting points were the *Physics* and the other texts of Aristotle. In a similar (almost mathematical) spirit, many of the same men carried logic even further than Ockham had done into the fields of logical calculus, paradoxes, and sophisms. Aristotle was not abandoned but expanded.

Modern developments. The Aristotelian tradition is here traced from the 15th through the mid-20th century.

From the Renaissance to the 18th century. In the 15th century Italy became the focal point at which various forms of Aristotelianism converged. Certain links be-

Aristotelianism in 15th-century Italy

tween Italian universities and religious schools and the University of Paris had flourished for a long time. In the late 14th century, Paolo Nicoletti (Paulus Venetus) returned from Oxford to Padua after absorbing the new logic and physics of the Mertonians and the radical terminism of Ockham and after increasing his acquaintance with the French Averroistic trend; works by the Englishmen and Paolo were textbooks in Italian universities for many generations. At the end of the century Spanish and Italian Jews were passing on, in Latin, more texts of Averroës on Aristotle and the Jews' own recent contributions to Aristotelian learning.

A more spectacular contribution of books, linguistic and didactic competence, and stimulating debates came with an influx of Greek scholars. They were attracted by the Humanists' craving for classical learning, the theological discussions between Orthodox and Roman Catholic leaders, and the relative freedom offered by the Republic of Venice and by Florence to those who were taking refuge from Turkish domination. Many manuscripts were taken to Italy, and many were transcribed in Italy by the Greeks, who also taught the Greek language to the Italian scholars. An editorial masterpiece by Aldus Manutius, an early printer, publisher, and editor, at the end of the century made accessible to many almost the complete Greek corpus of Aristotle's works. A great number of Greek and Latin scholars—such as Bessarion, John Argyropoulos, Leonardo Bruni, Lorenzo Valla—produced new translations of those texts; others translated many works on Aristotle previously unknown in Latin.

As soon as printing had been established (*c.* 1475), editorial activity was directed to the production of many complete and partial editions of the Latin versions of Aristotle and Averroës in the older and newer versions from the Greek, Arabic, and Hebrew. At the universities of Padua and Bologna and at Ferrara and Venice, Averroists such as Agostino Nifo and Nicoletto Vernia and independent interpreters such as Pietro Pomponazzi were dominating the philosophical scene. For Pomponazzi, Aristotle, whether right or wrong, had to be studied directly by way of his own works and not by way of his interpreters; knowledge, however, was not limited to Aristotle, and his mistakes concerning facts were to be exposed. Others followed Aristotle in his vast scientific achievements or searched his works for a clearer formulation of scientific methods. This scientific spirit kept alive the interest in Aristotle's methodology and his philosophy of nature down to the time, in the 17th century, when William Harvey, an English physician and anatomist, was lecturing on Aristotle's books on animals and Galileo was writing on logic.

Influence on methodology, cosmology, metaphysics, and politics

In a less apparent form, Aristotelianism was still strongly entrenched in most European schools affecting the most advanced "modern" philosophers. The methodology of Francis Bacon, English philosopher, scientist, and statesman, grew out of it, and his basic metaphysical concepts were borrowed from Aristotle. The Polish astronomer Copernicus was still attached to the perfection of circular movements. Gottfried Wilhelm Leibniz, a German Rationalist and mathematician, not only admired Aristotle's logic but also built his own metaphysics of individuals ("monads") around the theory of matter and form. Like Aristotle, political theorists such as Jean Bodin in France carried on their inquiries on the nature of the state by studying existing organizations and their natural backgrounds. In the literary field, Aristotle's *Poetics*, practically unknown until 1500, was now read and analyzed in the Greek text and Latin versions; its doctrines were compared and partly harmonized with the then-prevailing views of the Roman poet Horace, and Aristotle's view that art imitates nature prevailed for many over the conflicting theory that stressed the basic creativity of the poet. The doctrine of the unities of action, place, and time—though actually a later development resulting from forced interpretations of Aristotle—ruled over the work of many writers of tragedies (*e.g.*, Gian Giorgio Trissino in Italy, Jean Racine and Pierre Corneille in France, and, to a certain extent, Goethe in Germany). Many critics (including the English critics from Sir Philip Sidney to Matthew Arnold) accepted those rules, although few English poets—the great exception was John Milton—welcomed them. A lesser influence was exercised by Aristotle's *Rhetoric* outside the field of systematic theory.

Scholasticism and Aristotle's detractors

Scholasticism, in these centuries, belonged somehow to the history of Aristotelianism. All over western and central Europe and even in Spanish America, it ensured that higher education remained within Aristotelian frames of mind. Remarkable work was produced by Scholastics in the field of commentaries and detailed interpretation; Pedro de Fonseca, the "Portuguese Aristotle," in the 16th century and Sylvester Maurus in Rome in the 17th are notable examples. Insofar as the different Scholasticisms were living and interesting philosophical movements, however, they had more to do with newer philosophies than with Aristotle. Martin Luther's rebellion against Rome involved a rebellion against Scholastic philosophy and its distorted Aristotelian structure, not against Aristotle. In fact, the newer, more genuine, humanistic Aristotle emerged again as the great master of philosophy, independent of theology, when Luther's follower, Philipp Melanchthon, undertook to reorganize for the reformed Germany the curriculum for higher education. Aristotle had again, as in the early 13th century in Paris, pride of place particularly in logic and ethics, although less, as then, in metaphysics and natural philosophy.

The anti-Aristotelianism of the 16th to 18th century touched only a small part of the real Aristotle. On one side it was a reaction against Scholasticism, as though this had faithfully represented Aristotle's own philosophy. Thus, he was wrongly accused of extreme formalism, ruthless use of syllogisms consisting of empty or irrelevant concepts, a multiplication of pseudo-real entities, and the application of "scientific" methods to facts or supposed facts vouched for only by faith. On the other side, the whole of Aristotle was condemned because of his unsatisfactory account of local movement, with its consequences in mechanics, dynamics, cosmology, and astronomy. His downfall in the 17th century was the result, above all, of his failure to create, in the 4th century BC, a language that allowed him to describe the forms of things and events (*i.e.*, their knowable aspects) in precise or approximate mathematical formulas and of his failure to lay sufficient stress, in his philosophy of experience, on the need for experiments.

The 19th and 20th centuries. The anti-Aristotelian storm was weathered mainly by historical and philological scholarship. As Friedrich Adolf Trendelenburg, a German philosopher, saw it, Aristotle's personality and works must be known as exactly as possible because he is the indispensable historical basis of any serious philosophy. That type of study had declined after the great achievements of the 16th century. After the work done between the first new learned edition of the collected Greek texts of Aristotle by J.G. Buhle (1791 ff.) and a vast collection of all documentary material in the Aristoteles-Archiv at Berlin (1965 ff.), little, if anything, remains to be discovered concerning the original and deteriorated forms of Aristotle's traditional corpus. A monumental edition sponsored by the Prussian Academy from 1835 to 1870 was the basis of almost innumerable critical editions of individual works. A rich crop of fragments, identified and edited in the last centuries, including a large section of the *Constitution of Athens*, brought to light a previously almost unknown aspect of Aristotle's early activity: that it was principally Platonic. European and American academies have sponsored the editing of ancient and mediaeval commentaries and translations in Greek, Latin, Arabic, and Hebrew. Historical, philological, and philosophical exegesis has explored in great detail the contents and background of most of Aristotle's writings. Translations of all the works into English, German, and French and of many of them into most European languages and into Hebrew, Arabic, and Japanese have made Aristotle widely accessible. Historians of ideas have investigated Aristotle's relationship to Plato and to his contemporary Greece, his influence in following ages, and his own philosophical development.

Modern critical editions, translations, and studies

Philosophical Aristotelianism has been mainly confined to the German schools of Trendelenburg and Franz Brentano. Trendelenburg was concerned with a revaluation of Aristotle's metaphysics in the face of German Idealism; he had some influence in the United States on such thinkers as Felix Adler, George Sylvester Morris, and John Dewey. Franz Brentano's interpretation of Aristotle's theories of being and knowledge formed the basis of his "descriptive psychology" and his doctrine of human experience and contributed to Edmund Husserl's Phenomenology. Outside Germany, J.-G.-F.-L. Ravaisson-Mollien, a spiritualist philosopher, and William David Ross, editor of Aristotle's works, would acknowledge a debt to Aristotle, respectively, for their metaphysics and ethics; and the re-establishment of Aquinas, by Pope Leo XIII in 1879, as the great doctor of the church increased the interest in Aristotle as the "philosophical doctor" and his influence in the history of Christian thought. Contemporary philosophy in the Anglo-Saxon world is often associated with a keen interest in Aristotle: study groups are flourishing at Oxford, Harvard, and New York; and the name of the Aristotelian Society (London) reflects the view that good philosophy must be practiced in the spirit of Aristotle.

ASSESSMENT OF ARISTOTELIANISM

An assessment of the Aristotelian component of civilization would require an analysis, in extension and depth, of man's intellectual, and in many cases practical, approach in many fields of science, politics, morality, and art—both at the specialized and common, unsophisticated levels. Indications in that direction can be found in certain words that are indispensable for the articulate communication of thoughts, experiences, and problems. Some words still carry their Greek form, whereas others became established in their more important meanings as Latin equivalents of Aristotle's own words. The centuries-long impact of Aristotelian schooling on the intellectual dominant class lies at the root of the establishment of this vocabulary: "subject" and "predicate" in grammar and logic; "form" (information, transform) and "matter" as expressing the two correlative aspects of something that has acquired or acquires something else that is possibly essential to it; "energy" as the active power inherent in a thing; "potential" for what is latent but can be released; "substance" and "essence," "quantity" and "quality," "accidental," "relation," "cause" (and the many meanings of "because" corresponding to the four causes), "genus" and "species" (general, special), "individual," "-isible" (atomic)—these are a small sample of terms that still carry the mark of Aristotle's philosophy.

BIBLIOGRAPHY

Aristotelianism as covered in general histories: No extensive work of any value on Aristotelianism exists, apart from what is included in the fundamental history of philosophy by FRIEDRICH UEBERWEG, *Grundriss der Geschichte der Philosophie*, 11th–12th ed., 5 vol. (1923–28; Eng. trans. of 4th German ed., *A History of Philosophy, from Thales to the Present Time*, 2nd ed., 2 vol., 1875–76), with a vast bibliography. Useful histories of philosophy, general or partial are FREDERICK C. COPLESTON, *A History of Philosophy*, 8 vol. (1946–66); M.H. CARRE, *Phases of Thought in England* (1949), which is particularly good on Aristotelianism; J.H. RANDALL, *The Career of Philosophy*, 2 vol. (1962–65), imaginative and stimulating; ETIENNE GILSON, *History of Christian Philosophy in the Middle Ages* (1955), a personal interpretation with documentation and bibliography; MAURICE DE WULF, *Histoire de la philosophie médiévale*, 6th ed., 3 vol. (1934–47; Eng. trans. of vol. 1–2, *History of Mediaeval Philosophy*, 1952).

Aristotelianism in various periods or cultures: INGEMAR DURING, "Von Aristoteles bis Leibniz: Einige Hauptlinien in der Geschichte des Aristotelismus," *Antike und Abendland*, 4:118–154 (1954), mostly on Greek and medieval Aristotelianism; LORENZO MINIO-PALUELLO, *Opuscula: The Latin Aristotle* (1972), a collection of articles and essays concerning the Latin transmission of Aristotle's works; RICHARD MCKEON, "Aristotelianism in Western Christianity," in JOHN THOMAS MCNEILL, MATTHEW SPINKA, and HAROLD R. WILLOUGHBY (eds.), *Environmental Factors in Christian History* (1939). (*On Greek Aristotelianism*): EDUARD ZELLER, *Die Philosophie der Griechen*, 4th ed., vol. 2, pt. 2 and 5th ed., vol. 3, pt. 1–2 (1879–1920; Eng. trans. of early ed., *Plato and the Earlier Academy*, 1876–92), fundamental for the first eight centuries; PAUL MORAUX, *D'Aristote à Bessarion* (1970); FRITZ R. WEHRLI (ed.), *Der Peripatos in vorchristlicher Zeit* (1959); K. OEHLER, "Aristotle in Byzantium," in *Greek, Roman, and Byzantine Studies*, vol. 5 (1964); BASILE TATAKIS, *La Philosophie Byzantine*, (1949), an extensive survey, with rich bibliography. (*On Latin Aristotelianism*): Fundamental for pre-Renaissance are GEORGES LACOMBE *et al.*, *Aristoteles Latinus* (codices, 3 vol. and texts, 15 vol.; 1939–72, continuing); FERNAND VAN STEENBERGHEN, *Aristote en occident: les origines de l'Aristotélisme parisien* (1946; Eng. trans., *Aristotle in the West: The Origins of Latin Aristotelianism*, 1955), a scholarly survey of recent studies; RICHARD J. LEMAY, *Abu Ma'shar and Latin Aristotelianism in the Twelfth Century* (1962), important new contributions; D.A. CALLUS, "Introduction of Aristotelian Learning to Oxford," *Proceedings of the British Academy*, pp. 229–281 (1943), original, fundamental research; PAUL MORAUX *et. al.*, *Aristote et Saint Thomas d'Aquin* (1957), which includes some of the most reliable studies on the subject; M.D. CHENU, *La Théologie comme science au XIIIe siècle*, 3rd ed. rev. (1957), interplay of Aristotelian methodology and dogma; HASTINGS RASHDALL, *The Universities of Europe in the Middle Ages*, new ed., 3 vol. (1936), basic for Aristotelianism in the schools. (*On Syriac, Arabic, Jewish Aristotelianism*): ANTON BAUMSTARK, *Geschichte der syrischen Literatur* (1922), exhaustive factual information and bibliography, and *Aristoteles bei den Syrern* (1900), specialized research and texts; T.J. DE BOER, *Geschichte der Philosophie im Islam* (1901; Eng. trans., *The History of Philosophy in Islam*, 1903, reprinted 1970); CARL BROCKELMAN, *Geschichte der arabischen Literatur*, 2 vol. (1937–49), exhaustive factual information and bibliography; F.E. PETERS, *Aristoteles Arabus: The Oriental* [Syriac and Arabic] *Translations and Commentaries of the Aristotelian Corpus* (1968), to be supplemented with articles by R.R. WALZER *et al.*, in the *Encyclopaedia of Islam*, new ed. (1960–); ISAAC HUSIK, *A History of Mediaeval Jewish Philosophy* (1941, reprinted 1969); GEORGES VAJDA, *Introduction à la pensée juive du Moyen Âge* (1947), limited in scope, with good bibliography; H.A. WOLFSON, "Revised Plan for the Publication of a *Corpus Commentariorum Averrois in Aristotelem*," *Speculum*, 38:88–104 (1963), complete lists of Arabic, Latin, and Hebrew texts of Averroës' commentaries; and *Crescas' Critique of Aristotle* (1929); relevant articles in the new *Encyclopaedia Judaica* (in English), 16 vol. (1971). (*On Renaissance and later Aristotelianism*): PAUL OSKAR KRISTELLER, *Renaissance Philosophy and the Mediaeval Tradition* (1966), a brilliant survey, with bibliography; and *Studies in Renaissance Thought and Letters* (1956), many relevant essays; BRUNO NARDI, *Saggi sull'Aristotelismo padovano dal secolo XIV al XVI* (1958), one of several fundamental works by this author; PETER PETERSEN, *Geschichte der aristotelischen Philosophie in protestantischen Deutschland* (1921) and *Die Philosophie Friedrich Adolf Trendelenburgs: Ein Beitrag zur Geschichte des Aristotelismus im 19. Jahrhundert* (1913).

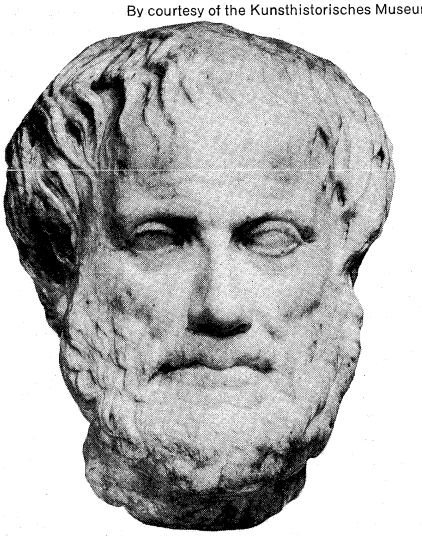
Aristotelianism in various areas or disciplines: (*On logic*): WILLIAM and MARTHA KNEALE, *The Development of Logic* (1962), an objective assessment of the Aristotelian and non-Aristotelian elements in the history of logic; I.M. BOCHENSKI, *Formale Logik* (1956; Eng. trans., *A History of Formal Logic*, 2nd ed., 1970), technical, with much bibliography. (*On science*): GEORGE SARTON, *Introduction to the History of Science*, 3 vol. (1927–48), fundamental, with extensive information and bibliography; RENE TATON (ed.), *Histoire générale des sciences*, 4 vol. (1957–64; Eng. trans., *A General History of the Sciences*, 4 vol., 1964–66); A.C. CROMBIE, *Robert Grosseteste and the Origins of Experimental Science*, 1100–1700 (1953), which upholds the view of Aristotelian impact on experimental method; ANNELIESE MAIER, *Studien zur Naturphilosophie der Spätscholastik*, 5 vol. (1949–58), fundamental research; ALEXANDRE KOYRE, *Études galiléennes*, 3 vol. (1939), indispensable for a proper evaluation of anti-Aristotelianism. (*On politics*): G.H. SABINE, *A History of Political Theory*, 3rd ed. (1963); ALESSANDRO PASSERIN D'ENTREVES, *The Mediaeval Contribution to Political Thought* (1939); GEORGES DE LAGARDE, *La Naissance de l'esprit laïque au déclin du Moyen Âge*, 3 vol. (1956–70), fundamental for the 14th century; HORST DREITZEL, *Protestantischer Aristotelismus und absoluter Staat: Die "Politica" des Henning Arnisaeus* (1970), excellent, with extensive bibliography on German Aristotelianism. (*Poetics and rhetoric*): BERNARD WEINBERG, *A History of Literary Criticism in the Italian Renaissance*, 2 vol. (1961), contains good surveys concerning Aristotle; LANE COOPER, *The Poetics of Aristotle, its meaning and influence* (1924); M.T. HERRICK, *The Fusion of Horatian and Aristotelian Literary Criticism, 1531–1555* (1946), and *The Poetics of Aristotle in England* (1930), indispensable complements to Cooper's book; C.S. BALDWIN, *Renaissance Literary Theory and Practice* (1939), useful for both poetics and rhetoric.

(L.M.-P.)

Aristotle

Aristotle, the most eminent representative of the Realist school of philosophy in the history of Western thought, is an essential source for what the Greco-Roman tradition has termed wisdom. A Greek philosopher, logician, and scientist, he, perhaps more than any other thinker, has characterized the orientation and content of all that is termed Western civilization. Aristotle was the author of an intellectual system that, through the centuries, became the support and vehicle for both medieval Christian and Islāmic Scholastic philosophy. His chief accomplishment, however, has been the establishment of the basis for logical thought; *i.e.*, that collection of methodological rules that enables discourse (*logos*) to achieve the greatest coherence and consequently the highest degree of effectiveness. An organizer as well as a classifier, he identified the categories that structure man's language and thought. More than his chief mentor, Plato, he understood the need to define language and to use it as a creative instrument (*organon*) of thought capable of an ordered structure and, therefore, of conferring its own inner law upon nature. An encyclopaedic thinker, he was able, on the one hand, to recognize the distinct kinds of knowledge, making contributions to specific disciplines, and on the other hand to perceive the overall unity of discourse that the forms of knowledge generate. The effect was to lay the groundwork for the evolution of empirical science and of speculative thought. Despite the limitations and imperfections now recognized in this achievement—*e.g.*, not until the turn of the 20th century could the development of mathematics have made comprehensive modal and symbolic logic possible—Aristotle's thought has molded the subsequent systematization of knowledge. Although the abstract area of his thinking, that of metaphysics, has not always been well understood, Aristotle nonetheless has shown the ultimate impossibility of conceptually unifying all of being, thus exposing the incomplete nature of every intellectual synthesis. He also exposed the qualified nature of any system of thought and the irreducibility of thought and existence to purely scientific technique with its analysis of the world of objects.

By courtesy of the Kunsthistorisches Museum, Vienna



Aristotle, marble bust with a restored nose, Roman copy of a Greek original, last quarter of the 4th century BC. In the Kunsthistorisches Museum, Vienna.

LIFE

Early years and background

Aristotle was born in the summer of 384 BC in the small Greek colony of Stagirus, on the Chalcidic peninsula of Macedonia, northern Greece. His father, Nicomachus, was court physician to Amyntas III, king of Macedonia, father of Philip II, and grandfather to Alexander the Great. As a doctor's son, he was heir to a scientific tradition about 200 years old. Through case histories contained in the *Epidēmiai* ("Epidemics") of Hippocrates, the father of medical science, Aristotle was introduced to Greek medicine and biology at an early age. As a physician, Nicomachus was a member of the guild of the Asclepiads, the so-called sons of Asclepius, the legendary founder and god of medicine. Medicine was traditional in certain families, being handed down from father to son by apprenticeship rather than by book learning. Thus, in all likelihood Aristotle learned the fundamentals of that practical skill he was afterward to display in his biological researches. Were he such an apprentice, he would have undergone a rigorous and varied training: he would have studied the role in therapy of diet, drugs, and exercise; he would have learned how to check the flow of blood, apply bandages, fit splints to broken limbs, reduce dislocations, and make poultices of flour, oil, and wine. Such, at least, were the skills of the trained physician. It is not known for sure that Aristotle acquired these skills; but it is known that medicine was studied in the Lyceum, Aristotle's own institute in Athens, and that later, in a distinctly snobbish vein, he considered a gentleman sufficiently educated if he knew medical theory but not its practice.

This early connection with medicine and with the rough-living Macedonian court largely explains both the predominantly biological cast of Aristotle's philosophical thought and the intense dislike of princes and courts to which he more than once gave expression.

First period: in the Academy at Athens. While Aristotle was still a youth his father died, and the young man became a ward of Proxenus, probably a relative of his father. He was sent to the Athenian Academy of Plato in 367 and there engaged in dialogue for 20 years. These years were the first of three main periods comprising Aristotle's intellectual development, years dominated by the formative influence of Socrates and Plato, who, together with Aristotle himself, were destined to form a compact intellectual dynasty.

Death of Plato: Aristotle's departure from Athens

On Plato's death in 348/347, his nephew Speusippus was named as head of the Academy. Aristotle then left Athens —in disgust, it is sometimes claimed—at not being appointed Plato's successor. This view of his motive, however, appears to lack foundation, for evidence suggests that he was ineligible to be the school's head because of his status as a resident alien who could not hold property legally. Studies have also shown that his departure from Athens was linked with an anti-Macedonian feeling that broke out in Athens after Philip had sacked the Greek city-state of Olynthus in 348. Nevertheless, Aristotle's 12-year absence from Athens indicates that he valued more the circle of friends who accompanied him on his travels—chief among them Theophrastus of Eresus, his pupil, colleague, and eventual successor as head of his Lyceum—than he did his relation with the Platonic Academy.

Second period: his travels. With him went another academician of note, Xenocrates of Chalcedon, whose lethargy Plato is reported to have contrasted with Aristotle's more energetic manner: "The one needs a spur, the other a bridle. . . . See what an ass I am training to compete with what a horse." The distinctive characters of the two men seem to have integrated well in establishing a new academy on the Asian side of the Aegean at the newly built town of Assus.

Directing philosophic circles at Assus and Mytilene. At Assus Hermeias of Atarneus, a Greek soldier of fortune, had acquired fiscal, then political, control of the Troad, northwest Asia Minor, as a vassal to his Persian overlords. After a visit to the Athenian Academy, he invited two of Plato's graduates to set up a small branch to help spread Greek rule and Greek philosophy to Asian soil. It was to this new intellectual centre that Aristotle came. There he assisted in the education of a philosopher-king, a task to be repeated later on a larger scale in Macedonia with the future Alexander the Great. To this period belong the first 12 chapters of Book 7 of Aristotle's treatise *Politica*. There he sketched the connection between philosophy and politics, viz., that the highest purpose of a city-state (*polis*) is to secure the conditions in which a minority who are capable of it can live the

philosophic life. Such a life, however, lies only within the capacity of the Greeks, whose racial superiority has qualified them to employ the non-Greek tribal peoples as serfs or slaves for the performance of all menial labour. Thus, citizenship and service in the armed forces were to be the exclusive rights of the Greeks. Nonetheless, Aristotle's espousal of an enlightened oligarchy actually constituted an advance over the political concepts flourishing at the time and should be viewed in its context as a positive development in the establishment of the noble civilization created by the Greeks.

At about the same time, Aristotle composed a work, *De regimine principum* (*On Kingship*), in which he clearly distinguished the function of the philosopher from that of the king. He altered Plato's dictum—for the better, it is said—by teaching that

> It was not merely unnecessary for a king to be a philosopher, but even a disadvantage. Rather a king should take the advice of true philosophers. Then he would fill his reign with good deeds, not with good words.

Aristotle thus strove to assure the independent role of the philosopher.

First marriage: familial contentment

On good terms with his patron, Hermeias, Aristotle was soon attracted to the monarch's adopted daughter or niece, Pythias, whom he married. She bore him a daughter, whom he called by her mother's name. In the *Politica*, cited above, Aristotle prescribed the ideal ages for marriage—37 for the man and 18 for the wife. Because Aristotle was himself 37 at this time, it may be that Pythias also approximated the marital age. It is also possible that their own marital relations are reflected in his further observation: "As for adultery, let it be held disgraceful for any man or woman to be found in any way unfaithful once they are married and call each other husband and wife." In his will Aristotle ordered that "Wherever they bury me, there the bones of Pythias shall be laid, in accordance with her own instructions."

Pythias did not live long, however; and after her death Aristotle chose another companion, Herpyllis, by whom he had a son, Nicomachus. She outlived him, and he made ample and considerate provision for her in his will "in recognition of the steady affection she has shown me." Generally, it appears that this early period was one of great personal happiness for Aristotle.

After three years at the young Assus Academy, Aristotle moved to the nearby island of Lesbos and settled in Mytilene, the capital city. With his friend Theophrastus, a native of that island, he established a philosophical circle patterned after the Academy. With the Earth and the sea then beginning to claim the greater share of his attention, he turned to biology, in which he undertook pioneering investigations. The landlocked lagoon of Pyrrha in the centre of Lesbos has been identified as one of his favourite haunts.

A marked shift of emphasis from astronomy and other theoretical, speculative modes of thinking to the inductive forms of natural observation thus arose in Aristotle's work. He even felt it necessary to justify this new attention to biology by dismissing the arguments that had classed it as an inferior, unattractive study. In his biological researches he focussed on a new type of causation, the intentional aim or end of nature, distinct from the mechanical causation also operative in the inorganic sphere. He noted that, if one burns an acorn, he destroys it in a mechanical way but that, if he gives it a chance, it turns itself into an oak. To admit the existence of such ends or aims in nature is to argue teleologically (Greek *telos*, "an end") or to admit the idea of a final cause (Latin *finis*, "end"). Though criticized for tending to substitute teleology for observation, Aristotle saw it as a valuable concept for coordinating the results of his researches. Thus, confessing his ignorance of the mode of generation of bees, he wrote in his treatise *De generatione animalium* (*On the Generation of Animals*):

> The facts have not yet been sufficiently established. If ever they are, then credit must be given to observation rather than to theories, and to theories only insofar as they are confirmed by the observed facts.

Associated with his researches into plant and animal life were his reflections on the relation of the soul to the body. As revealed by his tract *De anima* (*On the Soul*), he then began gradually to detach himself from the Platonic conception of the soul as an independently existing being temporarily resident in the body. With greater emphasis on the positive value of material existence, he suggested that the soul—analogously including all organic being—is the vital principle essentially united with the body to form the individual person. With some acknowledgment to Plato, he then defined the soul as the Form of the body and the body as the matter of the soul.

Psychological thought: reaction to Plato

Tutoring Alexander the Great. In late 343 or early 342, Aristotle, at about the age of 42, was invited by Philip II of Macedon to his capital at Pella to tutor his 13-year-old son Alexander. As the leading intellectual figure in Greece, Aristotle was commissioned to prepare Alexander for his future role as the military leader of the now-united Greek world against the Persian Empire. Using the model of the epic Greek hero, as in Homer's *Iliad*, Aristotle attempted to form Alexander as an embodiment of the classical valour of an Ajax or Achilles enlightened by the latest achievement of Greek civilization, philosophy. With his fixed conviction of the superiority of Greeks over Asians, he instructed Alexander to dominate the barbarians—*i.e.*, non-Greeks—and to hold them in servility by refraining from any physical intermixture with them. Despite this advice, however, Alexander later espoused intermarriage; he chose a wife from the Persian nobility and forced his high-ranking officers (and encouraged his troops) to do the same.

In other ways the influence that Aristotle had on Alexander was negligible. Although later, on his return to Athens, Aristotle enjoyed considerable political and economic support from the Macedonians and received assistance in the organization of his biological researches, it is not likely—as some have held—that Alexander collected and dispatched to Aristotle specimens of rare animals from Persia and India; in fact, Alexander's first penetration of the valley of the Indus did not occur until 328/327, less than six years before Aristotle's death. Indeed, the relation between the two was embittered by the execution of Aristotle's nephew, the historian Callisthenes of Olynthus, on a charge of treason while accompanying Alexander to Persia early in 328 to chronicle the campaign. It has even been reported that Alexander meditated revenge on Aristotle himself as a blood relative of the victim but was diverted by his preoccupation with the invasion of India. Clearly, in matters of political ideology, a gulf separated Aristotle and Alexander. Aristotle showed no awareness of the fundamental changes that Alexander's conquests were spawning in the Greek world; indeed, he was opposed in principle to Alexander's imperial policy that diminished the importance of the city-state. On the other hand, Alexander gratified his tutor by rebuilding the town of Stagirus, Aristotle's birthplace, which Philip II had destroyed earlier.

After three years at the Macedonian court, Aristotle withdrew to his paternal property at Stagirus (*c.* 339). There he continued the associations of his philosophical circle with Theophrastus and other old pupils of Plato.

Third period: founding and directing the Lyceum. Aristotle remained in Stagirus until 335, when, nearing 50 years of age, he returned to Athens and gave himself to purely scientific work. At this time the presidency of the Academy became vacant by the death of Speusippus. Possibly, Aristotle felt slighted when the school again passed him over and elected Xenocrates of Chalcedon, his old associate in biological research, as its new head. In any case, though he appears never to have wholly severed his relation with the Academy, he opened, in 335, a rival institution in the Lyceum, a gymnasium attached to the temple of Apollo Lyceus, situated in a grove just outside Athens. Frequented by other teachers—Plato mentions it as having been one of Socrates' haunts—the name came to be applied to Aristotle's school in particular. It was probably only after Aristotle's death that the school, under Theophrastus, acquired extensive property and achieved, like the Academy, the legal status of a *thiasos*, or religious association. From the fact that his

Return to academic world in Athens: the Lyceum

instruction was given in the *peripatos*, or covered walkway, of the gymnasium, the school has derived its name of Peripatetic. Informal as the school may have been under Aristotle, it was very important to him because, by coordinating the work of a number of scholars, he was able for the next 12 years to organize it as a centre for the prosecution of speculation and research in every department of inquiry and to compose numerous courses of lectures on scientific and philosophical questions. The chief difference in general character between the new school and the Academy was that, whereas the scientific interests of the Platonists centred in mathematics, the main contributions of the Lyceum lay in biology and history.

On the death of Alexander in 323, a brief but vigorous anti-Macedonian agitation broke out in Athens. Aristotle, from his Macedonian connections, naturally fell a victim, in spite of his lack of sympathy with the ideals of Philip and Alexander. Like Socrates, he was indicted on the capital charge of impiety, the pretext being that his poem on the execution of King Hermeias by the Persians, written 20 years before, was a virtual deification of his friend. The real offense, however, was political and lay in his connection with the Macedonian leader, Antipater. As condemnation was certain, the philosopher anticipated it by withdrawing with his disciples to Chalcis, now Khalkis, on the Porthmós Evrípou (Strait of Evripos) north of Athens. He died the following year from a stomach illness at the age of 62 or 63. It was reported that he abandoned Athens to save the Athenians from sinning twice against philosophy (referring to Socrates as the earlier victim).

PERSONALITY, CHARACTER, AND PHILOSOPHIC STANCE

The features of Aristotle, familiar from busts and engravings, are handsome, indicating refinement and acuteness. An ancient tradition, possibly from an unfriendly source, says, however, that Aristotle had spindleshanks and small eyes and spoke with a lisp. In compensation for these physical defects, he was notably well dressed. His cloak and sandals were of the best quality, and he sported rings. Presumably, he was rich, with large family holdings at Stagirus. One use that he made of his money was to support his habit of collecting books. Plato, with a touch of contempt for his devotion to reading and perhaps not without some envy of his affluence, called him "the mind." With equal justice he might have been called "the heart." If Aristotle strove for a fine balance between emotion and intellect, it seems that at times his feelings were the stronger. A story is told of Plato giving a reading of his *Phaedo*, a purported record of Socrates' last day. He had been condemned to death and was to drink the hemlock at sunset. His friends came to him in the morning, and he chose the immortality of the soul as an appropriate subject for their last day's talk. Though some of the argumentation is tedious, the dialogue, together with the narrative portions, is moving and solemn. As Plato read, however, his audience began to melt away. In the end, Aristotle alone was left. Probably fictitious, the anecdote was invented to enshrine a truth: Aristotle was, in fact, spellbound by the Socratic doctrine of immortality as expounded by Plato. It not only interested him intellectually but also absorbed him emotionally. His earliest works, dialogues written when he was a member of the Academy (now lost except for some fragments), were totally concerned with thoughts of the next world and the worthlessness of this one. These writings were in great vogue, as demonstrated by the numerous commentaries on them prepared in the centuries immediately following Aristotle's death.

Feelings about and interest in immortality

The anecdotes related of him reveal a kindly, affectionate character and show little trace of the self-importance that sometimes is detected in his works. His will, which has been preserved, exhibits the same traits in its references to his happy family life and its solicitous care for the future of his children and servants.

This personal happiness is reflected in his work *De philosophia* (*On Philosophy*), the last and greatest of his literary achievements, also surviving only in fragments. After he wrote this work, *c.* 348, he devoted his energies to research, teaching, and the writing of more technical treatises. The greatness of *De philosophia* lies in its influence on the thought of later antiquity; perhaps more than any other single work it established philosophy as a profession. In the extant part, *De philosophia* defines the specific role of the philosopher. Tracing the order in which the various elements of civilization originate as five main stages, Aristotle sees it as culminating in philosophy. First, men are compelled to devote themselves to the creation of the necessities without which they could not survive. Next come the arts that refine life, a stage followed by the discovery of the art of politics, the prerequisite of the good life as Aristotle conceived it. To these necessities and refinements of life is added the knowledge of their proper use in the fourth stage. With the emergence of the well-regulated state comes the leisure for philosophy, which at first occupies itself with the material causes of existing things, as the older thinkers did. Finally, comes the shift from natural to divine philosophy, when the mind lifts itself above the material world and grasps the formal and final causes of things, realizing the intelligible aspect of reality and the purpose that informs all change.

Concept of philosophy

This divine philosophy gave its attention to the new astral gods. Aristotle had experienced in Athens the long intellectual struggle to discover perfect order in the heavens. He had shared the relief of knowing that perfection was not to be confined to the mathematical abstractions, to which Plato had at first directed the attention of his pupils, but that the visible heavens themselves could be accepted as the embodiment of the divine. With the declaration of this intimacy between the deities and the work of their hands in the material universe, Aristotle proclaimed his manifesto, an optimistic affirmation of the values of this world; simultaneously, he rejected the Platonic doctrine of the imprisoned soul whose life purpose is to struggle free from the bonds of matter. By this stroke Aristotle established his own identity in the history of thought.

WRITINGS

Aristotle's writings fall into two groups: works published by Aristotle but now lost; and works not published by Aristotle, in fact not intended for publication but collected and preserved by others. In the first group are included (1) the writings that Aristotle himself termed "exoteric," or popular, works—that is, those written in dialogue or other current literary forms and meant for the general reading public—and (2) those that he termed "hypomnematic," or notes to aid the memory and collections of materials for further work. Of these, only fragments are extant because—as with many pre-Christian writings—the manuscripts, or rolls, were lost during the first centuries of the Christian Era. Finally, the group of writings that generally have survived, termed "acroamatic," or treatises (*logoi*, *methodoi*, *pragmateiai*), were meant for school use and were written in a concise and individualistic style. In later antiquity Aristotle's writings filled several hundred rolls; today the surviving block of 47 works attributed to Aristotle comprises two volumes in the Greek text of the standard Berlin edition of 1831. Three ancient catalogs list a total of more than 170 separate works by Aristotle, a figure corroborated by references and lists of titles in the extant treatises as well as by citations and paraphrases of early commentators. Cicero must have been alluding to Aristotle's popular dialogues when he described in the *Academica* "the suave style of Aristotle. . . . A river of gold," for such an impression can hardly be gained from the ponderous, clipped style of the tracts that still exist.

Lost works published by Aristotle. These lost popular works, which include surviving fragments of poetry and letters written by Aristotle, closely resembled the attractive Platonic dialogues. In the attempt to reconstitute their content, several problems confronted historical scholars. The lost dialogues, for example, appear to diverge widely from the doctrines of the surviving treatises. Indeed, they appear to outdo Plato in his own teaching. Thus, what is known of Aristotle's dialogue *Eudemus*, or

Popular dialogues

On the Soul, compares the relation of the soul to the body with an unnatural union, like that of the torture that the Tyrrhenian pirates inflicted on their prisoners by binding each of them to a corpse. Inasmuch as Aristotle in his extant treatises criticized his Platonist friends for alienating spirit from matter, Alexander of Aphrodisias, an authoritative Aristotelian commentator of the late 2nd century AD, raised the question whether he expressed "two truths," one "exoteric" for public consumption, the other "esoteric" and reserved for his students at the Lyceum. The present consensus of scholars is that Aristotle's popular writings generally derived from the early stage of his intellectual development during the first period of his work at Plato's Academy and are to be used as points of departure for discerning his subsequent development.

Chief among the lost works are: *Eudemus*, in the tradition of Plato's *Phaedo; De philosophia*, a type of philosophic program containing themes to be developed later in his *Metaphysica* (*Metaphysics*); the *Protrepticus*, or exhortation to the life of philosophy; *Gryllus*, or *On Rhetoric; De justitia* (*On Justice*), expressing nascent themes of his *Politica; De bono* (*On the Good*); and *De ideis* (*On Ideas*, or *On Form*).

Extant works assembled by Aristotelians. The 47 works that have been preserved comprise a mass of manuscripts, mostly, it seems, notes used in giving Lyceum courses. These are the "esoteric" writings of a concentrated, academic nature intended for the ears of the initiates. From classical antiquity romanticized accounts circulated of the way these manuscripts were preserved; *e.g.*, in Plutarch's *Sulla*, *c.* 26; and in Strabo's *Geography* 13:54. According to these versions, Aristotle's and Theophrastus' notes had been bequeathed to an old colleague, Neleus of Scepsis, whose heirs apparently were not interested in the contents but, to prevent them from being confiscated for the library of the kings of Pergamum, hid them in a cellar in Scepsis. Long afterward, in the 1st century BC, the descendants sold them to Apellicon of Teos, a philosopher, who brought them back to Athens. When Athens was conquered by Sulla in 86 BC, he appropriated the books and sent them to Rome, where they were purchased by Tyrannion the grammarian. The manuscripts suffered further maltreatment at the hands of copyists, subjective restoration of worm-eaten passages and interpretative, systematic ordering irrespective of chronology, until Andronicus of Rhodes, the last head of the Lyceum, acquired the copies and edited and published them *c.* 60 BC.

The story is somewhat improbable. It is difficult to imagine that the Lyceum, which lasted for more than 250 years after Aristotle, would have allowed the manuscripts of its founder to have been so carelessly looked after. Nevertheless, it is true that the Andronicus edition is the first publication of Aristotle's works, even if the story of its appearance was spread by Andronicus to emphasize its novelty. The form, titles, and order of Aristotle's texts that are studied today were given to them by Andronicus almost three centuries after the philosopher's death; and the long history of the wealth of commentaries they generated began at this stage.

Formation of existing works

These facts have affected the interpretation of Aristotle. In effect, the books of Aristotle that are known today were never edited by him. Thus, for example, Aristotle is not the author of the work called *Metaphysica;* rather, he wrote a dozen little treatises: on the theory of causes in the history of philosophy, on the chief philosophical problems, on the multiplicity of meanings, on act and potency, on being and essence, and on God. Those that the editors thought worth collecting were given the arbitrary title *Metaphysica; i.e.*, the tract that is to be read after the *Physica*. It is not surprising, then, that the *Metaphysica* and the other works of Aristotle appear as independent groupings of studies lacking any clear progression, repetitious and at times even contradictory. The texts suggest that students or subsequent members of the Lyceum even revised Aristotle's expressions. It is probable that Aristotle would never have released the work. Andronicus, assisted by previous editors, imposed a logical and didactic order upon all the writings, undoubtedly influenced by Aristotle's emphasis on logic as the propaedeutic (preparatory study) of all understanding. By ignoring the chronological order of the treatises and by grouping dissertations from different periods under the same title, the editors fashioned the Aristotelian corpus into an impersonal whole that loses its relationship to the living Aristotle. Because of the extrinsic circumstance of a plan of editing, the systematic character attributed to Aristotle by interpreters was born. Consequently, any interpretation of Aristotle must take into account the didactic finality injected into the corpus of his published works and contrast it with the specific manner of Aristotle's teaching. In the Socratic tradition the dialogue took precedence over the monologue, with the written text reflecting the interpersonal oral experience. If the text as composed lacked the critical responses of other minds, the author was obliged to parry with the opinions of other philosophers, past and present. Aristotle's treatises, rather than exhibiting a mélange of conflicting views on any question, reveal the philosopher at work, defining the problem he is to deal with, assessing the views of his predecessors, formulating his own preliminary opinion, considering whether this needs modifying in the light of difficulties and objections, rehearsing the arguments for different points of view—always searching, in short, for the most adequate solution or resolution of his problem. The reader, therefore, participates in the work of Aristotle, not dogmatically propounding a doctrine but often laboriously developing a perspective or insight that emerges from difficulties, contradictions, and paradoxes. Not surprisingly, few syllogisms appear in Aristotle's treatises; the reader should perceive them in a structure that Aristotle himself terms "dialectical"; *i.e.*, in the manner of a dialogue by an exchange of arguments for and against.

"Dialectical" nature of Aristotle's thought

THEORIES OF THE EVOLUTION OF ARISTOTLE'S THOUGHT

From the conclusions of Alexander of Aphrodisias in the latter part of the 2nd century, a distinction was established between the doctrine expressed in Aristotle's treatises (the technical writings that have come down to the present) and the popular Platonizing dialogues (the "exoteric" works surviving only in fragments). The orthodox view for 17 centuries was that the treatises were the sources for Aristotle's genuine thought; Valentin Rose, the 19th-century German editor, proposed that all of the lost dialogues are spurious because of their inconsistency of doctrine in relation to the treatises. The underlying assumption was that a man of such strict and systematic mind as Aristotle would maintain a strict constancy and never abandon opinions once formed.

Jaeger's developmental theory and rejoinders. In the first half of the 20th century, a developmental theory of Aristotle's thought was submitted by Thomas Case, an English philosophical scholar, and elaborated in detail by Werner Jaeger, a German historian of Greek philosophy, in his *Studien zur Entstehungsgeschichte der Metaphysik des Aristoteles* (1912; "Studies in the History of the Origin of Aristotle's Metaphysics") and later in his *Aristoteles: Grundlegung einer Geschichte seiner Entwicklung* (1923; Eng. trans., *Aristotle: Fundamentals of the History of His Development*, 2nd ed. 1948). Employing a historico-genetic methodology, Jaeger announced that the greater part of Aristotle's lost works represented his thought while he was still at the Academy under the immediate influence of Plato: the preponderance of such themes as the immortality of the soul, disdain for the material world, the doctrine of the reminiscence of Ideas, the heroicness of wisdom, asceticism, and the existence of God were Platonic. These dialogues addressed a wide audience and manifested an elegant literary style that fascinated classical authors. According to Jaeger, Aristotle gradually receded from Plato's position, even during his time at the Academy—rejecting certain Platonic arguments, adopting contrary positions, continually evolving from Platonic Idealism to a marked Empiricism.

Theory of Aristotle's early Platonism

In response to this evolutionary theory, critics have noted that the methodological analysis of Aristotle's works generates complex problems, as observed above in the account of the formation of the text of his existing

writings. The works edited by Andronicus of Rhodes are compilations of texts from differing periods. Thus, the *Metaphysica* covers almost the entire career of Aristotle, as does the *Politica*. Often within the same chapter, even in a single paragraph, one discerns strata representing different stages of Aristotle's thought: his early phase at the Academy, his subsequent maturity, his period of travel away from Athens, and his Lyceum experience, during which there purportedly existed even another distinction between his morning sessions with his best students and afternoon meetings with a wider audience from Athens. From these discrepancies one could theoretically construct an Aristotelian *Metaphysica I* and *Metaphysica II*, a first and second *Ethica*, as well as two *Physica*, two distinct tracts *De anima*, and a double treatment on *Politica*.

Because Aristotle's treatises lack the positive data of a Platonic dialogue, with its historical allusions and stylistic criteria, Jaeger's ingenious thesis assumes that Aristotle's concurrence with any Platonic tenet must have happened only during his early years at the Academy or, at the latest, by the close of the first Athenian period (347 BC). The argument suggests that he never could have sustained contradictory theses simultaneously. Hence, any distinction in doctrine represents a succession in time.

Certain scholars point out the arbitrary character of such a thesis. One could conceivably defend the converse: that Aristotle, in a self-assured youth, could have strongly challenged Plato but, having subsequently become conscious of the more profound significance of his master's philosophical postulates, did not hesitate to integrate with his own thought one or more Platonic theses. Indeed, in the logical treatise *Topica* (*Topics*), considered one of Aristotle's early works because it reflects discussions at the Academy, and in the *Ethica Eudemia* (*Eudemian Ethics*), the first version of Aristotle's course on ethics, there appears one of the strongest of anti-Platonic theses, the identification of being and the Good.

Evolution of thought on the soul

Jaeger's theory admittedly achieves near certitude in relation to Aristotle's psychology, as demonstrated by a study by François Nuyens, a Dutch historian of philosophy, in 1939. He held that in his early period, represented by the *Eudemus* and the *Protrepticus*, Aristotle began as a Platonist, describing the soul as a separate substance in an unnatural relationship with the body; next, in an intermediate stage, he described the body as the instrument of the soul, the function of which is analogous to that of a pilot steering a ship; finally, as expressed in the tract *De anima*, he advanced more clearly the concept of a substantial unity of body and soul in making the soul the Form (an intrinsic principle of operation) of the body. Some of Aristotle's followers advanced this development even one step further and submitted that the soul itself comprises a material nature.

Other authorities on Aristotle have observed that this instance of a discernible linear transition in his thought is of a type occurring rarely among his writings. Moreover, even in the works on psychology, Aristotle's concern for metaphysical thought is not wanting in his intermediate biological phase or in the alleged fruition of his empirical stage during his work at the Lyceum. Such a simultaneous preoccupation with both inductive and abstract thought is further manifested in an analysis of the sequence of books in the *Metaphysica*.

At the other extreme is the hypothesis of the German scholar Josef Zürcher, in *Aristoteles Werke und Geist* (1952; "The Works and Spirit of Aristotle"), who asserted that Aristotle's own thought always remained Platonic and that all of the characteristically Peripatetic philosophy came from his disciple and successor, Theophrastus, who is the true author of about 75 percent of the existing Aristotelian treatises. According to this theory, there never existed a young Aristotle and an elder Aristotle relative to any development of his thought; there was simply a Platonic Aristotle and an anti-Platonic Theophrastus, an Empiricist. In reaction to this interpretation is the suggestion made by Ingemar Düring, a Swedish classics scholar, in 1966, that Aristotle never really subscribed to the Platonic theory of transcendent forms but maintained lifelong coexisting interests in empirical data and metaphysical speculation; to Düring, any developmental theory of Aristotle's thought remains inconclusive.

Recent analyses of Aristotle's development and systematizing. Aristotelian scholars have generally concluded that a basis exists for a theory of evolution in his thought but that the determination of the chronology and the degree of change presents a difficult series of problems. It is quite possible to concur with Jaeger that during Aristotle's first years at the Academy he acknowledged Plato's teaching on subsisting Ideas. It is another matter, however, to suggest that in his later years he renounced such Platonically influenced doctrines as the immortality of the soul, a religious philosophy concluding in an ultimate being termed God, and a moral concept of life. Increased attention to data of the senses in subsequent phases of his life is not a sufficient argument for the emergence of an Empiricist Aristotle, who could only oppose a spiritualist and Idealist Plato. It is true that Aristotle later criticized the doctrine of the world of Ideas as inadequate and contradictory if related to a theory of knowledge that attempts to explain man's conceptual thought; but he continued, nevertheless, to recognize the effectiveness of metaphysical thought in arriving at the concept of a transcendent, nonmaterial, and subsistent intellect as the necessary explanation for the fact that anything exists. The consensus of modern commentators thus suggests that not every aspect of Platonic Idealism was annihilated by Aristotle's growing appreciation of empirical knowledge and of the dynamic aspects of matter. Rather, part and parcel with his experimental work in biology and physics was his continued teaching on the essential distinction between the phantasm (image) and the concept, thus confirming his position on the essentially spiritual nature of human intelligence, analogous to the Platonic view of the natural similarity between the soul and the Idea.

Thus, the attempts of certain interpreters to oppose Aristotle's empirical experience to his evidently contemplative trend toward reflection in terms of universal principles do not sufficiently attend to the systematizing aspect of Aristotle's thought. The inconsistencies, contrasts, and varying degrees of emphasis on different modes of thought throughout the Aristotelian corpus are not adequately explained either by positing intervening editors and copyists or simply by different stages in Aristotle's thought. He obviously attempted in all of the treatises to relate his own views to the whole history of thought before his time. On many occasions he was concerned, at the same point of time in the development of his thinking, to state different views seen as alternative possibilities. Often his method was deliberately aporetic, or skeptical; he raised difficulties that he knew had to be faced but for which he supplied no immediate or straightforward solution. Left by Plato with a vast body of problems, Aristotle conscientiously pursued an ideal of correcting and complementing the heritage of intellectual tradition bequeathed to him. Toward this end he often followed parallel but distinct paths of human investigation, not without apparent intrinsic contrarieties. His method was exploratory, and he used it on whatever fertile soil he was free to work. Only relatively late in life was he able to unify his results with varying degrees of success. The philosophy of Aristotle does not unfold simply by deducing consequences from assumed principles, nor does it seek to deduce a pluriform reality from an intuited unity; it was from the outset pluralist, and its unity was the singleness of purpose founded on inquiry. If his writings at times exhibit a disparate structure with gaps, it is because he experienced this beforehand in the nature of his thought.

Aporetic nature of Aristotle's thought

The evidence would seem to converge on an interpretation of Aristotle that, freed from the perspective imposed on thought by the later history of science and philosophy, sees him as searching for meanings against the background of the work of his predecessors. The effort to arrive at any generalized picture of his intellectual development should be accompanied simultaneously by a preoccupation with his systematizing approach and with the

successive steps that he took to arrive at his ultimate position. To date, however, the application of the historico-genetic method of textual analysis, while providing several valuable insights, has failed to formulate a universal expression of Aristotle's development that adequately responds to the complexity of the problem.

SYNOPSES OF THE ARISTOTELIAN CORPUS

Tracts on logic. The term logic does not derive from Aristotle but goes back to Xenocrates of the Academy. Aristotle, however, attributed more extensive significance to language (*logos*) and the rules of discourse; he emphasized that language alone is the specific distinctive of the human species. The significance of this intrinsic relationship affected even his definition of man as a rational animal, which in the Greek also means an animal possessing a language or speech or word.

In Aristotle's view, the purpose of language is to express the interior feelings and experience of the soul, and consequently words are signs or symbols of the various activities of the intelligence. He saw in terms, propositions, and syllogisms the reflection of concepts, judgments, and intellectual reasoning; and he considered that, in describing and handling these activities, he was not only being made aware of them but also being constrained to set down norms and methods for their proper and effective function.

The logical treatises of Aristotle make up the collection known as the *Organon* ("tool"), a title adopted by later commentators, who, in accordance with the well-established Peripatetic tradition, regarded logic as an instrument for doing philosophy. In Aristotle's preferred view logic was not included in the classification of the sciences at all but was treated as a preliminary to the study of each and every branch of knowledge. Aristotle's own name for logic was *analytika* (Greek: "to unravel"). The term logic, however, is found employed in a somewhat restricted sense in Aristotle's own writings; *e.g.*, in the *Topica* I, 14. And there is evidence that it was beginning to be used as the equivalent of dialectic or analytics almost immediately after Aristotle's death.

Contents of the *Organon*

The *Organon* contains the *Categoriae* (*Categories*), *De interpretatione* (*On Interpretation*), *Analytica priora* (*Prior Analytics*), *Analytica posteriora* (*Posterior Analytics*), *Topica*, and the *Sophistici elenchi* (*Sophistical Refutations*). The arrangement within the collection is meant to be systematic rather than chronological; it is not certain, however, whether the ordering is fully attributable to Andronicus of Rhodes or to Aristotle. The original chronological order can hardly be determined now with any certainty because Aristotle, or other editors, apparently used later insertions to supplement the original treatises. In a possible sequence of their composition, the *Categoriae*, *Topica*, and the *Sophistici elenchi* are listed earlier than *De interpretatione*, and this work, in turn, earlier than the *Analytica priora* and the *Analytica posteriora*. The chapters on modal logic in the *Analytica priora* are probably the last that Aristotle added to the body of the *Organon*. Outside of the *Organon*, the fourth book of the *Metaphysica* could be described as a logical work concerned with the principle of contradiction; it appears to have been written earlier than *De interpretatione*.

In the *Categoriae*, Aristotle distinguished expressions that exhibit propositional unity from expressions that do not; *i.e.*, the composite statement relating a subject to a predicate, as distinct from a simple term. This notion of propositional unity can perhaps be traced back to Plato, but the treatment of simple expressions was Aristotle's innovation. He considered simple expressions neither true nor false and held that they may signify one or another of the following categories: substance, quantity, quality, relation, place, time, position, state, action, and affection. It is by no means clear whether this classification is to be regarded as ontological (concerning the nature of reality) or as purely verbal—*i.e.*, whether it is about actual things or about words and expressions; the same ambiguity has been characteristic of practically every other scheme of categories suggested since his time.

As a part of a theory of reality, Aristotle later used the categories to criticize Plato's theory of Forms, or the world of subsisting Ideas. For Aristotle, Plato was involved in a confusion between the category of substance and the other categories when, for example, he attributed substantiality, or concrete existence, to predications of quantity, such as "being tall." But in chapter 5 of the *Categoriae*, Aristotle distinguished within the category of substance between "primary substance" and "secondary substance." Primary substances are particular men, horses, etc., and secondary substances are the species and genera to which the individuals belong. There Aristotle treated genus and species as substances and not as qualities. In the *Metaphysica*, however, he abandoned this position—*i.e.*, he definitely rejected the world of Ideas (theory of Forms) as inadequate for knowledge of the material world—in favour of the simpler contrast between substance and the universal concept, thereby indicating that the *Categoriae* was written at an early date and reflects to a degree a certain Platonic influence.

Anti-Platonic evolution of logic

The *Topica* appears to have been intended as a manual for participants in contests that involved argumentation. For the most part it consists of suggestions about how to look for an argument that will establish or refute a given thesis; thus, it elucidates general logical laws or rules.

The *Sophistici elenchi* exposes forms of reasoning that appear valid on the surface but actually conceal a delusion or deceit. Examples of fallacious arguments are "begging the question," or circular argument (*e.g.*, a statement that the soul continues to exist after death because it is immortal); the "fallacy of the consequent," or arguing from a consequent to its condition (*e.g.*, if a man is a drunkard he becomes destitute; therefore, if he is destitute, he is a drunkard); and the "fallacy of the irrelevant conclusion," wherein, instead of proving the fact in dispute, the arguer seeks to gain his point by diverting attention to some extraneous fact.

Importance of *De interpretatione*

The most important treatises of the *Organon* are *De interpretatione* and *Analytica priora*. Together they discuss various forms of logical statements and theories concerning their interrelationships. The significance of direct statements of fact is contrasted with the consequences of those containing expressions of condition or contingency (assertoric versus modal syllogisms). The meaning of affirmation and denial is also explored.

The *Analytica posteriora*, resembling a collection of notes, treats scientific method rather than logic proper.

Tracts on the philosophy of nature. In his treatise *Physica* Aristotle dealt with the natural body in general, or with all that is corporeal; special kinds of material bodies were discussed in his other physical works, such as *De caelo* (*On the Heavens*). The first book of the *Physica* is concerned with the intrinsic, constitutive elements of a natural body, those that he called "matter" and "form," or the substratum that perseveres in the phenomenon of change, and the active principle that defines the nature of the product of change. The second book treats mainly the different types of cause studied by the physicist, the material and the formal causes just mentioned, and the final and efficient causes, or the purpose and the responsible agent by which anything comes into being, respectively. Books 3 through 7 deal with movement, or motion, and the notions implied in it—such as space, position, and time, their magnitudes and continuity. The subject of Book 8 is the first mover, which, though not itself a natural body, is the cause of movement in natural bodies; its necessary attributes—immovability, eternalness, etc. —are also examined.

Whatever the virtues or defects, clarity or obscurity, of Aristotle's physical treatises, they assume that the distinction between physics and metaphysics (the "first philosophy," or the science of being as being) is valid. Although the conception of a continuous scale of nature from inorganic substances to biological and psychological phenomena is basic in all of his science, explanation does not consist in running uniformly up the hierarchy of beings to God nor in reducing all functioning to some material organ. The sciences of Aristotle are based on a multiple system of classification, not on a simple scheme of mutually exclusive and independently existent genera

and species. The very distinction of causes in existing and mutable things permits the differentiation of the subject matters of the natural sciences.

Natural science based on threefold classification

The *Physica* investigates the "natures" of things: elemental bodies and their compounds, plants and their parts, animals and their parts. These things are considered within the context of change and in terms of the internal principles of motion and rest that they possess. The treatment of the basic principles of natural science in the *Physica* depends on a triple classification: (1) A classification is posited according to "causes." Four causes are relevant to the problems of the physicist, although three of them—the formal, efficient, and final causes—frequently coincide, and physical explanation therefore consists usually of stating either the "end" (*i.e.*, the final cause) or the antecedent "necessity" of a process. (2) A classification is also derived according to "substance." Three kinds of substances determine the three branches of study in physics: one, of things that are incapable of motion, the second of things in motion but indestructible, the third of destructible things. (3) Finally, a classification is determined according to "change." Four categories of change are held to occur: generation and destruction in the category of substance and three kinds of motion—alteration in the category of quality, increase and diminution in the category of quantity, and locomotion in the category of place.

The tract culminates in the study of the first of the three "substances" with which natural science is concerned; *i.e.*, with the study of things incapable of motion, the unmoved mover, and the final cause of motion.

De caelo considers the universe as a whole in terms of its parts. It thus treats of bodies, conceived as three-dimensional magnitudes, and their motions in place, differentiating the eternal circular motion of heavenly bodies (planets and stars) from the straight-line motion of terrestrial or sublunary bodies, conceived as three-dimensional magnitudes. The author inferred from this distinction that the four elements of terrestrial bodies (air, water, earth, and fire) differ, because they are either heavy or light, from the "primary body" (later called the "fifth-body" or "quintessence") of which the heavens are formed, each orbit constituting a distinct sphere. The problems were formulated and treated in terms of the properties of simple bodies or elements—that is, of material causes. The finitude of all bodies was stated, although heavenly bodies are ungenerated and indestructible. Aristotle thus proposed a mechanical, geocentric model of the cosmic universe with the Earth at rest, broadly suggesting gravitational theory. The entire structure is more than strictly mechanical, however, because it is constituted of intrinsic, dynamic elements conferring impetus (physical and intellectual) and purpose.

The mutability of earthly, material bodies is the subject of *De generatione et corruptione* (*On Generation and Corruption*). Unlike heavenly bodies, these undergo changes (as the title, paraphrased "On Coming-to-be and Passing-away," suggests)—changes in size, quality, and even substance. The inadequacies of theories about corporeal change found in other schools of thought were analyzed, and the Aristotelian response was founded on the four causes of things.

The doctrine of causes was applied to the problem of the interrelation between the constitution of things in various degrees of structural complexity and their qualities and functions. Thus, in *Meteorologica* (*Meteorology*) Aristotle discussed the intermediate region of bodily things; *i.e.*, the region between the encompassing motions of the heavenly bodies and the least motions of material elements. These material phenomena include simple elements—such "homogeneous bodies" as metals, plants, and animals and such "heterogeneous bodies" as organs and parts of animals or humans.

The operation of causes within this bodily world was also investigated. Among the more complex organs, the ends or purposes of their operations are expressed by means of final causes. In the case of regular compounds the formal cause is applied, and among the simple elements the explanation is in terms of the material cause.

Biological problems, finally, are raised in the remaining intermediate region of zoology in the tracts on the natural world. Aristotle submitted that inquiry into biological phenomena depends on the one hand on a theory of life and the soul and on the other hand on the characteristics of organs compounded from homogeneous bodies, such as tissues and bones, and from the elements. The chief problems about which the elaborate observations and analyses of Aristotle's four biological treatises revolve are precisely the functions of life that were mentioned without further development of their characteristics in his psychological treatise *De anima:* generation (together with such related phenomena as heredity, sex, adaptation to environment, the struggle for existence), locomotion, growth, and nutrition. The orderly interrelations of the kinds and functions of animals are set forth, together with rules and exemplifications of the method adapted to their study. The diverse bases for the classification of biological phenomena are set forth in the *Historia animalium* (*History of Animals*), sometimes regarded as a monumental work, in which numerous sets of classifications of animals are positioned against equally numerous sets of classifications of their parts to indicate the various grades of likeness to be found among animals relative to their distinct parts and functions. *De partibus animalium* (*On the Parts of Animals*) shows the application of the method of physical inquiry to biological studies, reiterating the distinction between two kinds of explanation that might be sought: first, deductive reasoning, which is based whenever possible on the end or purpose of animal function; and, second, inductive or necessary conclusions derived by arguing to antecedents from observed consequents. The remainder of the work treats of the biological phenomena that can be explained in terms of the material parts and organs of animals.

Biological classifications according to structure and function

In *De incessu animalium* (*On the Progression of Animals*) the problems of local motion are considered by constructing a kind of animate mechanics of limbs according to their numbers and their kinds of joints. As he indicated at the beginning of *De generatione animalium*, (*On the Generation of Animals*), Aristotle recapitulated the four causes, remarking that all of them had thus far been used except the efficient cause; and he thereafter proceeded to treat of the problems of reproduction, to which the efficient cause is applicable.

Tracts on psychology. The relation between the active principle and the passive continuum (or between form and matter) that is operational in sentient and intellectual life was examined in the tract *De anima.* After exploring the concept and conditions of life, Aristotle related the function of matter and Form (body and soul) in human life to all of life's biological and psychological phenomena while rejecting Platonic transcendentalist and pre-Socratic Materialist theories on the nature of the soul. The complexities and universal conditions constituting birth, nutrition, and growth (plants, animals), sense experience and movement (animals, man), and thought (man, God) were viewed in the perspective of life's material and formal causes. The vital principle was identified with the soul (the formal and, in a special sense, also the efficient and final cause of living); and the various processes involved in sensation and thought were elaborated in relation to their proper objects; *e.g.*, sight is specified by visible objects and by the process of seeing. Aristotle maintained the essential immateriality of the soul and its intellectual powers by insisting on the distinctions between thought, perception, and imagination.

Psychology of thought and perception

In the *Parva naturalia,* the medieval designation for a collection of short treatises on natural functions, the formal analysis of some of the processes of life as exemplified in human beings is supplemented by specific inquiry into experiences of sensation and imagination in animals. In this work Aristotle investigated the related material components, the common characteristics of animal bodies, their comparable organs, and the similar objects of their activities in questions that arose relative to senses and sensibles, memory and reminiscence, the length and brevity of life, youth and old age, life and death, and respiration.

Tracts on metaphysics. The study of metaphysics, the function and content of which have generated neither conviction nor consensus of opinion on the scope of its subject matter, is—together with the logical syllogism and the differentiation of kinds of premises—an innovation of Aristotle and is therefore conditioned by the methodological instruments of his logic. In his *Metaphysica* the doctrines that Aristotle sometimes referred to as comprising "wisdom" and sometimes as comprising "first philosophy" or "theology" were developed. Its task is that of employing the systematic method delineated in the *Organon*, with which he prefaced all scientific inquiry, and applying its logical tools; thus, metaphysics is probably best approached in the perspective of his logical methodology. In this manner one avoids the spurious excursions that have appeared in the history of thought since Aristotle, which have extended beyond the limits of his theory and tended toward the transcendental pretensions and mystical connotations later associated with his name.

Critique of pre-Aristotelian theories of causes

Aristotle recounted the successive attempts in earlier thought to explain the various causes of existing things, of continuity, and of change. Indeed, his *Metaphysica* constitutes a principal source for pre-Socratic and Platonic philosophy. In it his predecessors' interpretations and uses of causal principles are delineated. During the period from Thales to Democritus, phenomena and things both had been explained by their constituent elements, which Aristotle called material principles and causes. He enumerated four causes to explain any thing or situation, although on any occasion only two or three may be accessible to investigation: the material cause, out of which a thing comes to be (as the silver of a bowl); the form or definition (as the shape of the bowl); the efficient cause (as the hammering of the silversmith); and the final cause (as the purpose for which the bowl is intended). Subsequent to the Atomists a number of schools (*e.g.*, the Pythagoreans) had recognized the formal cause, whereas others (*e.g.*, the Eleatics) had varied between the two, relying now on matter, now on form; and still others (*e.g.*, the Platonists) had used them both. But Aristotle considered his own singular contribution to be the discussion of efficient and final causes, and he found only vague, though ingenious, anticipations of the efficient cause in philosophers such as Empedocles and Anaxagoras, while the traces of inquiry into the final cause were even more tenuous.

After sketching the main problems of philosophy, Aristotle took up the essential nature of a science, viz., a systematic body of knowledge determined by its object of study and by its assumed basic principles, the demonstration and justification of which are relegated to another science. Metaphysics, or "first philosophy," is consequently assigned the responsibility of establishing and criticizing the first principles of all scientific knowledge, just as its object of investigation transcends the individual classes of things specifying each special science and focusses upon undifferentiated existence, thus characterizing it as the science of being qua being.

Nature of metaphysics as a science

The examples of scientific principles that Aristotle cited most frequently from his predecessors are definitions. Because his initial formulation of the problems of "first philosophy" (in *Metaphysica*, Book 1) was in terms of the insights and insufficiencies of the analyses of his predecessors, the basic metaphysical difficulties were intimated in what he said of the definitions of philosophers and of the methods that they used to establish definitions.

Aristotle distinguished two kinds of definitions and methods, the dialectical method of the Socratic–Platonic tradition and the "physical" method of Democritus and other exponents of Atomism (whose descriptions and definitions of the physical world derive from the observation of the behaviour of matter, without directly inquiring into the nature of definitions). Aristotle sought to combine the virtues of these two methods to give the Forms of Plato a material embodiment and to disclose formal (intrinsic) traits beyond the qualities of simple and compound elements in the matter of Democritus. In his view, both methods reduce philosophy to a single inquiry: the followers of Plato made all philosophy mathematics, while, at the other extreme, the Atomists reduced metaphysics and all of the sciences to physics because all things were found to be material and all changes to be chance motions in space.

The *Metaphysica* maintains a delicately poised argument concerning the relation of the principles and causes of things to the principles and axioms of knowledge. Pinpointing the difficulties that particularly enmeshed Plato (his tendency to make substances of universals and to suppose that the terms in formulas are always used in the same significance), Aristotle repeatedly and specifically warned of the numerous meanings of each of the terms that are introduced; thus all of Book 5 is devoted to a lexicon of the meanings of 30 such terms. With careful distinctions between how men think about things and how they state what they know (thus embracing with one stroke the meanings of knowledge, statement, and being), Aristotle transmuted the principles of formal and material causality into those of actuality and potentiality, or into the active and passive principles involved in any process such as motion or knowledge. Then, uniting these principles in the universality of being—viz., into the three kinds of substance: temporal sensible, eternal sensible, and immovable (*i.e.*, perfect and unchanging)—and stating the nature of substance in terms of actuality and potentiality, Aristotle led his argument beyond such considerations to the conditions that determine the relation of actuality and potentiality in a unified universe in which each substance is involved in constant and unending motion. This consideration of motion leads by way of one of the most famous arguments in philosophy to the final cause of substances and of change, the Unmoved Mover, the First Cause of being and motion.

These distinctions in kinds of substance reflect those made in the *Physica*, inasmuch as the distinctions made in the *Metaphysica* relevant to unity and numbers and to continuity and dimensions are dialectically involved in the fundamental conceptions of Aristotle's mathematics. (The *Physica*, indeed, introduces an additional classification of substance according to kinds of change.)

Metaphysics as dialectical inquiry

The method of inquiry in Aristotle's metaphysics is dialectical: the subject matter is determined by starting with what men think actually exists or is and inquiring into the conditions of being; the science does not treat of one class of being; the meanings of important terms are changed systematically; the principles that apply to all things are not literal and univocal but analogical.

From the consideration of their subject matters and their purposes, however, two of the basic classifications that differentiate the sciences themselves and the manner of their pursuit have arisen. By means of their subject matter Aristotle thus differentiated three sciences: metaphysics, mathematics, and physics, or natural science. The purpose of all three is simply knowledge, and they are therefore grouped together as theoretical sciences, to be distinguished by their purpose from the practical sciences (ethics and politics), which are directed to action and conduct, and from the productive sciences (art and rhetoric), which are intended for the making of things. These distinctions are amplified and applied in all of the works of Aristotle. There is only slight evidence concerning how he would have handled mathematics; how he did treat it in his lost works can only be conjectured from his incidental remarks about mathematics in tracts devoted to other subjects (as his perusal, for example, in Book 13 of the *Metaphysica*).

Tracts on ethics and politics. In emphasizing the crucial differences in the purposes of the theoretical and practical sciences, Aristotle indicated that the practical disciplines, unlike the theoretical, are for the sake of doing or making something and not for the sake of contemplating, defining, or knowing it. Thus, at the start of his *Ethica Nicomachea* (*Nicomachean Ethics*), he explained how the practical sciences are incapable of the exactness of the theoretical sciences, for their subject matters are not limited to things or natures, which are subject to precise definition, but involve habits and skills, which can be acquired and lost, and associations and institutions, which

Nature of ethics as a practical science

in their changes affect the accomplishment of political actions and the practicability of moral ends. And however precise biological or psychological definitions may be, man varies as moral agent and as citizen according to environmental determination, educational background, and the influences of family, economic position, social class, means of livelihood, and even the associations of his casual leisure.

Relating ethics to politics, Aristotle set out to demonstrate by definitions and distinctions from physics and metaphysics that problems of morality as they affect the individual cannot be separated from each other or from problems of political association. The *Ethica Nicomachea* and the *Politica* therefore do not develop separate sciences or independent subject matters but rather supplement each other by treating a common field according to different aspects.

Although he treated moral problems in terms of the potentialities of individual men, the practicability and actualization of these potentialities are conditioned by political circumstances. Therefore, in the first chapters of Book 1 of the *Ethica Nicomachea*, Aristotle began by introducing moral considerations into the broad context of political philosophy, and he ended by returning, in the concluding Book 8, from the examination of happiness and the contemplative life to a shrewd statement of the contribution of law to moral questions, which forms the transition from ethics to politics.

At first investigating ethics from the perspective of man's knowledge and action for the sake of some good, Aristotle treated the question dialectically because verbal agreement can be found in the general identification of happiness with the good. After further examination disclosed general disagreement concerning the nature of happiness, however, he abandoned (at the end of Book 1) the attempt to define happiness as an end to be achieved and identified it rather as activity of a certain kind. Specifically, it is activity in accordance with virtue; and the kinds of virtues are determined by the powers of the soul, irrational and rational. Both of the classes of virtues corresponding to these powers—the moral and intellectual virtues—are habits or qualities of the soul. Finally, when the question of happiness is taken up again on the foundations furnished by these inquiries into the actions, passions, and powers of man and by the formal definitions of virtue, it is treated together with other possible motivations to action and such other ends as pleasure, honour, praise, continence, and friendship, which might be associated or confused with virtue and happiness.

Happiness as virtuous activity

The *Politica* takes up the problems of human action and association as they bear on the ends of communal life encompassed in living well, but the choice of political ends requires a complex examination of possible criteria capable of application to the vast diversity of men and conditions of men. Grounding his argument on the premise that man is "naturally" a political animal, Aristotle developed the theory of the state, distinguished six kinds, and considered the best state for the particular circumstances, character, and conditions of the citizens. The tract concludes by treating the conditions under which the characters of the citizens are inconsistent with the ends of the state, as the *Ethica Nicomachea* had ended by associating the ends of man with the regulations of the state.

Tracts on art and rhetoric. Aristotle depicted rhetoric in terms of its end or final cause, which is persuasion. Like dialectic, he stated, it is not a science, and therefore it has no specific subject matter, no single method, and no proper set of principles. It is simply the faculty or power of observing in any given case the available means of persuasion. According to Aristotle, there are three modes of persuasion that a speaker may exercise: the persuasive power of his own character, the excitation of desired emotions in the audience, and proof or apparent proof.

In the *Poetica* (*Poetics*) Aristotle's analysis of poetry provides for careful isolation of the specific or, in the modern sense of the term, "aesthetic" character of poetry while permitting the analogy of poetry to oratory as well as the consideration of the moral, political, and educational effects of both. Tragedy, however, which is the only kind of poetry analyzed in the extant portions of Aristotle's work, is defined in terms of its form, not in terms of its purpose, as a kind of imitation rather than as a mode of persuasion or excitation. Thus, in the famous definition of the sixth chapter, it imitates a serious action of great magnitude in a dramatic form to accomplish the catharsis of emotions.

Nature of tragic poetry

Applying this definition as the basis for the discussion of poetry in a part–whole analysis—as in the analysis of the universe in terms of its parts in *De caelo* and in that of animals in terms of their parts in *De partibus animalium*—Aristotle considered poetic art in terms of the characteristics and interrelations of the six parts of tragedy: plot, character, and thought (the objects of imitation); diction and melody (the means of imitation); and spectacle (the manner of imitation).

The last four chapters of the *Poetica* return to more general questions of value and to final causes by means of detailed comparisons of tragedy with comparable poetic works and specifically with the epic.

ON READING ARISTOTLE

Goethe compared Aristotle's philosophy to a pyramid rising on high in regular form from a broad base on the Earth. Because each part of Aristotle's philosophy contributes to the understanding of other parts, it is generally true of him—even more than of most philosophers—that he cannot be read initially with sure and well-grounded understanding but must be reread for the sake of perceiving the primary conceptual and methodological relationships. Faced with the mass of materials that constitute the imposing body of his works, the reader might prudently start with the treatment of problems that are relevant, interesting, or important to him. Acting on the basis of effective motivation technique, Aristotle often reiterated the suggestion that one concentrate first on sense experience as something that is better known to the inquirer and attend only afterward to the essential concepts of things in the effort to organize knowledge and constitute the principles of the particular sciences. He frequently distinguished the process of learning from that of proof, inquiry and discovery from demonstration and knowledge.

Interrelationship of all Aristotle's writings

Aristotle's thought, consequently, can be said to be known in two ways: (1) as remnants of his doctrines constituting the speech of Western culture, in the tradition of Western thought, and in the history of its sciences; (2) as it is known from the attentive study of his writings. What he discovered in his intellectual inquiries and what he said are most readily intelligible to someone of Greco-Roman culture in the modified form in which it still constitutes part of that thought and conviction. It is suggested, therefore, that Aristotle be read for the first time in an order the reverse of that in which his works have traditionally been arranged and that his conclusions and analyses be examined before inquiring about his principles. It is true, for example, that what Aristotle said concerning the poetic and rhetorical arts is more complex in manner of analysis and more difficult in systematic construction than what modern writers might say on such subjects; but it still approximates more nearly to contemporary thought in this area than do his works on physics and metaphysics. And, for the same reason, his moral and political theory led him, in the course of its development, to many distinctions and statements that a modern reader would be disposed to accept or to reject without too lengthy critical discussion.

Once the manner of Aristotle's analysis is more firmly grasped and better appreciated, the reader can take up his logical works in the *Organon* and his investigations of space, time, and motion in the *Physica*. Through later consultation with his more complex thought in the *Metaphysica* and in his tract *De anima*, he can review the earlier tentative conclusions made in the study of ethics and literary theory in the light of the deeper insights, acute distinctions, and strength of argument thus acquired.

When a student approaches Aristotle's conclusions in

the light of his principles, allowing the text to illuminate his own experiences, Aristotle can then be appreciated not only for his expression of a philosophy but also as a help in the cultivation of the mind, a task to which he thought all men were dedicated and to which his philosophy remains a unique contribution.

MAJOR WORKS

MEMORANDA AND COLLECTIONS: *Divisiones quae vulgo dicuntur Aristoteleae* (published 1906), a collection of semantic observations; *Athenaion politeia* (*The Athenian Constitution*, trans. by H. Rackham, 1935); *Aristotelis Problemata Physica* (1922; *Aristotle's Book of Problems*, trans. 1710), scientific enquiry.

SCIENTIFIC AND PHILOSOPHICAL TREATISES: (LOGICAL WORKS): collectively known as the *Organon*, in Greek and English (1938), containing *Categories*, *On Interpretation*, and *Prior Analytics; Analytica posteriora* (published 1534; *Aristotle's Posterior Analytics*, trans. by E.S. Bouchier, 1901); *Sophistici elenchi* (Greek and English, *Sophistical Refutations*, 1866); *Topica* (published in Latin 1602). (PHYSICAL WORKS): *Aristotle: The Physics* (Greek and English trans. by P.H. Wicksteed, 1929); *On the Heavens* (Greek and English trans. by W.K.C. Guthrie, 1939); *Aristotelous peri geneseōs kai phthoras* (1922; *Aristotle: Meteorologica*, Greek and English trans. by H.D.P. Lee, 1952). (PSYCHOLOGICAL WORKS): *Aristotle: De Anima* (Greek and English trans. by R.D. Hicks, 1907). *Parva naturalia*, the whole published in 1955, is a collective title covering *Aristotle: De Sensu* and *De Memoria* (Greek and English trans. by G.R.T. Ross, 1906), and *De somno et vigilia*, *De insomniis*, *De divinatione per somnia*, *De longitudine et brevitate vitae*, *De juventute et senectute*, *De vita et morte*, and *De respiratione*. (PHILOSOPHICAL WORKS): *Aristotle's Metaphysics*, 2 vol. (published in 1924; trans. by R. Hope, 1952); *Aristotle: The Nichomachean Ethics* (Greek and English trans. by H. Rackham, 1926); *Aristotelis Ethica Eudemia* (1851, Greek and Latin, *Eudemian Ethics*, trans. by H. Rackham, 1935); *The Politics* (Greek and English trans. by H. Rackham, 1932); *Aristotle: The Poetics* (Greek and English trans. by W. Hamilton Fyfe, 1927). (WORKS ON NATURAL HISTORY): *Aristotelis De animalibus historia* (published 1907; *Aristotle's History of Animals*. In ten books, trans. by R. Creswell, 1862); *Aristotelis De partibus animalium* (1868; *Aristotle on the Parts of Animals*, trans. by W. Ogle, 1882); *De motu animalium*, in Latin only (1566); *De incessu animalium; Aristotle: Generation of Animals* (Greek and English trans. by A.L. Peck, 1943).

POPULAR WRITINGS: Philosophical dialogues of which mere fragments remain: *Eudemus; Protrepticus; On Philosophy; On the Good; On Ideas* or *On Forms; On Justice; On Monarchy; Gryllus; Menexenus; Nerinthus; Alexander*, or *On Colonization; Peplus* (published 1798), extant collection of epitaphs for Homeric heroes; possibly the work of another Aristotle, not the philosopher.

BIBLIOGRAPHY. The standard edition of the Greek text of Aristotle is that of the Berlin Academy, *Aristotelis Opera*, ed. by IMMANUEL BEKKER, 5 vol. (1831–70; new ed. of vol. 1, 2, 4, and 5 by OTTO GIGON, 1960–61). This is supplemented by *Aristotelis Fragmenta*, ed. by VALENTIN ROSE (Leipzig, 1886, reprinted 1967); *Aristotelis Dialogorum Fragmenta*, ed. by RICHARD WALZER (Florence, 1934); and *Aristotelis Fragmenta selecta*, ed. by W.D. ROSS (Oxford, 1956). The text of the Berlin Academy edition has been improved upon for most of the treatises by various editors in the "Teubner Series" (1868–1961), in the "Oxford Classical Texts" (1894–1968), in the "Budé Series" (with French trans., 1926–68), and in the "Loeb Series" (with Eng. trans., 1926–66). The best English version remains *The Works of Aristotle Translated into English*, under the editorship of W.D. ROSS, 12 vol. (1908–52). Medieval Latin translations are printed in *Aristotles latinus*, planned to comprise some 34 vol. (1939–), and in *Opera cum Averrois commentariis*, 12 vol. in 14 (1562–74, reprinted 1962). Works of the Greek commentators are printed in *Commentaria in Aristotelem graeca*, together with *Supplementum aristotelicum*, 26 vol. (1812–1900).

Important modern editions and commentaries include *Prior and Posterior Analytics*, ed. by W.D. ROSS (1949); *De Anima*, ed. by W.D. ROSS (1961); *De Generatione et corruptione*, ed. by HAROLD H. JOACHIM (1922); *L'Éthique à Nicomaque*, ed. by R.A. GAUTHIER and J.Y. JOLIF, 3 vol. (1958); *Metaphysics*, ed. by W.D. ROSS, 2 vol. (1924); *Parva Naturalia*, ed. by W.D. ROSS (1955); *Physics*, ed. by W.D. ROSS (1936); *Aristotle's Poetics: The Argument*, commentary by GERALD F. ELSE (1957); *Politics*, ed. by W.L. NEWMAN, 4 vol. (1887–1902); *Politics*, Eng. trans. by ERNEST BARKER (1946).

For Aristotle's life and writings, see INGEMAR DURING, *Aristotle in the Ancient Biographical Tradition* (1957); and PAUL MORAUX, *Les Listes anciennes des ouvrages d'Aristote* (1951).

General accounts, including development, are WERNER W. JAEGER, *Aristoteles: Grundlegung einer Geschichte seiner Entwicklung* (1923; Eng. trans., *Aristotle: Fundamentals of the History of His Development*, 2nd ed., 1948); W.D. ROSS, *Aristotle*, rev. ed. (1964); ETTORE BIGNONE, *L'Aristotele perduto . . .*, 2 vol. (1936); HAROLD CHERNISS, *Aristotle's Criticism of Plato and the Academy* (1944, reprinted 1962); DONALD J. ALLAN, *The Philosophy of Aristotle* (1952); ENRICO BERTI, *La Filosofia del primo Aristotele* (1962); INGEMAR DURING, *Aristoteles, Darstellung und Interpretation seines Denkens* (1966).

On Aristotle's logic, see FRIEDRICH SOLMSEN, *Die Entwicklung der aristotelischen Logik und Rhetorik* (1929); J.M. LE BLOND, *Logique et méthode chez Aristote* (1939); JAN LUKASIEWICZ, *Aristotle's Syllogistic from the Standpoint of Modern Formal Logic*, 2nd ed. (1957); GUNTHER PATZIG, *Die aristotelische Syllogistik*, 2nd ed. (1963; Eng. trans., *Aristotle's Syllogistic*, 1968); and WILLIAM and MARTHA KNEALE, *The Development of Logic* (1962). On his physics, see AUGUSTE MANSION, *Introduction à la physique aristotélicienne*, 2nd ed. (1946); FRIEDRICH SOLMSEN, *Aristotle's System of the Physical World* (1960); and WOLFGANG WIELAND, *Die aristotelische Physik* (1962). On his psychology, see JOHN I. BEARE, *Greek Theories of Elementary Cognition . . .* (1906); and FRANCOIS NUYENS, *Ontwikkelingsmomenten in de Zielkunde van Aristoteles* (1939; French trans., *L'Évolution de la psychologie d' Aristote*, 1948). On his metaphysics, see JOSEPH OWENS, *The Doctrine of Being in the Aristotelian Metaphysics*, 2nd ed. (1963); and PIERRE AUBENQUE, *Le Problème de l'être chez Aristote*, 2nd ed. (1966). On his theology, see W.K.C. GUTHRIE, "Development of Aristotle's Theology," *Classical Quarterly*, vol. 27 (1933). On his ethics and politics, see ERNEST BARKER, *The Political Thought of Plato and Aristotle* (1906, reprinted 1959); JEAN LEONARD, *Le Bonheur chez Aristote* (1948); PIERRE AUBENQUE, *La Prudence chez Aristote* (1963); and W.F.R. HARDIE, *Aristotle's Ethical Theory* (1968). On his biology, see THOMAS E. LONES, *Aristotle's Researches in Natural Science* (1912); and MARJORIE GRENE, *A Portrait of Aristotle* (1963). On his mathematics, see THOMAS L. HEATH, *Mathematics in Aristotle* (1949); and H.G. APOSTLE, *Aristotle's Philosophy of Mathematics* (1952).

(A.H.Ao.)

Arithmetic

Arithmetic (derived from the Greek *arithmos*, "number") refers generally to the elementary aspects of the theory of numbers, arts of mensuration (measurement), and numerical computations. Such meaning, however, has not been uniform in mathematical usage. An eminent German mathematician, Carl Friedrich Gauss (*q.v.*), in *Disquisitiones Arithmeticae* (1801), and certain contemporary mathematicians have used the term to include some aspects of number theory that are considerably more advanced. The reader interested in the latter is referred to NUMBER THEORY.

FUNDAMENTAL DEFINITIONS AND LAWS

In a collection (or set) of 1, 2, 3, or, generally, n objects (or elements), the act of determining the number of objects present is called counting. For an empty set, no object is present, and the count yields the number 0. The numbers n thus obtained are called natural numbers; there are some mathematicians who include 0 among the natural numbers, whereas there are others who do not include it.

If n_1 and n_2 are the numbers of objects in the two sets S_1 and S_2 respectively, then those objects in S_1 and S_2 may possibly be paired in such a way that each pair contains one object from S_1 and one from S_2. The two numbers n_1, n_2 would then be said to be equal; that is, $n_1 = n_2$. The two sets, in such a case, are said to be equivalent. The concept of equivalent sets is basic to the foundations of modern mathematics and, through the techniques of the "new math," has been introduced into primary education, as shown in Figure 1. The theory of sets is discussed in the articles SET THEORY and MATHEMATICS, FOUNDATIONS OF.

Taking the two sets S_1 and S_2 together to form a new set S_3, it is clear that S_3 contains $n_1 + n_2 = n_3$ objects. The number n_3 is called the sum of n_1, n_2; and each of the latter is called a summand. The operation of forming the sum is called addition, the symbol $+$ being read as plus.

From the definition of counting it is evident that two fundamental laws hold for the operation of addition (see

Box, laws 1 and 2), in which the order of the summands can be changed and the order of the operation of addition can be changed, when applied to three summands, without affecting the sum. These are called the laws of commutativity and associativity for addition. To dispel any notion that these laws are trivially self-evident, it is notable that the first of these, for example, does not hold for the addition of ordinary (finite) rotations. If a closed book that is thin in comparison with its page size is lying flat on a table in an orientation ready to be read, and if a denotes a rotation of the book through 90° about a horizontal axis fixed at the upper edge of the book, and b denotes a similar subsequent rotation about a vertical axis, then adding the rotations $a + b$, in this order, will readily be found to lead to a final position of the book in which the latter is vertical, upside-down, and with the "thin" dimension facing the reader. Performing the sum $b + a$, on the other hand, will cause the book's thin dimension to rest on the table, the front cover facing toward (or away from) the reader. Hence, in this case, $a + b$ is clearly not equal to $b + a$.

Symbols > and <

If there exists a natural number k such that $a = b + k$, it is said that a is greater than b (written $a > b$), and that b is less than a (written $b < a$). If a and b are any two natural numbers, then it is the case that either $a = b$ or $a > b$ or $a < b$.

From the above laws it is evident that a repeated sum such as $5 + 5 + 5$ is independent of the way in which the summands are grouped and is written 3×5. Thus, a second binary operation called multiplication is defined. The number 5 is called the multiplicand, the number 3, which denotes the number of summands, is called the multiplier, and the result 3×5 is called the product. The symbol $\times$ of this operation is read "times." If such letters as a and b are used to denote the numbers, the product $a \times b$ is often written $a \cdot b$ or simply ab.

If three rows of five dots each are written, as illustrated below,



it is clear that the total number of dots in the array is 3×5, or 15. This same number of dots can evidently be written in five rows of three dots each, whence $5 \times 3 = 15$. The argument is general, leading to the law (see 3) that the order of the multiplicands does not affect the product, called the commutative law of multiplication. Here again, it is notable that this law does not apply to all mathematical entities. Indeed, much of the mathematical formulation of modern physics, for example, depends crucially on the fact that some entities do not commute.

By the use of a three-dimensional array of dots, it becomes evident that the order of multiplication when applied to three numbers does not affect the product

(1) The commutative law of addition: $a+b=b+a$

(2) The associative law of addition: $a+(b+c)=(a+b)+c$

(3) The commutative law of multiplication: $ab=ba$

(4) The associative law of multiplication: $a(bc)=(ab)c$

(5) The distributive law: $(a+b)c=ac+bc$

(6) $a^m a^n = a^{m+n}$

(7) $(a^m)^n = a^{mn}$

(8) $a^m b^m = (ab)^m$

(9) $a^m \div a^n = a^{m-n} \qquad m > n$

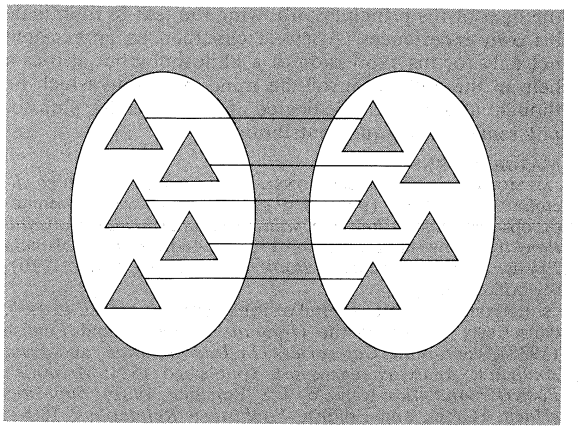
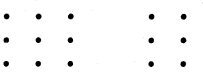


Figure 1: Page from a first grade workbook typical of "new math" might state: "Draw connecting lines from triangles in the first set to triangles in the second set. Are the two sets equivalent in number?"

(see 4). Such a law is called the associative law of multiplication. If the 15 dots written above have separated into two sets, namely,



then the first set consists of three columns of three dots each, or 3×3 dots; the second set consists of two columns of three dots each, or 2×3 dots; the sum $(3 \times 3) + (2 \times 3)$ consists of $3 + 2 = 5$ columns of three dots each, or $(3 + 2) \times 3$ dots. In general, one may prove that the multiplication of a sum by a number is the same as the sum of two appropriate products. Such a law (see 5) is called the distributive law.

Subtraction has not been introduced for the simple reason that it can be defined as the inverse of addition. Thus, the difference $a - b$ of two numbers a and b is defined as a solution x of the equation $b + x = a$. If a number system is restricted to the natural numbers, differences need not always exist, but if they do, the five laws of arithmetic, as already discussed, can be used to prove that they are unique. Furthermore, the laws of operations of addition and multiplication can be extended to apply to differences. The natural numbers (including zero) can be extended to include the solution of $1 + x = 0$, that is, the number -1, as well as all products of the form $-1 \times n$, in which n is a natural number. The extended collection of numbers is called the integers, of which the positive integers are the same as the natural numbers (excluding zero). The numbers that are newly introduced in this way are called negative integers.

Just as a repeated sum $a + a + \cdots + a$ of k summands is written ka, so a repeated product $a \times a \times \cdots \times a$ of k factors is written a^k. The number k is called the exponent, and a the base of the power a^k.

The fundamental laws (see 6–8) of exponents follow easily from the definitions, and other laws (see 9) are immediate consequences of the fundamental ones. The first five laws are often called the five fundamental laws of arithmetic.

THEORY OF DIVISORS

Extension of classes of numbers

At this point an interesting development occurs, for, so long as only additions and/or multiplications are performed with integers, the resulting numbers are invariably again integers; that is, numbers of the same kind as their antecedents. This characteristic changes drastically, however, as soon as division is introduced. Performing division leads to results, called quotients or fractions, which surprisingly include numbers of a new kind, namely, rationals. These, though arising from the combination of integers, patently constitute a distinct extension of the natural-number and integer concepts as defined above. By means of the application of the division operation, the domain of the natural numbers becomes

extended and enriched immeasurably beyond the integers (see below *Rational numbers*).

The preceding illustrates, simply but clearly, one of the proclivities often associated with mathematical thought: relatively simple concepts (integers), initially based on very concrete operations (counting), are found to be capable of assuming novel meanings, and potential uses, extending far beyond the limits of the concept as originally defined. A similar extension of basic concepts, with even more powerful results, will be found with the introduction of irrationals (see below *Irrational numbers*).

A second example of this is presented by the following: Under the primitive definition, with k equal to either zero or a fraction, a^k would, at first sight, appear to be utterly devoid of meaning. Clarification is needed before writing a repeated product of either zero factors or a fractional number of factors. Yet, limiting attention to the case $k = 0$ here (k = fraction will be considered below), a little reflection shows that a^0 can, in fact, assume a perfectly precise meaning, coupled with an additional and quite extraordinary property. Since the result of dividing any (nonzero) number by itself is unity, it follows that $a^m \div a^m = a^{m-m} = a^0 = 1$. Not only can the definition of a^k be extended to include the case $k = 0$, but the ensuing result also possesses the noteworthy property that it is independent of the particular (nonzero) value of the base a. A similar argument may be given to show that a^k is a meaningful expression even when k is negative, namely, $a^{-k} = 1/a^k$. The original concept of exponent is thus broadened to a great extent.

Fundamental theory. If three positive integers a, b, and c are in the relation $ab = c$, it is said that a and b are divisors or factors of c, or that a divides c (written $a \mid c$), and b divides c. The number c is said to be a multiple of a and a multiple of b.

The number 1 is called the unit, and it is clear that 1 is a divisor of every positive integer. If c can be expressed as a product ab in which a and b are positive integers each greater than 1, then c is called composite. A positive integer neither 1 nor composite is called a prime. Thus, 2, 3, 5, 7, 11, 13, 17, 19, $\cdots$ are prime numbers. Euclid proved that the number of prime numbers is infinite (*Elements*, book IX, proposition 20).

The fundamental theorem of arithmetic was proved by Gauss in his *Disquisitiones Arithmeticae*. It states that every composite number can be expressed as a product of prime numbers and that, save for the order in which the factors are written, this representation is unique. This theorem follows rather directly from a theorem of Euclid (*Elements*, book VII, proposition 30) to the effect that if a prime divides a product, it divides one of its factors, and the fundamental theorem is therefore sometimes credited to Euclid (*q.v.*).

Gauss's contribution

For every finite set $a_1, a_2, \cdots, a_k$ of positive integers, there exists a largest integer d that divides each of these numbers, called their greatest common divisor (GCD). If $d = 1$, the numbers are said to be relatively prime. There also exists a smallest positive integer m that is a multiple of each of the numbers. This is called their least common multiple (LCM).

If $p_1, p_2, \cdots, p_h$ are the distinct primes that divide all of the numbers $a_1, a_2, \cdots, a_k$, and if e_i is the smallest exponent to which p_i occurs in any of them, then (see 10) the product of the powers formed with typical number p_i as base and e_i as exponent is the GCD of $a_1, a_2, \cdots, a_k$. If $p_1, p_2, \cdots, p_t$ are the distinct primes that divide any one or more of the numbers $a_1, a_2, \cdots, a_k$, and if n_i is the largest exponent to which p_i occurs in any of them, then (see 11) the product of the powers formed with typical number p_i as base and n_i as exponent is the LCM of $a_1, a_2, \cdots, a_k$. An example (see 12) is easily constructed. When only two numbers are involved, the GCD and the LCM combine to give the same product as the product of the original numbers.

If a and b are two positive integers, $a > b$, by means of the division algorithm two integers q and r can be determined such that (see 13) a is the sum of q numbers of magnitude b plus a number r that is less than b.

The number q is called the partial quotient (the quotient if $r = 0$), and r is called the remainder. The GCD of a and b is equal to the GCD of b and r. If the division algorithm is applied successively, a remainder 0 must ultimately appear. The last positive remainder is the GCD of a and b. Thus, if $a = 554$, $b = 119$, a simple calculation (see 14) shows that the GCD of 544 and 119 is 17. This process is known as the Euclidean algorithm. By means of it, the GCD can be obtained without first factoring the numbers a and b into prime factors.

(10) $$d = p_1^{e_1} p_2^{e_2} \cdots p_h^{e_h}$$

(11) $$m = p_1^{n_1} p_2^{n_2} \cdots p_t^{n_t}$$

(12) $$a_1 = 2^3 3^1 5^3 7^4 \qquad a_2 = 2^1 3^3 7^3 (11)^2$$
$$d = 2^1 3^1 7^3 \qquad m = 2^3 3^3 5^3 7^4 (11)^2$$

(13) $$a = bq + r \qquad 0 \leq r < b$$

(14) $$544 = 4(119) + 68 \qquad 119 = 1(68) + 51$$
$$68 = 1(51) + 17 \qquad 51 = 3(17)$$

(15) $$\frac{a}{d} = \frac{ka}{kd}$$

(16) $$\frac{a}{d} = \frac{ka}{kd} = \frac{ka}{m}, \qquad \frac{b}{e} = \frac{lb}{le} = \frac{lb}{m}$$

(17) $$\frac{a}{d} \pm \frac{b}{e} = \frac{ka}{m} \pm \frac{lb}{m} = \frac{ka \pm lb}{m}$$

Rational numbers. From a less abstract point of view, the notion of division, or fraction, may also be considered to arise as follows: if the time duration of a given process is required to be known to an accuracy of better than 1 hour, the number of minutes may be specified; or, if the hour is to be retained as the fundamental unit, each minute may be represented by $\frac{1}{60}$, or by 1/60.

In general, the fractional unit $1/d$ is defined by the property $d \times (1/d) = 1$. The number $a \times (1/d)$ is written a/d and is called a common fraction. It may be considered as the quotient of a divided by d. The number d is called the denominator (it determines the fractional unit or denomination), and a is called the numerator (it enumerates the number of fractional units that are taken). The numerator and denominator together are called the terms of the fraction. A positive fraction a/d is said to be proper if $a < d$; otherwise it is improper.

Common fraction

The numerator and denominator of a fraction are not unique, since for every positive integer k, the numerator and denominator of a fraction can each simultaneously be multiplied by the integer k without altering the fractional value (see 15). Thus, every fraction can be written as the quotient of two relatively prime integers. In this form it is said to be in lowest terms.

The integers and fractions constitute what are called the rational numbers. The five fundamental laws stated earlier with regard to the positive integers can be generalized to apply to all rational numbers.

Adding and subtracting fractions. From the definition of fraction it follows that the sum (or difference) of two fractions having the same denominator is a fraction with this denominator, the numerator of which is the sum (or difference) of the numerators of the given fractions. Two fractions having different denominators may be added by first reducing them to fractions with the same denominator. Thus, to add a/d and b/e the LCM m of d and e, often called the least common denominator of the fractions, is determined. It follows that there exist numbers k and l such that $m = kd = le$, and both fractions can be written (see 16) with common denominator m, so that their sum or difference (see 17) is accomplished by adding or subtracting the numerator only.

Multiplying and dividing fractions. The product of two fractions (see 18) is a fraction the numerator of which is the product of the numerators of the factors, and the denominator of which is the product of the denominators of the factors. The quotient of two fractions (see 19) is equal to the product of the dividend by the divisor inverted; that is, the divisor with its terms interchanged.

Theory of rationals. A method of introducing the positive rational numbers that is free from intuition (that is, with all logical steps included) was given by a German mathematician, Ernst Steinitz, in 1910. In considering the set of all number pairs (a, d), (b, e), $\cdots$ in which $a, b, d, e, \cdots$ are positive integers, the equals relation $(a, d) = (b, e)$ is defined to mean that $ae = bd$, and the two operations $+$ and $\times$ are defined so that the sum of a pair (see 20) is a pair and the product of a pair (see 21) is a pair. It can be proved that, if these sums and products are properly specified, the fundamental laws (see 1–5) hold for these pairs, and that the pairs of the type $(a, 1)$ are abstractly identical with the positive integers a. Moreover, $d \times (a, d) = a$, so that the pair (a, d) is abstractly identical with the fraction a/d.

Irrational numbers. It was known to the Pythagoreans that, given a straight line segment a and a unit segment u, it is not always possible to find a fractional unit such that both a and u are multiples of it. Thus, the hypotenuse of an isosceles right triangle the sides of which are taken as the unit u must by the Pythagorean theorem have a length the square of which is 2. But there exists no rational number the square of which is 2.

Eudoxus' contribution

Eudoxus of Cnidus (*q.v.*), a contemporary of Plato, established the technique necessary to extend numbers beyond the rationals. His contribution, one of the most important in the history of mathematics, was included in Euclid's *Elements*, Book V and elsewhere, and then lay dormant until the modern period of growth in mathematical analysis in Germany in the 19th century.

It is customary to assume on an intuitive basis that, corresponding to every line segment and every unit length, there exists a number (called a positive real number) that represents the length of the line segment. Not all such numbers are rational, but every one can be approximated arbitrarily closely by a rational number. That is, if α is real and ε is any positive rational number no matter how small, it is possible to find two positive rational numbers a and b within ε distance from each other such that (see 22) α is between them. In problems in mensuration, real numbers are usually replaced by rational approximations.

A rigorous development of the irrational numbers is beyond the scope of arithmetic. They are most satisfactorily introduced by means of Dedekind cuts, as introduced by the German mathematician Richard Dedekind (*q.v.*), or sequences of rationals, as introduced by Eudoxus and developed by the German mathematician Georg Cantor (*q.v.*). These methods are discussed in the theory of functions of a real variable (see ANALYSIS, REAL).

The employment of irrational numbers greatly increases the scope and usefulness of arithmetic. For instance, if n is any whole number and a is any positive real number, there exists a unique positive real number $\sqrt[n]{a}$, called the nth root of a, whose nth power is a. The root symbol $\sqrt{\ }$ is a conventionalized r for radix, or root. The term evolution is sometimes applied to the process of finding a rational approximation to an nth root.

The symbol $\sqrt{\ }$

Surds. A number (see 23) that is obtainable from rational numbers by a finite number of additions, multiplications, divisions, and root extractions is called a surd. An irrational number of the form $\sqrt[n]{a}$ in which a is rational is called a pure surd of index n. For $n = 2$, it is called a quadratic surd or square root and is written simply $\sqrt{a}$. A surd that is not pure is called mixed.

A pure surd of index n may be reduced to a rational multiple of the nth root of a positive integer no prime factor of which occurs to an exponent as high as n. Such a surd is in lowest terms. Thus consider $\sqrt[3]{125/18}$. Since $18 = 2(3^2)$, it becomes a perfect cube upon multiplication by $2^2(3) = 12$. Then (see 24) the cube root of $125/18$ can be expressed as a rational multiple of the cube root of 12.

Two surds that can be written as rational multiples of the same surd are called similar. Their sum is similar to each of them, its rational coefficient being the sum of the rational coefficients of the summands.

Properties of surds. The elementary properties of surds are readily provable (see 25–29) and guide one in calculating products of roots (see 25), roots of roots (see 26), powers of roots (see 27), roots of powers (see 28), and roots of fractions (see 29).

The theory of surds can be much simplified by the introduction of negative and fractional exponents, in which (see 30) the zeroth power of a number is unity, the negative power of a number is the reciprocal of the number, and a fractional power is expressed as a surd.

It can readily be shown that the fundamental laws of exponents, together with the properties of the rational fractions, are equivalent to the elementary properties (see 25–28) of surds. Exemplary verification (see 31, 32) is not difficult. (C.C.MacD.)

(18) $$\frac{a}{d} \times \frac{b}{e} = \frac{ab}{de}$$

(19) $$\frac{a}{d} \div \frac{b}{e} = \frac{a}{d} \times \frac{e}{b} = \frac{ae}{db}$$

(20) $$(a, d) + (b, e) = (ae + bd, de)$$

(21) $$(a, d) \times (b, e) = (ab, de)$$

(22) $$a < \alpha < b \qquad a - b < \epsilon$$

(23) $$\sqrt{(\sqrt{2} + \sqrt[5]{3})/\sqrt{5}}$$

(24) $$\sqrt[3]{\frac{125}{18}} = \sqrt[3]{\frac{125(12)}{18(12)}} = \sqrt[3]{\frac{5^3(12)}{6^3}} = \frac{5}{6}\sqrt[3]{12}$$

(25) $$\sqrt[n]{a}(\sqrt[n]{b}) = \sqrt[n]{ab}$$

(26) $$\sqrt[n]{\sqrt[m]{a}} = \sqrt[mn]{a}$$

(27) $$(\sqrt[m]{a})^k = \sqrt[m]{a^k}$$

(28) $$\sqrt[mn]{a^{kn}} = \sqrt[m]{a^k}$$

(29) $$\sqrt[n]{\frac{a}{b}} = \frac{\sqrt[n]{a}}{\sqrt[n]{b}}$$

(30) $$a^0 = 1 \qquad a^{-n} = \frac{1}{a^n} \qquad a^{p/q} = \sqrt[q]{a^p} \qquad a \neq 0$$

(31) $$\sqrt[n]{a}(\sqrt[n]{b}) = a^{1/n}b^{1/n} = (ab)^{1/n} = \sqrt[n]{ab}$$

(32) $$\sqrt[m]{a^k} = a^{k/m} = a^{kn/mn} = \sqrt[mn]{a^{kn}}$$

NUMBER SYSTEMS AND NOTATION

Just as in language, where infinitely many words can be composed of but a small variety of different characters, so also in arithmetic are numbers, infinitely many of them, composed of but a small variety of numerals. The devising of a scheme whereby an infinitude of things can be represented by means of but a small number of symbols must be ranked among the most important achievements of the human intellect, for without it an advanced development of either language or arithmetic is unimaginable. The key ideas are two: the positional principle (to be described below) and the symbol zero. "The idea is so simple," wrote the 18th–19th-century mathematician Pierre-Simon, marquis de Laplace, speaking of the

(33) $p \times a^0 + q \times a^1 + r \times a^2 + s \times a^3 + \cdots$

(34) $p \times 1 + q \times 10 + r \times 100 + s \times 1000 + \cdots$

positional principle, "that this very simplicity is the reason for our not being sufficiently aware how much admiration it deserves." Similarly, the invention, probably by the Hindus, of the digit zero has been described as one of the greatest importance in the history of mathematics. Hindu literature gives evidence that the zero may have been known before the birth of Christ, but no inscription has been found with such a symbol before the 9th century.

The number zero

Applied to the construction of numbers, the positional principle operates thus: the sequence of digits $\cdots srqp$ is defined to signify a number the magnitude of which is a sum of products involving powers of a number a (see 33), in which a is called the base or radix; that is, the position of each of the coefficients $p, q, r, s, \cdots$ is associated, in reverse order to the representation $srqp$, with the zeroth, first, second, third, $\cdots$ powers of the base a. The number of distinct numerals required in this notation is readily seen to be a.

Each numeral, when part of a number, can therefore be said to have two values: an intrinsic value, which is simply that signified by the isolated symbol itself; and a local value, which is that possessed by virtue of its position, or location, within the given number.

Systems are called binary, ternary, quaternary, quinary, senary, septenary, octenary (or octal), nonary, denary (or decimal), undenary, duodenary, hexadecimal, vigesimal, and sexagesimal, corresponding to values of $a = 2$, 3, 4, 5, $\cdots$ 20, and 60, respectively. The pair system, in which the counting goes "one, two, two and one, two twos, two and two and one," etc., is found among the ethnologically oldest tribes of Australia, in many Papuan languages of Torres Strait and the adjacent coast of New Guinea, among some African Pygmies, and in various South American tribes. Other tribes of Tierra del Fuego and the South American continent use number systems with bases three and four. The quinary scale, or number systems with base five, is very old but in pure form seems to be used at present only in Saraveca, a South American Arawakan language; elsewhere it is combined with the decimal or the vigesimal system, where the base is 20. Similarly, the pure base six scale seems to occur only sparsely in Northwest Africa and is otherwise combined with the duodecimal, or base 12, system.

In the course of history the decimal system finally overshadowed all others, and it is now found in all nations of high culture on the entire globe, except those of Mexico and Central America, where the number 20 was used in astronomy and thus became firmly entrenched. Nevertheless, there are still many vestiges of other systems in nations of high culture, chiefly in commercial and domestic units, where change always meets the resistance of tradition. Thus, there occurs 12 as the number of inches in a foot, pence in a shilling, months in a year, and twice 12 hours in a day; and both dozen and gross measure by twelves. In English the base 20 occurs chiefly in the word score ("Four score and seven years ago . . ."), but other traces are found in pre-English Celtic, Gaelic, Danish, and Welsh.

Ancient number systems

The Babylonians developed (2000–3000 BC) a positional system with base 60—a sexagesimal system. With such a large base it would be awkward to have unrelated names for the digits 0, 1, $\cdots$, 59, so a simple grouping system to base 10 was used for these numbers. For example, the number 258,458 is written in Figure 2 as it would have appeared in Babylon. The base 60 still occurs in measurement of time and angles.

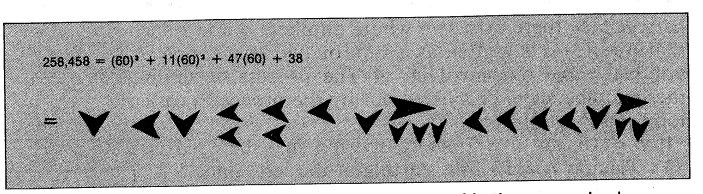


Figure 2: The number 258,458 expressed in the sexagesimal system of the Babylonians and in cuneiform.

In addition to being somewhat cumbersome because of the large base chosen, the Babylonian system suffered until very late from the lack of a zero symbol; the resulting ambiguities may well have bothered the Babylonians as much as later translators.

In the course of early Spanish expeditions into Yucatán it was discovered that the Mayans, at an early but still undated time, had a well-developed positional system, complete with zero. It seems to have been used primarily for the calendar rather than for commercial or other computation; this is reflected in the fact that although the base is 20, the third digit from the end does not signify multiples of 20^2 but of 18×20, thus giving their year a simple number of days. The digits 0, 1, $\cdots$, 19 are, as in the Babylonian, formed by a simple grouping system, in this case to base 5; the groups were written vertically, as in Figure 3.

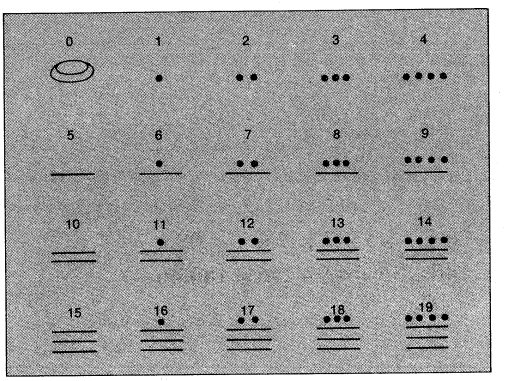


Figure 3: The Mayan vigesimal number system (see text).

The earliest numerals of which there is definite record were simple straight marks for the small numbers, with some special form for 10. These symbols appear in Egypt as early as *c.* 3400 BC and in Mesopotamia as early as *c.* 3000 BC. These dates long precede the first known inscriptions containing numerals in India (*c.* 300 BC), in China (3rd century BC), and in Crete (*c.* 1200 BC). Some ancient symbols for 1 and 10 are given in Figure 4. This special position occupied by 10 stems from the number of human fingers, of course, and is still evident in modern usage not only in the logical structure of the decimal system but in the English names for the numbers. Thus, eleven comes from Old English *endleofan*, literally meaning "[ten and] one left [over]," and twelve from *twelf*, meaning "two left"; the endings -teen and -ty both refer to 10, of course, and hundred comes originally from a pre-Greek term meaning "ten times [ten]."

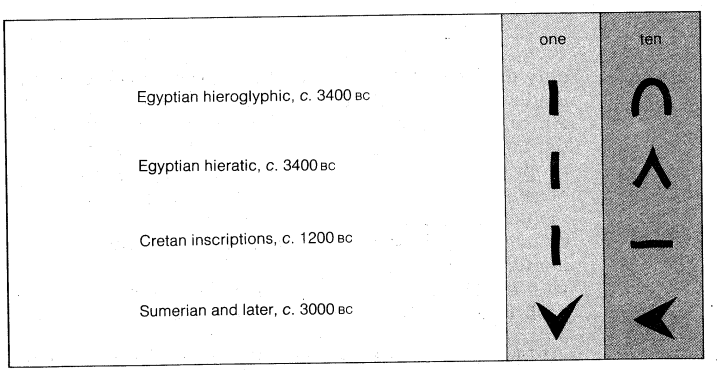


Figure 4: Some ancient symbols for 1 and 10.

In the decimal system the above number $\cdots srqp$ has a value expressed as a sum of products involving powers of 10 (see 34). The ten numerals 1, 2, $\cdots$ 8, 9, 0 are the

(35) $N = 0 + u \times a^{-1} + v \times a^{-2} + w \times a^{-3} + \cdots$

(36)
$$\begin{array}{r} 47.65 \\ 5.473 \\ \underline{649.8} \end{array}$$

Sum calculated as follows:

$$\begin{array}{r} 4(10) + 7 + 6(10^{-1}) + 5(10^{-2}) \\ 5 + 4(10^{-1}) + 7(10^{-2}) + 3(10^{-3}) \\ \underline{6(10^{2}) + 4(10) + 9 + 8(10^{-1})\phantom{+ 3(10^{-3})}} \\ 6(10^{2}) + 8(10) + 21 + 18(10^{-1}) + 12(10^{-2}) + 3(10^{-3}) \end{array}$$

(37) $7(10^2) + 0(10) + 2 + 9(10^{-1}) + 2(10^{-2}) + 3(10^{-3}) = 702.923$

(38)
$$\begin{array}{r} \underline{21\ 1} \\ 47.65 \\ 5.473 \\ \underline{649.8} \\ 702.923 \end{array}$$

(39)
$$\begin{array}{r} 563.142 \\ \underline{-\ 170.8} \\ 392.342 \end{array}$$

(40) $25.78 \times 7 = 20 \times 7 + 5 \times 7 + .7 \times 7 + .08 \times 7$
$= 140 + 35 + 4.9 + .56 = 180.46$

(41)
$$\begin{array}{r} \underline{45\ 5} \\ 25.78 \\ \underline{7} \\ 180.46 \end{array}$$

(42)
$$\begin{array}{r} 25.78 \\ \underline{45.7} \\ 18\ 046 \\ 128\ 90 \\ \underline{1\ 031\ 2} \\ 1\ 178.146 \end{array}$$

only distinct symbols needed here. (These symbols, originated in large part in India and widely used by the Arabs, are in several instances stylized forms of the initial characters of the words employed to denote the appropriate numbers. They are called Hindu-Arabic numerals.) This notation is capable of being expanded to include the representation of fractional parts of numbers by means of placing a point (decimal point when $a = 10$) immediately to the right of the coefficient associated with a^0. Thus, $0.uvw \cdots$ represents the sum of products, each term of which includes a negative power of a (see 35).

With the advent of electronic high-speed computers, the choice $a = 2$ and, hence, binary arithmetic have become of great practical importance, though, as has been mentioned above, the binary (or pair) system has been used in tribal, nontechnical cultures. The fundamental reason for the new importance is the basic mode of operation of computing machines, in which an electric current pulse, or the direction of magnetization of an element, has one of two values. The numerals used in the binary system are 1 and 0. As an example, the magnitude of the number $N_2 = 100111$ to the base 2 (indicated by the subscript on N) is $1 \times 2^0 + 1 \times 2^1 + 1 \times 2^2 + 0 \times 2^3 + 0 \times 2^4 + 1 \times 2^5$, which in decimal notation becomes $N_{10} = 1 + 2 + 4 + 0 + 0 + 32 = 39$. It is apparent that the binary system's use of only two distinct symbols is partly offset by the requirement of many more digits (6 in N_2, as compared with only 2 in N_{10} in the above example). Octal and hexadecimal modes also find frequent application in electronic computing devices.

Binary numbers

In octal notation the above $N_{10} = 39$ is readily seen to become $N_8 = 47$ because $7 \times 8^0 + 4 \times 8^1 = 7 \times 1 + 32 = 39$ (in decimal notation). (D.E.S./W.J.LeV.)

ARITHMETIC CALCULATION WITH DECIMALS

Addition and subtraction of decimals. Numbers in decimal notation can easily be added by adding the coefficients of corresponding powers of 10 and then adjusting, by the process known as carrying, the coefficients that exceed 9. A typical example (see 36) would be the addition of the numbers 47.65, 5.473, and 649.8. One first adds coefficients of equal powers of 10; *e.g.*, one adds 5 and 7 to obtain 12 as a coefficient of 10^{-2}. After adjustment of the coefficients (see 37), this sum becomes 702.923.

The process of addition is commonly carried out as follows (see 38): Beginning at the right, the sum of the coefficients of 10^{-3}, which is 3, is written below the answer line; the sum of the coefficients of 10^{-2} is 12, the digit 2 being placed to the left of the 3 in the sum; however, the 10 becomes a 1 in the preceding position and may be placed at the top of the column of coefficients of 10^{-1}, etc. In practice the carrying is usually performed mentally.

Subtraction is performed by reversing the above procedure. Thus, to subtract 170.8 from 563.142 (see 39), for example, the smaller number is placed under the larger, with the respective decimal points aligned vertically. Again, starting from the right, since $0.8 = 0.800$, 0 subtracted from 2 leaves 2 (which is written below the line), and 0 from 4 leaves 4; but since 8 exceeds 1, a unit is borrowed from the unit's place (which changes the 3 to 2), and thus the 1 in the tenth's place becomes 11; then 8 subtracted from 11 leaves 3, 0 from 2 leaves 2, and so forth.

Another way of performing subtractions consists in finding the number, digit by digit, that, when added to the second, yields the first as sum.

Multiplication of decimals. Multiplication of decimal numbers is based upon the distributive law. Thus, the multiplication (see 40) of 25.78 by 7 is accomplished as a sum of four products, each involving 7 as a factor.

By a process known as short multiplication, this may be carried out as follows (see 41): Since $7 \times 8 = 56$, the 6 is written below the answer line, starting at the right, and the 5 is placed above the 7 in the multiplicand; then $7 \times 7 = 49$, to which is added the 5 that was carried, making 54; the 4 is written in the answer, the 5 being carried as before. The numbers written above the top line are usually carried mentally.

If the multiplier has more than one digit, long multiplication is used (see 42). This process also is based upon the distributive law. The first partial product is 18.046, which was obtained by multiplying 25.78 by .7 by short multiplication. The second partial product is $25.78 \times 5 = 128.90$; the third is $25.78 \times 40 = 1{,}031.2$. The product is the sum of these partial products. It is customary to ignore the decimal point in these partial products but to indent each partial product one more place than the partial product above it. In the product as many digits are pointed off from the right as the sum of the number of digits to the right of the decimal point in the multiplicand and the number of digits to the right of the decimal point in the multiplier.

Division of decimals. If a and b are two whole numbers, $a > b$, there exist two whole numbers q and r such that (see 13) a is written as a sum of two terms, one involving q and one involving r. The process of finding these numbers can be carried out by the processes called long and short division.

If $b < 10$, say $b = 7$, short division may be used (see 43). Then, if $a = 53{,}149$, starting at the left, since $5 < 7$ and considering the first two digits, 53, the largest multiple of 7 that is less than or equal to 53 is $49 = 7 \times 7$.

The 7 is written below the 3, and the remainder, $53 - 49 = 4$, above the 3. This remainder with the next digit in the dividend is 41; the largest multiple of $7 \leqslant 41$ is $35 = 7 \times 5$. The 5 is written below the 1, and the remainder, $41 - 35 = 6$, above the 1. The last remainder is 5. Usually the remainders, which have been written above the dividend, are carried mentally.

It is notable that this process involves guessing the largest multiple of the divisor that is less than a certain number, but only a few guesses are ever required and no practical difficulty is encountered.

Because $53{,}149 = 7{,}592 \times 7 + 5$, one may write (see 44) the ratio of 53,149 and 7 so that 7,592 is the partial quotient and 5 is the remainder.

(43)
$$\begin{array}{r} 4\,615 \\ 7\overline{)53\,149} \\ 7\,592 \end{array}$$

(44)
$$\frac{53{,}149}{7} = 7{,}592\frac{5}{7}$$

(45)
$$\begin{array}{l} 621)\ 83\,742\ (134 \\ \underline{62\,1\ \ } \\ 21\,64 \\ \underline{18\,63} \\ 3\,012 \\ \underline{2\,484} \\ 528 \end{array}$$

(46)
$$83{,}742 = 621 \times 134 + 528$$

(47)
$$\begin{array}{l} 621)\ 83\,742\ (134 \\ 21\,64 \\ 3\,012 \\ 528 \end{array}$$

(48)
$$\frac{83\ 742}{621} = 134\frac{528}{621}$$

(49)
$$\begin{array}{l} 621)\ 83\,742\ (134.8502 \\ 21\,64 \\ 3\,012 \\ 528\,0 \\ 31\,20 \\ 1\,500 \\ 258 \end{array}$$

(50)
$$\frac{83742.0000}{621} = 134.8502$$

(51)
$$\frac{83742}{621} = \frac{837.42}{6.21} = 134.8502$$

(52)
$$\begin{array}{r|l} & 15\ 78.42\ (39.72 \\ & \underline{9\ \ \ \ \ \ \ \ } \\ 69 & 6\ 78 \\ & \underline{6\ 21\ \ \ } \\ 787 & 57\ 42 \\ & \underline{55\ 09\ \ } \\ 7942 & 2\ 33\ 00 \\ & \underline{1\ 58\ 84} \\ & 74\ 16 \end{array}$$

If the divisor exceeds 10, long division is preferable (see 45). To divide 83,742 by 621, the 1 is written as the first digit of the quotient and 621×1 is subtracted from 837 since the largest multiple of $621 \leqslant 837$ is 621×1. Actually what is done is to subtract 621×100 from 83,742, leaving a remainder of 21,642. Since only the 2,164 will be used in the next step, the final 2 of the dividend need not be brought down yet. The largest multiple of $621 \leqslant 2{,}164$ is $621 \times 3 = 1{,}863$, and the remainder is 301. Now the final 2 must be brought down to make 3,012. The largest multiple of $621 \leqslant 3{,}012$ is $621 \times 4 = 2{,}484$, and the remainder is 528. Long division

Thus, the dividend (see 46) is the sum of a product and a remainder.

The preceding method may be considerably abbreviated by mentally subtracting the partial product as it is formed (see 47). Again, the resultant quotient may be expressed as the sum of an integer and a proper fraction (see 48). If it is desired to express the result as a decimal fraction, carried to a given number of decimal places, the above process is merely continued (see 49). Any two decimal numbers may be divided in this manner. The number of digits to the right of the decimal point in the quotient is equal to the number of such digits in the dividend, diminished by the number of such digits in the divisor, with care being taken to add to the dividend all the 0's that are brought down. Thus, in the present example (49) the proper fraction (see 50) is replaced by a decimal fraction. An alternative and exceptionally clear and simple method for determining the position of the decimal point in the quotient is (see 51) to divide both numerator and denominator by an appropriate power of 10; *i.e.*, expressing the divisor as a number lying between 0 and 10, which renders it obvious that the quotient lies between 100 and 1,000.

Divisibility rules. In both theory and practical application of arithmetic it is often important to factor a natural number; that is, to decompose it into numbers that, when multiplied together, will yield the given number as product.

The following tests for divisibility are therefore given. A composite number is divisible

- by 2 if it is even (*i.e.*, if it ends in 0, 2, 4, 6, or 8);
- by 3 if the sum of its digits is divisible by 3;
- by 4 if the number formed by its last two digits is divisible by 4;
- by 5 if it ends in 0 or 5;
- by 6 if it is even and the sum of its digits is divisible by 3;
- by 8 if the number formed by its last 3 digits is divisible by 8;
- by 9 if the sum of its digits is divisible by 9;
- by 10 if it ends with 0;
- by 11 if the difference between the sum of its digits in the odd places and that of the digits in the even places is either 0 or divisible by 11.

Similar rules are easily devised with respect to other divisors by means of appropriate combinations of the rules given above. For example, if a number is to be divisible by 132, say, then it must satisfy the test for divisibility by 3, by 4, and by 11, respectively; *i.e.*, $3 \times 4 \times 11 = 132$.

Evolution. An algorithm for the determination of the square root of a decimal number, such as 1,578.42, is carried out as follows (see 52): Starting at the decimal point, the number is separated into periods of two digits each. The leftmost period is 15, and the largest square $\leqslant 15$ is $9 = 3^2$. The 3 is written at the right and the remainder $15 - 9 = 6$ is brought down. The next period 78 is brought down beside the 6, giving 678. The 3 is doubled and written at the left. By trial it is found that 9 is the largest digit, such that $69 \times 9 = 621 \leqslant 678$. The 9 is written after the 3 in the answer, and the difference, 57, is brought down, followed by the next period, 42. The partial answer 39 is doubled and written at the left under the 69. By trial it is found that $787 \times 7 = 5{,}509 \leqslant 5{,}742$ while $788 \times 8 > 5{,}742$. The next digit in the answer is therefore 7. The process is continued until the desired degree of accuracy is attained. In the present example (see 52) the last remainder, 7,416, exceeds half Method of square root

of the last trial divisor, 7,942, so that 3 is a better approximation than 2 for the last digit. In fact, $39.73^2 = 1{,}578.4729$.

The above process is based upon the relation $(a + b)^2 = a^2 + (2a + b)b$. In each step the part of the square root already obtained is a; the part remaining to be found is b. In the example (see 52), $(a + b)^2 = 1{,}578.42$, $a = 30$. Then $2ab + b^2 = 678.42$. In order to determine the largest integer n in b, one must find the largest integer n (namely 9), such that $(2a + n)n \leqslant 678.42$. The trial divisor is $2a + n = 69$. In the next step $a = 39$, $2ab + b^2 = 57.42$, etc.

Cube and higher roots. The cube root of a number may be calculated by a similar algorithm based upon the relation (see 53) that expresses the difference of two cubes as a product involving a quadratic as a factor. Thus, it is possible to find the approximate cube root of 279,463 by proceeding as in the example (see 54). This last remainder is too large but much closer than the remainder resulting from using 7 as the last digit.

(53) $$(a+b)^3 - a^3 = (3a^2 + 3ab + b^2)b$$

(54)

	279 463.000 (65.38 —
	216
10 800	63 463
900	
25	
11 725	58 625
1 267 500	4 838 000
5 850	
9	
1 273 359	3 820 077
127 922 700	1 017 923 000
156 720	
64	
128 079 484	1 024 635 872

A fourth root is easily obtained as the square root of the square root. Fifth and higher roots can be obtained by an algorithm similar to those just given, based upon the expansion of $(a + b)^n$ by the binomial theorem. The method is cumbersome and is seldom used, for the method of logarithms is easy and rapid.

LOGARITHMS

Napier's contribution

Another marvel of human inventiveness in mathematics is the logarithm, introduced with virtually no precedent in 1614 by the Scottish mathematician John Napier (*q.v.*). As a concept, its hallmarks are simplicity, economy, and utility.

Formally, if $N = b^p$, then $\log_b N = p$, which is read as "the logarithm of the number N to the base b is p." Thus, $\log_{10} 100 = 2$ because $100 = 10^2$, and $\log_2 8 = 3$ because $8 = 2^3$. The simplicity is self-evident. Next, it is noted that if $N = 100{,}000{,}000 = 10^8$, then $\log_{10} N = 8$, and the impressive economy effected in the representation of the number N is observed, cut down as it is from no fewer than nine digits in conventional notation to but one in terms of the corresponding logarithm. These characteristics alone are striking. The full measure of the value of this concept, however, becomes apparent when it is realized that, apart from its simplicity and economy, it reduces all multiplications and divisions to mere additions and subtractions, and all involution (raising to a power) and evolution (determination of square root) to mere multiplications, respectively. The importance of this is that the number of operations generally involved in multiplying (or dividing) two numbers always exceeds the number of operations involved in adding (or subtracting) them. Furthermore, this discrepancy in labour rapidly increases with the magnitudes of the particular numbers involved.

As for illustrative examples, if $N = 10^a$, and $M = 10^b$, then, using the basic results of arithmetic (see above *Fundamental definitions and laws* and *Theory of divisors*), it will be found that $N \cdot M = 10^a \cdot 10^b = 10^{a+b}$, and hence, $\log (N \cdot M) = a + b$, a simple sum. (In the preceding and below, the convention is used that if no base is specified, log refers to base 10.) Similarly, then, $N/M = 10^{a-b}$, so that $\log (N/M) = a - b$, a simple difference.

Next, in considering involution and evolution, in which, for the sake of clarity, it is assumed that n is a real number, if $P = N^n$, then $\log P = \log (N^n) = \log ((10^a)^n) = \log (10^{an}) = an = n \cdot \log N$, and it is seen that involution has been transformed into a mere product. (N has been used here as defined in the preceding paragraph.) Similarly, if $R = N^{1/n}$, then $\log R = \log ((10^a)^{1/n}) = \log (10^{a/n}) = a/n = (1/n) \cdot \log N$, and the tedious process of determining the value of the nth root of N has been reduced to a mere multiplication (or, depending on one's point of view, to a simple division) at most. It is to be noted that in each of these examples the quantities a, b, n need not be positive (or negative) integers: instead, in accord with above *Fundamental definitions and laws* and *Theory of divisors*, they may be fractions, proper or improper, positive or negative. Indeed, they may also be irrational numbers, or logarithms themselves; for example, $\log N^{\log M} = \log N^b = b \cdot \log N = ab$.

The expression $\log 1 = 0$ holds for any base b because $b^0 = 1$ for all (nonzero) values of b. Similarly, for any base b ($\neq 0$), $\log_b b = 1$ because $b^1 = b$.

Sometimes it is necessary to transform logarithms from one base a to another, say b. To see how this transformation can be accomplished, N may now be taken to be a number such that $\log_a N = x$, so that $a^x = N$. It follows that $\log_b N = \log_b (a^x) = x \cdot \log_b a = (\log_b a) \cdot \log_a N$. A case of particular interest is that in which $a = 10$, $b = e$, the symbol e denoting the base of the so-called natural or Naperian logarithms. From the result above, it follows at once that $\log_e N = (\log_e 10) \times (\log_{10} N)$, but $\log_e 10$ is a well-known constant $(2.30258 \cdots)$. The preceding equation thus enables one to compute $\log_e N$ if $\log_{10} N$ is known, and conversely. (C.C.MacD.)

BIBLIOGRAPHY. MUNRO LEAF, *Arithmetic Can Be Fun* (1949); and ISAAC ASIMOV, *Realm of Numbers* (1959), are introductory presentations at an elementary level. HAROLD D. LARSEN, *Arithmetic for Colleges* (1958); SIDNEY G. HACKER, WILFRED E. BARNES, and CALVIN T. LONG, *Fundamental Concepts of Arithmetic* (1963); and CARL B. ALLENDOERFER, *Mathematics for Parents* (1965), are written from the point of view of education. G.H. HARDY and E.M. WRIGHT, *An Introduction to the Theory of Numbers*, 4th ed. (1960); HAROLD DAVENPORT, *The Higher Arithmetic* (1952); and WILLIAM J. LEVEQUE, *Elementary Theory of Numbers* (1962), discuss fundamental concepts associated with the theory of numbers. Works emphasizing the history of arithmetic as a part of number theory include: OYSTEIN ORE, *Number Theory and Its History* (1948); LEONARD E. DICKSON, *History of the Theory of Numbers*, 3 vol. (1919–23, reprinted 1952); TOBIAS DANTZIG, *Number: The Language of Science*, 4th ed. rev. (1959); and DAVID E. SMITH and AUGUSTUS DE MORGAN, *Rara Arithmetica*, 4th ed. (1970). The following are classic and advanced studies: CARL FARBER, *Arithmetik* (1911); CARL F. GAUSS, *Disquisitiones Arithmeticae*, 2nd ed. (1870; Eng. trans., 1966); GOTTLOB FREGE, *Die Grundlagen der arithmetik* (1884; Eng. trans., *Foundations of Arithmetic*, 1968); FELIX KLEIN, *Elementarmathematik vom höheren Standpunkte*, 3rd ed., vol. 1 (1924; Eng. trans., *Elementary Mathematics from an Advanced Standpoint*, vol. 1, *Arithmetic, Algebra, Analysis*, 1932, reprinted 1968); JEAN-PIERRE SERRE, *Cours d'Arithmetique* (1970).

(C.C.MacD./D.E.S./W.J.LeV.)